ECONOMETRIC SOCIETY MONOGRAPHS

# Analysis of Panel Data

Third Edition

Cheng Hsiao

CAMBRIDGE

## Analysis of Panel Data, Third Edition

This book provides a comprehensive, coherent, and intuitive review of panel data methodologies that are useful for empirical analysis. Substantially revised from the second edition, it includes two new chapters on modeling cross-sectionally dependent data and dynamic systems of equations. Some of the more complicated concepts have been further streamlined. Other new material includes correlated random coefficient models, pseudo-panels, duration and count data models, quantile analysis, and alternative approaches for controlling the impact of unobserved heterogeneity in nonlinear panel data models.

Cheng Hsiao is Professor of Economics at the University of Southern California and adjunct professor at Xiamen University. He received his PhD in Economics from Stanford University. He has worked mainly in integrating economic theory with econometric analysis. Professor Hsiao has made extensive contributions in methodology and empirical analysis in the areas of panel data, time series, cross-sectional data, structural modeling, and measurement errors, among other fields. He is the author of the first two editions of *Analysis of Panel Data* and was a co-editor of the *Journal of Econometrics* from 1991 to 2013.

## Econometric Society Monographs

*Editors:*

Professor Donald W. K. Andrews, Yale University
Professor Jeffrey C. Ely, Northwestern University

The Econometric Society is an international society for the advancement of economic theory in relation to statistics and mathematics. The Econometric Society Monograph series is designed to promote the publication of original research contribution of high quality in mathematical economics and theoretical and applied econometrics.

Other Titles in the Series:

G. S. Maddala, *Limited dependent and qualitative variables in econometrics*, 780521241434, 9780521338257
Gerard Debreu, *Mathematical economics: Twenty papers of Gerard Debreu*, 9780521237369, 9780521335614
Jean-Michel Grandmont, *Money and value: A reconsideration of classical and neoclassical monetary economics*, 9780521251419, 9780521313643
Franklin M. Fisher, *Disequilibrium foundations of equilibrium economics*, 9780521378567
Andreu Mas-Colell, *The theory of general equilibrium: A differentiable approach*, 9780521265140, 9780521388702
Truman F. Bewley, Editor, *Advances in econometrics – Fifth World Congress (Volume I)*, 9780521467261
Truman F. Bewley, Editor, *Advances in econometrics – Fifth World Congress (Volume II)*, 9780521467254
Hervé Moulin, *Axioms of cooperative decision making*, 9780521360555, 9780521424585
L. G. Godfrey, *Misspecification tests in econometrics: The Lagrange multiplier principle and other approaches*, 9780521424592
Tony Lancaster, *The econometric analysis of transition data*, 9780521437899
Alvin E. Roth and Marilda A. Oliviera Sotomayor, Editors, *Two-sided matching: A study in game-theoretic modeling and analysis*, 9780521437882
Wolfgang Härdle, *Applied nonparametric regression*, 9780521429504
Jean-Jacques Laffont, Editor, *Advances in economic theory – Sixth World Congress (Volume I)*, 9780521484596
Jean-Jacques Laffont, Editor, *Advances in economic theory – Sixth World Congress (Volume II)*, 9780521484602
Halbert White, *Estimation, inference and specification*, 9780521252805, 9780521574464
Christopher Sims, Editor, *Advances in econometrics – Sixth World Congress (Volume I)*, 9780521444590, 9780521566100
Christopher Sims, Editor, *Advances in econometrics – Sixth World Congress (Volume II)*, 9780521444606, 9780521566094
Roger Guesnerie, *A contribution to the pure theory of taxation*, 9780521629560
David M. Kreps and Kenneth F. Wallis, Editors, *Advances in economics and econometrics – Seventh World Congress (Volume I)*, 9780521589833
David M. Kreps and Kenneth F. Wallis, Editors, *Advances in economics and econometrics – Seventh World Congress (Volume II)*, 9780521589826
David M. Kreps and Kenneth F. Wallis, Editors, *Advances in economics and econometrics – Seventh World Congress (Volume III)*, 9780521580137, 9780521589819

*Continued on page following the index*

# Analysis of Panel Data

## Third Edition

Cheng Hsiao
University of Southern California

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

32 Avenue of the Americas, New York, NY 10013-2473, USA

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107657632



First edition published 1986
Second edition 2003
Third edition 2014

Printed in the United States of America

*A catalog record for this publication is available from the British Library.*

*Library of Congress Cataloging in Publication Data*
Hsiao, Cheng, 1943–
Analysis of panel data / Cheng Hsiao. – Third edition.
pages cm. – (Econometric society monographs)
Includes bibliographical references and index.
ISBN 978-1-107-03869-1 (hardback) – ISBN 978-1-107-65763-2 (paperback)
1. Econometrics. 2. Panel analysis. 3. Analysis of variance. I. Title.
HB139.H75 2014
330.01′5195–dc23 2014006652

ISBN 978-1-107-03869-1 Hardback
ISBN 978-1-107-65763-2 Paperback

*To my wife, Amy Mei-Yun*
*and my children*
*Irene Chiayun*
*Allen Chenwen*
*Michael Chenyee*
*Wendy Chiawen*

# Contents

*Preface to the Third Edition* *page* xvii
*Preface to the Second Edition* xix
*Preface to the First Edition* xxi

1 Introduction 1
1.1 Introduction 1
1.2 Advantages of Panel Data 4
1.3 Issues Involved in Utilizing Panel Data 10
1.3.1 Unobserved Heterogeneity across Individuals and over Time 10
1.3.2 Incidental Parameters and Multidimensional Statistics 13
1.3.3 Sample Attrition 13
1.4 Outline of the Monograph 14

2 Homogeneity Tests for Linear Regression Models (Analysis of Covariance) 17
2.1 Introduction 17
2.2 Analysis of Covariance 18
2.3 An Example 24

3 Simple Regression with Variable Intercepts 31
3.1 Introduction 31
3.2 Fixed-Effects Models: Least-Squares Dummy Variable Approach 34
3.3 Random Effects Models: Estimation of Variance-Components Models 39
3.3.1 Covariance Estimation 40
3.3.2 Generalized Least-Squares (GLS) Estimation 41
3.3.3 Maximum-Likelihood Estimation 45

3.4 Fixed Effects or Random Effects 47
3.4.1 An Example 47
3.4.2 Conditional Inference or Unconditional (Marginal) Inference 48
3.5 Tests for Misspecification 56
3.6 Models with Time- and/or Individual-Invariant Explanatory Variables and Both Individual- and Time-Specific Effects 58
3.6.1 Estimation of Models with Individual-Specific Variables 58
3.6.2 Estimation of Models with Both Individual and Time Effects 61
3.7 Heteroscedasticity and Autocorrelation 64
3.7.1 Heteroscedasticity 64
3.7.2 Models with Serially Correlated Errors 65
3.7.3 Heteroscedasticity Autocorrelation Consistent Estimator for the Covariance Matrix of the CV Estimator 68
3.8 Models with Arbitrary Error Structure – Chamberlain $\pi$-Approach 69
Appendix 3A: Consistency and Asymptotic Normality of the Minimum-Distance Estimator 75
Appendix 3B: Characteristic Vectors and the Inverse of the Variance–Covariance Matrix of a Three-Component Model 77

4 Dynamic Models with Variable Intercepts 80
4.1 Introduction 80
4.2 The CV Estimator 82
4.3 Random-Effects Models 84
4.3.1 Bias in the OLS Estimator 85
4.3.2 Model Formulation 86
4.3.3 Estimation of Random-Effects Models 89
4.3.4 Testing Some Maintained Hypotheses on Initial Conditions 106
4.3.5 Simulation Evidence 107
4.4 An Example 108
4.5 Fixed-Effects Models 111
4.5.1 Transformed Likelihood Approach 112
4.5.2 Minimum Distance Estimator 114
4.5.3 Relations between the Likelihood-Based Estimator and the GMM 116
4.5.4 Issues of Random versus Fixed-Effects Specification 119

4.6 Estimation of Dynamic Models with Arbitrary Serial Correlations in the Residuals 121
4.7 Models with Both Individual- and Time-Specific Additive Effects 122
Appendix 4A: Derivation of the Asymptotic Covariance Matrix of Feasible MDE 129
Appendix 4B: Large $N$ and $T$ Asymptotics 130

5 Static Simultaneous-Equations Models 136
5.1 Introduction 136
5.2 Joint Generalized Least-Squares Estimation Technique 140
5.3 Estimation of Structural Equations 144
5.3.1 Estimation of a Single Equation in the Structural Model 144
5.3.2 Estimation of the Complete Structural System 149
5.4 Triangular System 152
5.4.1 Identification 153
5.4.2 Estimation 155
5.4.3 An Example 162
Appendix 5A 164

6 Variable-Coefficient Models 167
6.1 Introduction 167
6.2 Coefficients that Vary over Cross-Sectional Units 170
6.2.1 Fixed-Coefficient Model 170
6.2.2 Random-Coefficient Model 172
6.3 Coefficients that Vary over Time and Cross-Sectional Units 180
6.3.1 The Model 180
6.3.2 Fixed-Coefficient Model 182
6.3.3 Random-Coefficient Model 183
6.4 Coefficients that Evolve over Time 186
6.4.1 The Model 186
6.4.2 Predicting $\beta_t$ by the Kalman Filter 188
6.4.3 Maximum-Likelihood Estimation 191
6.4.4 Tests for Parameter Constancy 192
6.5 Coefficients that Are Functions of Other Exogenous Variables 193
6.6 A Mixed Fixed- and Random-Coefficients Model 196
6.6.1 Model Formulation 196
6.6.2 A Bayes Solution 198
6.6.3 Random or Fixed Differences? 201
6.7 Dynamic Random-Coefficients Models 206
6.8 Two Examples 212

6.8.1 Liquidity Constraints and Firm Investment Expenditure 212
6.8.2 Aggregate versus Disaggregate Analysis 217
6.9 Correlated Random-Coefficients Models 220
6.9.1 Introduction 220
6.9.2 Identification with Cross-Sectional Data 221
6.9.3 Estimation of the Mean Effects with Panel Data 223
Appendix 6A: Combination of Two Normal Distributions 228

7 Discrete Data 230
7.1 Introduction 230
7.2 Some Discrete-Response Models for Cross-Sectional Data 230
7.3 Parametric Approach to Static Models with Heterogeneity 235
7.3.1 Fixed-Effects Models 236
7.3.2 Random-Effects Models 242
7.4 Semiparametric Approach to Static Models 246
7.4.1 Maximum Score Estimator 247
7.4.2 A Root-*N* Consistent Semiparametric Estimator 249
7.5 Dynamic Models 250
7.5.1 The General Model 250
7.5.2 Initial Conditions 252
7.5.3 A Conditional Approach 255
7.5.4 State Dependence versus Heterogeneity 261
7.5.5 Two Examples 264
7.6 Alternative Approaches for Identifying State Dependence 270
7.6.1 Bias-Adjusted Estimator 270
7.6.2 Bounding Parameters 274
7.6.3 Approximate Model 276

8 Sample Truncation and Sample Selection 281
8.1 Introduction 281
8.2 An Example – Nonrandomly Missing Data 292
8.2.1 Introduction 292
8.2.2 A Probability Model of Attrition and Selection Bias 292
8.2.3 Attrition in the Gary Income-Maintenance Experiment 296
8.3 Tobit Models with Random Individual Effects 298
8.4 Fixed-Effects Estimator 299
8.4.1 Pairwise Trimmed Least-Squares and Least Absolute Deviation Estimators for Truncated and Censored Regressions 299

8.4.2 A Semiparametric Two-Step Estimator for the Endogenously Determined Sample Selection Model 311
8.5 An Example: Housing Expenditure 313
8.6 Dynamic Tobit Models 317
8.6.1 Dynamic Censored Models 317
8.6.2 Dynamic Sample Selection Models 324

9 Cross-Sectionally Dependent Panel Data 327
9.1 Issues of Cross-Sectional Dependence 327
9.2 Spatial Approach 329
9.2.1 Introduction 329
9.2.2 Spatial Error Model 332
9.2.3 Spatial Lag Model 333
9.2.4 Spatial Error Models with Individual-Specific Effects 334
9.2.5 Spatial Lag Model with Individual-Specific Effects 335
9.2.6 Spatial Dynamic Panel Data Models 336
9.3 Factor Approach 337
9.4 Group Mean Augmented (Common Correlated Effects) Approach to Control the Impact of Cross-Sectional Dependence 342
9.5 Test of Cross-Sectional Independence 344
9.5.1 Linear Model 344
9.5.2 Limited Dependent-Variable Model 348
9.5.3 An Example – A Housing Price Model of China 350
9.6 A Panel Data Approach for Program Evaluation 352
9.6.1 Introduction 352
9.6.2 Definition of Treatment Effects 352
9.6.3 Cross-Sectional Adjustment Methods 354
9.6.4 Panel Data Approach 359

10 Dynamic System 369
10.1 Panel Vector Autoregressive Models 370
10.1.1 "Homogeneous" Panel VAR Models 370
10.1.2 Heterogeneous Vector Autoregressive Models 377
10.2 Cointegrated Panel Models and Vector Error Correction 379
10.2.1 Properties of Cointegrated Processes 379
10.2.2 Estimation 381
10.3 Unit Root and Cointegration Tests 386
10.3.1 Unit Root Tests 386
10.3.2 Tests of Cointegration 394
10.4 Dynamic Simultaneous Equations Models 397
10.4.1 The Model 397

10.4.2 Likelihood Approach 398
10.4.3 Method of Moments Estimator 401

11 Incomplete Panel Data 403
11.1 Rotating or Randomly Missing Data 403
11.2 Pseudo-Panels (or Repeated Cross-Sectional Data) 408
11.3 Pooling of Single Cross-Sectional and Single Time Series Data 411
11.3.1 Introduction 411
11.3.2 The Likelihood Approach to Pooling Cross-Sectional and Time Series Data 413
11.3.3 An Example 416
11.4 Estimating Distributed Lags in Short Panels 418
11.4.1 Introduction 418
11.4.2 Common Assumptions 419
11.4.3 Identification Using Prior Structure on the Process of the Exogenous Variable 421
11.4.4 Identification Using Prior Structure on the Lag Coefficients 425
11.4.5 Estimation and Testing 428

12 Miscellaneous Topics 430
12.1 Duration Model 430
12.2 Count Data Model 438
12.3 Panel Quantile Regression 445
12.4 Simulation Methods 448
12.5 Data with Multilevel Structures 453
12.6 Errors of Measurement 455
12.7 Nonparametric Panel Data Models 461

13 A Summary View 464
13.1 Benefits of Panel Data 464
13.1.1 Increasing Degrees of Freedom and Lessening the Problem of Multicollinearity 464
13.1.2 Identification and Discrimination between Competing Hypotheses 465
13.1.3 Reducing Estimation Bias 467
13.1.4 Generating More Accurate Predictions for Individual Outcomes 468
13.1.5 Providing Information on Appropriate Level of Aggregation 468
13.1.6 Simplifying Computation and Statistical Inference 469

13.2 Challenges for Panel Data Analysis 469
13.2.1 Modeling Unobserved Heterogeneity 469
13.2.2 Controlling the Impact of Unobserved Heterogeneity in Nonlinear Models 470
13.2.3 Modeling Cross-Sectional Dependence 471
13.2.4 Multidimensional Asymptotics 472
13.2.5 Sample Attrition 472
13.3 A Concluding Remark 473

*References* 475

*Author Index* 507

*Subject Index* 513

# Preface to the Third Edition

Panel data econometrics is one of the most exciting fields in econometrics today. The possibility of modeling more realistic behavioral hypotheses and challenging methodological issues, together with the increasing availability of panel data have led to the phenomenal proliferation of studies on panel data. This edition is a substantial revision of the second edition. Two new chapters on modeling cross-sectionally dependent data and the dynamic system of equations have been added. Some of the more complicated concepts have been further streamlined and new material on correlated random-coefficients models, pseudo-panels, duration and count data models, quantile analysis, alternative approaches for controlling the impact of unobserved heterogeneity in nonlinear panel data models, inference with data having both large cross section and long time series, etc. have been incorporated into existing chapters. It is hoped that the present version can provide a reasonably comprehensive, coherent, and intuitive review of panel methodologies that are useful for empirical analysis. However, no single monograph can do justice to the huge amount of literature in this field. I apologize for any omissions of the important contributions in panel data analysis.

I would like to thank the former and current Cambridge University Press publisher, Scott Parris and Karen Maloney, for their encouragement and support for this project. I am grateful to *Econometrica*, International Monetary Fund, *Financial Times, Journal of the American Statistical Association, Journal of Applied Econometrics, Journal of Econometrics, Regional Science and Urban Economics, Review of Economic Studies*, the University of Chicago Press, and Elsevier for permission to reproduce some of the materials published here. Thanks to Kristin Purdy and Kate Gavino for assistance in obtaining the copyright permissions and K. Bharadwaj, S. Shankar, J. Penney, and T. Kornak for their excellent work on copyediting and typesetting. During the process of preparing this monograph I have benefited from the excellent working conditions provided by the University of Southern California, Xiamen University, the City University of Hong Kong, and Hong Kong University of Science and Technology and the partial research support of the China Natural Science Foundation grant #71131008. I am grateful to Sena Schlessinger for her excellent

typing of various drafts of the monograph; R. Matzkin and two referees; and J. C. Duan, R. Koenker, C. Lambache, L. F. Lee, X. Lu, M. H. Pesaran, H. R. Moon, L. Su, T. Wansbeek, and J. H. Yu for helpful comments on some parts of the book. I would like to thank Q. Zhou for pointing out many typos and oversights in an early version of the manuscript; Michael Hsiao for preparing Tables 1.1, 6.9–6.11, and 9.1; Shui Wan for preparing Tables 9.2–9.5 and Figures 9.1–9.4; and T. Wang for kindly making the source files for Table 12.1 and Figures 12.1 and 12.2 available. In spite of their help, no doubt errors remain. I apologize for the errors and would appreciate being informed of any that are spotted.

# Preface to the Second Edition

Since the publication of the first edition of this monograph in 1986, there has been a phenomenal growth of articles dealing with panel data. According to the *Social Science Citation Index*, there were 29 articles related to panel data in 1989. But in 1997 there were 518; in 1998, 553; and in 1999, 650. The increasing attention is partly due to the greater availability of panel data sets, which can better answer questions of substantial interest than a single set of cross-sectional or time series data can, and partly due to the rapid growth in computational power of the individual researcher. It is furthermore motivated by the internal methodological logic of the subject (e.g., Trognon (2000)).

The current version is a substantial revision of the first edition. The major additions are essentially on nonlinear panel data models of discrete choice (Chapter 7) and sample selection (Chapter 8); a new Chapter 10 on miscellaneous topics such as simulation techniques, large $N$ and $T$ theory, unit root and cointegration tests, multiple level structure, and cross-sectional dependence; and new sections on estimation of dynamic models (4.5–4.7), Bayesian treatment of models with fixed and random coefficients (6.6–6.8), and repeated cross-sectional data (or pseudopanels), etc. In addition, many of the discussions in old chapters have been updated. For instance, the notion of strict exogeneity is introduced, and estimators are also presented in a generalized method of moments framework to help link the assumptions that are required for the identification of various models. The discussion of fixed and random effects is updated in regard to restrictions on the assumption about unobserved specific effects, etc.

The goal of this revision remains the same as that of the first edition. It aims to bring up to date a comprehensive analytical framework for the analysis of a greater variety of data. The emphasis is on formulating appropriate statistical inference for issues shaped by important policy concerns. The revised edition of this monograph is intended as neither an encyclopedia nor a history of panel data econometrics. I apologize for the omissions of many important contributions. A recount of the history of panel data econometrics can be found in Nerlove (2000). Some additional issues and references can also be found in a survey by Arellano and Honoré (2001) and in four recent

edited volumes – Matyás and Sevester (1996); Hsiao, Lahiri, Lee, and Pesaran (1999); Hsiao, Morimune, and Powell (2001); and Krishnakumar and Ronchetti (2000). Software is reviewed by Blanchard (1996).

I would like to thank the editor, Scott Parris, for his encouragement and assistance in preparing the revision, and Andrew Chesher and two anonymous readers for helpful comments on an early draft. I am also very grateful to E. Kyriazidou for her careful and detailed comments on Chapters 7 and 8, S. Chen and J. Powell for their helpful comments and suggestions on Chapter 8, H. R. Moon for the section on large panels, Sena Schlessinger for her expert typing of the manuscript except for Chapter 7, Yan Shen for carefully proofreading the manuscript and for expertly typing Chapter 7, and Siyan Wang for drawing the figures for Chapter 8. Of course, all remaining errors are mine. The kind permissions to reproduce parts of articles by James Heckman, C. Manski, Daniel McFadden, Ariel Pakes, *Econometrica, Journal of the American Statistical Association, Journal of Econometrics, Regional Science and Urban Economics, Review of Economic Studies*, the University of Chicago Press, and Elsevier Science are also gratefully acknowledged.

# Preface to the First Edition

Recently, empirical research in economics has been enriched by the availability of a wealth of new sources of data: cross sections of individuals observed over time. These allow us to construct and test more realistic behavioral models that could not be identified using only a cross section or a single time series data set. Nevertheless, the availability of new data sources raises new issues. New methods are constantly being introduced, and points of view are changing. An author preparing an introductory monograph has to select the topics to be included. My selection involves controlling for unobserved individual and/or time characteristics to avoid specification bias and to improve the efficiency of the estimates. The more basic and more commonly used methods are treated here, although to some extent the coverage is a matter of taste. Some examples of applications of the methods are also given, and the uses, computational approaches, and interpretations are discussed.

I am very much indebted to C. Manski and to a reader for Cambridge University Press, as well as to G. Chamberlain and J. Ham, for helpful comments and suggestions. I am also grateful to Mario Tello Pacheco, who read through the manuscript and made numerous suggestions concerning matters of exposition and corrections of errors of every magnitude. My appreciation also goes to V. Bencivenga, A. C. Cameron, T. Crawley, A. Deaton, E. Kuh, B. Ma, D. McFadden, D. Mountain, G. Solon, G. Taylor, and K. Y. Tsui, for helpful comments, and Sophia Knapik and Jennifer Johnson, who patiently typed and retyped innumerable drafts and revisions. Of course, in material like this it is easy to generate errors, and the reader should put the blame on the author for any remaining errors.

Various parts of this monograph were written while I was associated with Bell Laboratories, Murray Hill, Princeton University, Stanford University, the University of Southern California, and the University of Toronto. I am grateful to these institutions for providing me with secretarial and research facilities and, most of all, stimulating colleagues. Financial support from the National Science Foundation, U.S.A., and from the Social Sciences and Humanities Research Council of Canada is gratefully acknowledged.

CHAPTER 1

# Introduction

## 1.1 INTRODUCTION

A longitudinal, or panel, data set is one that follows a given sample of individuals over time, and thus provides multiple observations on each individual in the sample. Panel data have become widely available in both the developed and developing countries. In the United States, two of the most prominent panel data sets are the National Longitudinal Surveys of Labor Market Experience (NLS) and the University of Michigan's Panel Study of Income Dynamics (PSID).

The NLS was initiated in 1966. The surveys include data about a wide range of attitudes, behaviors, and events related to schooling, employment, marriage, fertility, training, child care, health, and drug and alcohol use. The original four cohorts were men aged 45 to 59 in 1966, young men aged 14 to 24 in 1966, women aged 30 to 44 in 1967, and young women aged 14 to 24 in 1968. Table 1.1 summarizes the size and the span of years each group of these original samples has been interviewed, as well as the currently ongoing surveys (the NLS Handbook 2005 U.S. Department of Labor, Bureau of Labor Statistics). In 1979, the NLS expanded to include a nationally representative sample of 12,686 young men and women who were 14 to 22 years old. These individuals were interviewed annually through 1994 and are currently interviewed on a biennial basis (NLS79). In 1986, the NLS started surveys of the children born to women who participated in the National Longitudinal Survey of Youth 1979 (NLS79 Children and Young Adult). In addition to all the mother's information from the NLS79, the child survey includes additional demographic and development information. For children aged 10 years and older, information has been collected from the children biennially since 1988. The National Longitudinal Survey of Youth 1997 (NLS97) consists of a nationally representative sample of youths who were 12 to 16 years old as of December 31, 1996. The original sample includes 8,984 respondents. The eligible youths continued to be interviewed on an annual basis. The survey collects extensive information on respondents' labor market behavior and educational experiences. The survey also includes data on the youths' families and community backgrounds. It

Table 1.1. *The span and sample sizes of the National Longitudinal Surveys*

| Cohorts | Age | Birth year | Beginning year/ ending year/ | Beginning sample size | Number interviewed in year |
|---|---|---|---|---|---|
| Older men | 45–59 | 4/2/1907–4/1/1921 | 1966/1990 | 5,020 | 2,092 (1990) |
| Mature women | 30–44 | 4/2/1923–4/1/1937 | 1967/2003 | 5,083 | 2,237 (2003) |
| Young men | 14–24 | 4/2/1942–4/1/1952 | 1966/1981 | 5,225 | 3,398 (1981) |
| Young women | 14–24 | 1944–1954 | 1968/2003 | 5,159 | 2,287 (2003) |
| NLS79 | 14–21 | 1957–1964 | 1979/– | 12,686 | 7,724 (2002) |
| NLS79 children | 0–14 | — | 1986/– | 5,255 | 7,467 (2002) |
| NLS79 young adult | 15–22 | — | 1994/– | 980 | 4,238 (2002) |
| NLS97 | 12–16 | 1980–1984 | 1997/– | 8,984 | 7,756 (2004) |

*Source:* Bureau of Labor Statistics, *National Longitudinal Surveys Handbook (2005)*.

documents the transition from school to work and from adolescence to adulthood. Access on NLS data and documentation is available online at the NLS *Product Availability Center* at NLSinfo.org.

The PSID began in 1968 with collection of annual economic information from a representative national sample of about 6,000 families and 15,000 individuals and their descendants and has continued to the present. The PSID gathers data on the family as a whole and on individuals residing within the family, emphasizing the dynamic and interactive aspects of family economics, demography, and health. The data set contains more than 5,000 variables, including employment, income, and human capital variables, as well as information on housing, travel to work, and mobility. PSID data were collected annually from 1968 to 1997 and biennially after 1997. They are available online in the PSID Data Center at no charge (PSID.org). In addition to the NLS and PSID data sets there are several other panel data sets that are of interest to economists, and these have been cataloged and discussed by Borus (1981) and Juster (2001); also see Ashenfelter and Solon (1982) and Becketti et al. (1988).[1]

In Europe, various countries have their annual national or more frequent surveys: the Netherlands Socio-Economic Panel (SEP), the German Social Economics Panel (GSOEP), the Luxembourg's Social Economic Panel (PSELL), the British Household Panel Survey (BHPS), and so forth. Starting in 1994, the National Data Collection Units (NDU) of the Statistical Office of the European Communities, "in response to the increasing demand in the European Union for comparable information across the member states on income, work and employment, poverty and social exclusion, housing, health, and many other diverse social indicators concerning living conditions of private households and persons" (Eurostat 1996), have begun coordinating and linking existing national panels with centrally designed standardized multipurpose annual

[1] For examples of marketing data, see Beckwith (1972); for biomedical data, see Sheiner, Rosenberg, and Melmon (1972); for a financial-market database, see Dielman, Nantell, and Wright (1980).

longitudinal surveys. For instance, the Manheim Innovation Panel (MIP) and the Manheim Innovation Panel-Service Sector (MIP-S), started in 1993 and 1995, respectively, contain annual surveys of innovative activities such as product innovations, expenditure on innovations, expenditure on research and development (R&D), factors hampering innovations, the stock of capital, wages and skill structures of employees, and so on of German firms with at least five employees in manufacturing and service sectors. The survey methodology is closely related to the recommendations on innovation surveys manifested in the *Oslo Manual* of the Organisation for Economic Co-operation and Development (OECD) and Eurostat, thereby yielding international comparable data on innovation activities of German firms. The 1993 and 1997 surveys also become part of the European Community Innovation Surveys CIS I and CIS II (for details, see Janz et al. 2001). Similarly, the European Community Household Panel (ECHP) is meant to represent the population of the European Union (EU) at the household and individual level. The ECHP contains information on demographics, labor force behavior, income, health, education and training, housing, migration, and so forth. With the exception of Sweden, the ECHP now covers 14 of the 15 countries (Peracchi 2000). Detailed statistics from the ECHP are published in Eurostat's reference data based New Cronos in three domains, namely health, housing, and "ILC" – income and living conditions.[2]

Panel data have also become increasingly available in developing countries. In these countries, there may not have a long tradition of statistical collection. It is especially important to obtain original survey data to answer many significant and important questions. Many international agencies have sponsored and helped to design panel surveys. For instance, the Dutch non-government organization (NGO), Investing in Children and their Societies (ICS), Africa collaborated with the Kenya Ministry of Health have carried out a Primary School Deworming Project (PDSP). The project took place in a poor and densely settled farming region in western Kenya – the Busia district. The 75 project schools include nearly all rural primary schools in this area, with more than 30,000 enrolled pupils between the ages of 6 and 18 years from 1998 to 2001. The World Bank has also sponsored and helped to design many panel surveys. For instance, the Development Research Institute of the Research Center for Rural Development of the State Council of China, in collaboration with the World Bank, undertook an annual survey of 200 large Chinese township and village enterprises from 1984 to 1990 (Hsiao et al. 1998).

There is also a worldwide concerted effort to collect panel data about aging, retirement, and health in many countries. It started with the biannual panel data of the Health and Retirement Study in the USA (HRS; http://www.rand.org/labor/aging/dataprod/, http://hrsonline.isr.umich.edu/), followed by the English

[2] Potential users interested in the ECHP can access and download the detailed documentation of the ECHP users' database (ECHP UDP) from the ECHP website: http://forum.europa.eu.int/irc/dsis/echpane/info/data/information.html.

Longitudinal Study of Aging (ELSA; http://www.ifs.org.uk/elsa/), and the Survey of Health, Aging and Retirement in Europe (SHARE; http://www.share-project.org/), which covers 11 continental European countries, but more European countries, as well as Israel, will be added. Other countries are also developing similar projects, in particular several Asian countries. These data sets are collected with a multidisciplinary view and are set up such that the data are highly comparable across countries. They contain a great deal of information about people of (approximately) 50 years of age and older and their households. Among others, this involves labor history and present labor force participation, income from various sources (labor, self-employment, pensions, social security, assets), wealth in various categories (stocks, bonds, pension plans, housing), various aspects of health (general health, diseases, problems with activities of daily living and mobility), subjective predictions of retirement, and actual retirement. Using these data, researchers can study various substantive questions that cannot be studied from other (panel) studies, such as the development of health at older age and the relation between health and retirement. Furthermore, owing to the highly synchronized questionnaires across a large number of countries, it becomes possible to study the role of institutional factors, such as pension systems, retirement laws, and social security plans, on labor force participation and retirement, and so forth (for further information, see Wansbeek and Meijer 2007).

## 1.2 ADVANTAGES OF PANEL DATA

A panel data set for economic research possesses several major advantages over conventional cross-sectional or time series data sets (e.g., Hsiao 1985a, 1995, 2001, 2007) such as:

1. More accurate inference of model parameters. Panel data usually give researchers a large number of data points, increasing the degrees of freedom and reducing the collinearity among explanatory variables – hence improving the efficiency of econometric estimates.
2. Greater capacity for constructing more realistic behavioral hypotheses. By blending interindividual differences with intraindividual dynamics, longitudinal data allow a researcher to analyze a number of important economic questions that cannot be addressed using cross-sectional or time series data sets. For instance, a typical assumption for the analysis using cross-sectional data is that individuals with the same conditional variables, $\mathbf{x}$, have the same expected value, $E(y_i \mid \mathbf{x}_i = \mathbf{a}) = E(y_j \mid \mathbf{x}_j = \mathbf{a})$. Under this assumption, if a cross-sectional sample of married women is found to have an average yearly labor force participation rate of 50 percent, it would imply that each woman in a homogeneous population has a 50 percent chance of being in the labor force in any given year. Each woman would be expected

to spend half of her married life in the labor force, and half out of the labor force, and job turnover would be expected to be frequent, with an average job duration of two years. However, as Ben-Porath (1973) illustrated that the cross-sectional sample could be drawn from a heterogeneous population, 50 percent of the women were from the population that always works and 50 percent from the population that never works. In this case, there is no turnover, and current information about work status is a perfect predictor of future work status. The availability of panel data makes it possible to discriminate between these two models. The sequential observations for a number of individuals allows a researcher to utilize individual labor force histories to estimate the probability of participation in different subintervals of the life cycle.

The difficulties of making inferences about the dynamics of change from cross-sectional evidence are seen as well in other labor market situations. Consider the impact of unionism on economic behavior (e.g., Freeman and Medoff, 1981). Those economists who tend to interpret the observed differences between union and nonunion firms/employees as largely real believe that unions and the collective bargaining process fundamentally alter key aspects of the employment relationship: compensation, internal and external mobility of labor, work rules, and environment. Those economists who regard union effects as largely illusory tend to posit that the real world is close enough to satisfying the conditions of perfect competition; they believe that the observed union/nonunion differences are due mainly to differences between union and nonunion firms/workers prior to unionism or post-union sorting. Unions do not raise wages in the long run, because firms react to higher wages (forced by the union) by hiring better quality workers. If one believes the former view, the coefficient of the dummy variable for union status in a wage or earning equation is a measure of the effect of unionism. If one believes the latter view, then the dummy variable for union status could be simply acting as a proxy for worker quality. A single cross-sectional data set usually cannot provide a direct choice between these two hypotheses, because the estimates are likely to reflect interindividual differences inherent in comparisons of *different* people or firms. However, if panel data are used, one can distinguish these two hypotheses by studying the wage differential for a worker moving from a nonunion firm to a union firm, or vice versa. If one accepts the view that unions have no effect, then a worker's wage should not be affected when he moves from a nonunion firm to a union firm, if the quality of this worker is constant over time. On the other hand, if unions truly do raise wages, then, holding worker quality constant, the worker's wage should rise as he moves to a union firm from a nonunion firm. By following given

individuals or firms over time as they change status (say from nonunion to union, or vice versa), one can construct a proper recursive structure to study the before/after effect.

3. Uncovering dynamic relationships. Because of institutional or technological rigidities or inertia in human behavior, "economic behavior is inherently dynamic" (Nerlove 2000). Microdynamic and macrodynamic effects typically cannot be estimated using a cross-sectional data set. A single time series data set often cannot provide good estimates of dynamic coefficients either. For instance, consider the estimation of a distributed-lag model:

$$y_t = \sum_{\tau=0}^{h} \beta_\tau x_{t-\tau} + u_t, \quad t = 1, \ldots, T, \tag{1.2.1}$$

where $x_t$ is an exogenous variable and $u_t$ is a random disturbance term. In general, $x_t$ is near $x_{t-1}$, and still nearer $2x_{t-1} - x_{t-2} = x_{t-1} + (x_{t-1} - x_{t-2})$; fairly strict multicollinearities appear among $h + 1$ explanatory variables, $x_1, x_{t-1}, \ldots, x_{t-h}$. Hence, there is not sufficient information to obtain precise estimates of any of the lag coefficients without specifying, a priori, that each of them is a function of only a very small number of parameters [e.g., Almon lag, rational distributed lag, Malinvaud (1970)]. If panel data are available, we can utilize the interindividual differences in $x$ values to reduce the problem of collinearity, thus allowing us to drop the ad hoc conventional approach of constraining the lag coefficients $\{\beta_\tau\}$ and to impose a different prior restriction to estimate an unconstrained distributed-lag model.

4. Controlling the impact of omitted variables (or individual or time heterogeneity). The use of panel data provides a means of resolving or reducing the magnitude of a key econometric problem that often arises in empirical studies, namely, the often heard assertion that the real reason one finds (or does not find) certain effects is because of omitted (mismeasured, not observed) variables that correlate with explanatory variables. By utilizing information on both the intertemporal dynamics and the individuality of the entities being investigated, one is better able to control in a more natural way for the effects of missing or unobserved variables. For instance, consider a simple regression model:

$$\begin{aligned} y_{it} = \alpha^* + \boldsymbol{\beta}'\mathbf{x}_{it} + \boldsymbol{\rho}'\mathbf{z}_{it} + u_{it}, \quad & i = 1, \ldots, N, \\ & t = 1, \ldots T, \end{aligned} \tag{1.2.2}$$

where $\mathbf{x}_{it}$ and $\mathbf{z}_{it}$ are $k_1 \times 1$ and $k_2 \times 1$ vectors of exogenous variables; $\alpha^*$, $\boldsymbol{\beta}$, and $\boldsymbol{\rho}$ are $1 \times 1$, $k_1 \times 1$, and $k_2 \times 1$ vectors of constants, respectively; and the error term $u_{it}$ is independently, identically

distributed over $i$ and $t$, with mean zero and variance $\sigma_u^2$. It is well known that the least-squares regression of $y_{it}$ on $\mathbf{x}_{it}$ and $\mathbf{z}_{it}$ yields unbiased and consistent estimators of $\alpha^*$, $\boldsymbol{\beta}$, and $\boldsymbol{\rho}$. Now suppose that $\mathbf{z}_{it}$ values are unobservable, and the covariances between $\mathbf{x}_{it}$ and $\mathbf{z}_{it}$ are nonzero. Then the least-squares regression coefficients of $y_{it}$ on $\mathbf{x}_{it}$ are biased. However, if repeated observations for a group of individuals are available, they may allow us to get rid of the effects of $\mathbf{z}$ through a linear transformation. For example, if $\mathbf{z}_{it} = \mathbf{z}_i$ for all $t$ (i.e., $\mathbf{z}$ values stay constant through time for a given individual but vary across individuals), we can take the first difference of individual observations over time and obtain

$$y_{it} - y_{i,t-1} = \boldsymbol{\beta}'(\mathbf{x}_{it} - \mathbf{x}_{i,t-1}) + (u_{it} - u_{i,t-1}), \quad \begin{matrix} i = 1, \ldots, N, \\ t = 2, \ldots, T. \end{matrix} \tag{1.2.3}$$

Similarly if $\mathbf{z}_{it} = \mathbf{z}_t$ for all $i$ (i.e., $\mathbf{z}$ values stay constant across individuals at a given time, but exhibit variation through time), we can take the deviation from the mean across individuals at a given time and obtain

$$y_{it} - \overline{y}_t = \boldsymbol{\beta}'(\mathbf{x}_{it} - \overline{\mathbf{x}}_t) + (u_{it} - \overline{u}_t), \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{matrix} \tag{1.2.4}$$

where $\overline{y}_t = (1/N)\sum_{i=1}^N y_{it}$, $\overline{\mathbf{x}}_t = (1/N)\sum_{i=1}^N \mathbf{x}_{it}$ and $\overline{u}_t = (1/N)\sum_{i=1}^N u_{it}$. Least-squares regression of (1.2.3) or (1.2.4) now provides unbiased and consistent estimates of $\boldsymbol{\beta}$. Nevertheless, if we have only a single cross-sectional data set ($T = 1$) for the former case ($\mathbf{z}_{it} = \mathbf{z}_i$), or a single time series data set ($N = 1$) for the latter case ($\mathbf{z}_{it} = \mathbf{z}_t$), such transformations cannot be performed. We cannot get consistent estimates of $\boldsymbol{\beta}$ unless there exist instruments that correlate with $\mathbf{x}$ but do not correlate with $\mathbf{z}$ and $u$.

MaCurdy's (1981) work on the life cycle labor supply of prime age males under certainty is an example of this approach. Under certain simplifying assumptions, MaCurdy shows that a worker's labor supply function can be written as (1.2.2), where $y$ is the logarithm of hours worked, $\mathbf{x}$ is the logarithm of the real wage rate, and $z$ is the logarithm of the worker's (unobserved) marginal utility of initial wealth, which, as a summary measure of a worker's lifetime wages and property income, is assumed to stay constant through time but to vary across individuals (i.e., $z_{it} = z_i$). Given the economic problem, not only does $\mathbf{x}_{it}$ correlate with $z_i$, but every economic variable that could act as an instrument for $\mathbf{x}_{it}$ (such as education) also correlates with $z_i$. Thus, in general, it is not possible to estimate $\boldsymbol{\beta}$ consistently from a

cross-sectional data set,[3] but if panel data are available, one can consistently estimate $\boldsymbol{\beta}$ by first differencing (1.2.2).

The "conditional convergence" of the growth rate is another example (e.g., Durlauf 2001; Temple 1999). Given the role of transitional dynamics, it is widely agreed that growth regressions should control for the steady-state level of income (e.g., Barro and Sala-i-Martin 1995; Mankiw, Romer, and Weil 1992). Thus, the growth rate regression model typically includes investment ratio, initial income, and measures of policy outcomes such as school enrollment and the black market exchange rate premium as regressors. However, an important component, the initial level of a country's technical efficiency, $z_{i0}$, is omitted because this variable is unobserved. Because a country that is less efficient is also more likely to have lower investment rate or school enrollment, one can easily imagine that $z_{i0}$ correlates with the regressors and the resulting cross-sectional parameters estimates are subject to omitted variable bias. However, with panel data one can eliminate the influence of initial efficiency by taking the first difference of individual country observations over time as in (1.2.3).

5. Generating more accurate predictions for individual outcomes. Pooling the data could yield more accurate predictions of individual outcomes than generating predictions using the data on the individual in question if individual behaviors are similar conditional on certain variables. When data on individual history are limited, panel data provide the possibility of learning an individual's behavior by observing the behavior of others. Thus, it is possible to obtain a more accurate description of an individual's behavior by supplementing observations of the individual in question with data on other individuals (e.g., Hsiao, Appelbe, and Dineen 1993; Hsiao, Mountain, Tsui, and Chan 1989).
6. Providing micro-foundations for aggregate data analysis. In macro analysis economists often invoke the "representative agent" assumption. However, if micro-units are heterogeneous, not only can the time series properties of aggregate data be very different from those of disaggregate data (e.g., Granger 1980; Lewbel 1992; Pesaran 2003), but also policy evaluation based on aggregate data may be grossly misleading. Furthermore, the prediction of aggregate outcomes using aggregate data can be less accurate than the prediction based on aggregating micro-equations (e.g., Hsiao, Shen, and Fujiki 2005). Panel data containing time series observations for a number of individuals are ideal for investigating the "homogeneity" versus "heterogeneity" issue.

[3] This assumes that there are no other variables, such as consumption, that can act as a proxy for $z_i$. Most North American data sets do not contain information on consumption.

7. Simplifying computation and statistical inference. Panel data involve at least two dimensions, a cross-sectional dimension and a time series dimension. Under normal circumstances one would expect that the computation of panel data estimator or inference would be more complicated than estimators based on cross-sectional or time series data alone. However, in certain cases, the availability of panel data actually simplifies computation and inference. For instance:
   a. Analysis of nonstationary time series. When time series data are not stationary, the large sample approximations of the distributions of the least-squares or maximum likelihood estimators are no longer normally distributed (e.g., Anderson 1959; Dickey and Fuller (1979, 1981); Phillips and Durlauf 1986). But if panel data are available, one can invoke the central limit theorem across cross-sectional units to show that the limiting distributions of many estimators remain asymptotically normal and the Wald type test statistics are asymptotically chi-square distributed. (e.g., Binder, Hsiao, and Pesaran 2005; Im, Pesaran, and Shin 2003; Levin, Lin, and Chu 2002; Phillips and Moon 1999).
   b. Measurement errors. Measurement errors can lead to underidentification of an econometric model (e.g., Aigner, Hsiao, Kapteyn, and Wansbeek 1984). The availability of multiple observations for a given individual or at a given time may allow a researcher to make different transformations to induce different and deducible changes in the estimators, and hence to identify an otherwise unidentified model (e.g., Biørn 1992; Griliches and Hausman 1986; Wansbeek and Koning 1989).
   c. Dynamic Tobit models. When a variable is truncated or censored, the actual realized value is unobserved. If an outcome variable depends on previous realized value and the previous realized value are unobserved, one has to take integration over the truncated range to obtain the likelihood of observables. In a dynamic framework with multiple missing values, the multiple integration is computationally infeasible. For instance, consider a dynamic Tobit model of the form

$$y_{it}^* = \gamma y_{i,t-1}^* + \beta x_{it} + \epsilon_{it} \tag{1.2.5}$$

   where $y^*$ is unobservable, and what we observe is $y$, where $y_{it} = y_{it}^*$ if $y_{it}^* > 0$ and 0 otherwise. The conditional density of $y_{it}$ given $y_{i,t-1} = 0$ is much more complicated than the case if $y_{i,t-1}^*$ is known because the joint density of $(y_{it}, y_{i,t-1})$ involves the integration of $y_{i,t-1}^*$ from $-\infty$ to 0. Moreover, when there are a number of censored observations over time, the full implementation of the maximum likelihood principle is almost impossible. However, with panel data, the estimation of $\gamma$ and $\beta$ can be simplified considerably by simply focusing on the subset of data where

$y_{i,t-1} > 0$ because the joint density of $f(y_{it}, y_{i,t-1})$ can be written as the product of the conditional density $f(y_{i,t} \mid y_{i,t-1})$ and the marginal density of $y_{i,t-1}$. But if $y^*_{i,t-1}$ is observable, the conditional density of $y_{it}$ given $y_{i,t-1} = y^*_{i,t-1}$ is simply the density of $\epsilon_{it}$ (Arellano, Bover, and Labeaga 1999).

## 1.3 ISSUES INVOLVED IN UTILIZING PANEL DATA

### 1.3.1 Unobserved Heterogeneity across Individuals and over Time

The oft-touted power of panel data derives from their theoretical ability to isolate the effects of specific actions, treatments, or more general policies. This theoretical ability is based on the assumption that economic data are generated from controlled experiments in which the outcomes are random variables with a probability distribution that is a smooth function of the various variables describing the conditions of the experiment. If the available data were in fact generated from simple controlled experiments, standard statistical methods could be applied. Unfortunately, most panel data come from the very complicated process of everyday economic life. In general, different individuals may be subject to the influences of different factors. In explaining individual behavior, one may extend the list of factors ad infinitum. It is neither feasible nor desirable to include all the factors affecting the outcome of all individuals in a model specification because the purpose of modeling is not to mimic the reality but to capture the essential forces affecting the outcome. It is typical to leave out those factors that are believed to have insignificant impacts or are peculiar to certain individuals. However, when important factors peculiar to a given individual are left out, the typical assumption that economic variable $y$ is generated by a parametric probability distribution function $F(y \mid \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is an $m$-dimensional real vector, identical for all individuals at all times, may not be a realistic one. If the conditional density of $y_{it}$ given $\mathbf{x}_{it}$ varies across $i$ and over $t$, $f_{it}(y_{it} \mid \mathbf{x}_{it})$, the conditions for the fundamental theorems for statistical analysis, the law of large numbers and central limit theorem, may not hold. The challenge of panel data analysis is how to model the heterogeneity across individuals and over time that are not captured by $\mathbf{x}$. A popular approach to control the unobserved heterogeneity is to let the parameters characterizing the conditional distribution of $y_{it}$ given $\mathbf{x}_{it}$ to vary across $i$ and over $t$, $f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\theta}_{it})$. However, if no structure is imposed on $\boldsymbol{\theta}_{it}$, there will be more unknown parameters than the number of available sample observations. To allow the inference about the relationship between $y_{it}$ and $\mathbf{x}_{it}$, $\boldsymbol{\theta}_{it}$ is often decomposed into two components, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_{it}$, where $\boldsymbol{\beta}$ is assumed identical across $i$ and over $t$, and $\boldsymbol{\gamma}_{it}$ is allowed to vary with $i$ and $t$. The common parameters, $\boldsymbol{\beta}$, are called *structural parameters* in the statistical literature. When $\boldsymbol{\gamma}_{it}$ are treated as random variables, it is called the *random effects* model (e.g., Balestra and Nerlove 1966). When $\boldsymbol{\gamma}_{it}$ are treated as fixed unknown constants, it is called the fixed effects model (e.g., Kuh 1963). The parameters $\gamma_{it}$ vary with $i$ and $t$ and are

called *incidental parameters* in the statistical literature because when sample sizes increase, so do the unknown $\boldsymbol{\gamma}_{it}$. There is also an issue about whether $\boldsymbol{\gamma}_{it}$ correlates with the conditional variables (or regressors) (e.g., Mundlak 1978a; Hausman 1978; Chamberlain 1984).

The focus of panel data analysis is how to control the impact of unobserved heterogeneity to obtain valid inference on the common parameters, $\boldsymbol{\beta}$. For instance, in a linear regression framework, suppose unobserved heterogeneity is individual specific and time invariant. Then this individual-specific effect on the outcome variable $y_{it}$ could either be invariant with the explanatory variables $\mathbf{x}_{it}$ or interact with $\mathbf{x}_{it}$. A linear regression model for $y_{it}$ to take account of both possibilities with a single explanatory variable $x_{it}$ could be postulated as

$$\begin{aligned} y_{it} = \alpha_i^* + \beta_i x_{it} + u_{it}, \quad & i = 1, \ldots, N, \\ & t = 1, \ldots, T, \end{aligned} \tag{1.3.1}$$

where $u_{it}$ is the error term, uncorrelated with $x$, with mean zero and constant variance $\sigma_u^2$. The parameters $\alpha_i^*$ and $\beta_i$ may be different for different cross-sectional units, although they stay constant over time. Following this assumption, a variety of sampling distributions may occur. Such sampling distributions can seriously mislead the least-squares regression of $y_{it}$ on $x_{it}$ when all $NT$ observations are used to estimate the model:

$$\begin{aligned} y_{it} = \alpha^* + \beta x_{it} + u_{it}, \quad & i = 1, \ldots, N, \\ & t = 1, \ldots, T. \end{aligned} \tag{1.3.2}$$

For instance, consider the situation that the data are generated as either in case 1 or case 2:

**Case 1:** Heterogeneous intercepts ($\alpha_i^* \neq \alpha_j^*$), homogeneous slope ($\beta_i = \beta_j$). We use graphs to illustrate the likely biases of estimating (1.3.2) because $\alpha_i^* \notin \alpha_j^*$ and $\beta_i = \beta_j$. In these graphs, the broken-line circles represent the point scatter for an individual over time, and the broken straight lines represent the individual regressions. Solid lines serve the same purpose for the least-squares regression of (1.3.2) using all $NT$ observations. A variety of circumstances may arise in this case, as shown in Figures 1.1, 1.2, and 1.3. All of these figures depict situations in which biases arise in pooled least-squares estimates of (1.3.2) because of heterogeneous intercepts. Obviously, in these cases, pooled regression ignoring heterogeneous intercepts should never be used. Moreover, the direction of the bias of the pooled slope estimates cannot be identified a priori; it can go either way.

**Case 2:** Heterogeneous intercepts and slopes ($\alpha_i^* \neq \alpha_j^*, \beta_i \neq \beta_j$). In Figures 1.4 and 1.5 the point scatters are not shown, and the circled numbers signify the individuals whose regressions have been included in the analysis. For the example depicted in Figure 1.4, a straightforward pooling of all $NT$ observations, assuming identical parameters for all cross-sectional units, would lead

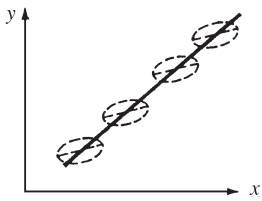


Fig. 1.1

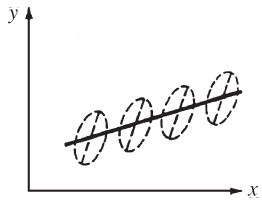


Fig. 1.2

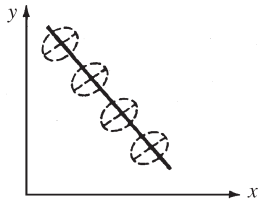


Fig. 1.3

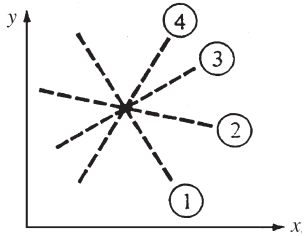


Fig. 1.4

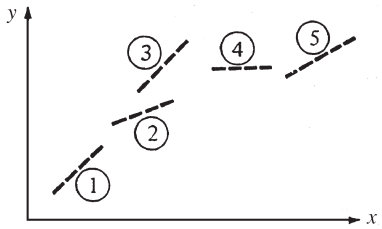


Fig. 1.5

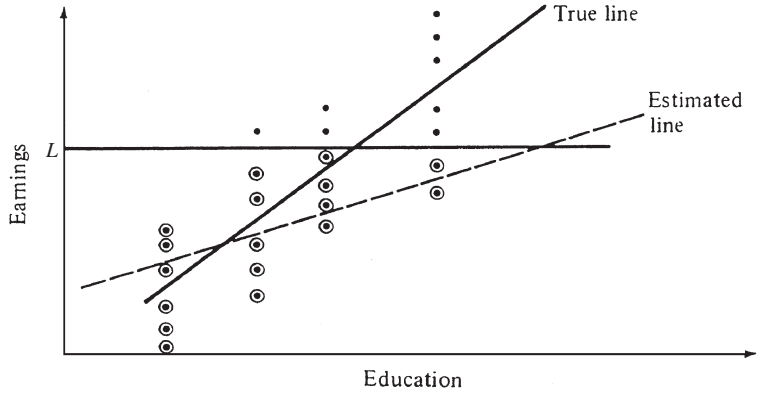


Fig. 1.6

to nonsensical results because they would represent an average of coefficients that differ greatly across individuals. Nor does Figure 1.5 make any sense, because it gives rise to the false inference that the pooled relation is curvilinear. In either case, the classic paradigm of the "representative agent" simply does not hold and a common slope parameter model makes no sense.

These are some of the likely biases when parameter heterogeneities among cross-sectional units are ignored. Similar patterns of bias will also arise if the intercepts and slopes vary through time, even though for a given time period they are identical for all individuals. More elaborate patterns than those depicted here are, of course, likely to occur (e.g., Chesher and Lancaster 1983;

Kuh 1963). Moreover, if $\gamma_{it}$ is persistent over time, say $\gamma_{it}$ represents the impact of individual time-invariant variables so $\gamma_{it} = \gamma_i$, then $E(y_{it} \mid \mathbf{x}_{it}, \gamma_i, y_{i,t-1}) = E(y_{it} \mid \mathbf{x}_{it}, \gamma_i) \neq E(y_{it} \mid \mathbf{x}_{it})$ and $E(y_{it} \mid \mathbf{x}_{it}) \neq E(y_{it} \mid \mathbf{x}_{it}, y_{i,t-1})$. The latter inequality could lead an investigator to infer falsely that there is state dependence. However, the presence of $y_{i,t-1}$ improves the prediction of $y_{it}$ because $y_{i,t-1}$ serves as a proxy for the omitted $\gamma_i$. The observed state dependence is spurious (e.g., Heckman 1978a, 1981b).

### 1.3.2 Incidental Parameters and Multidimensional Statistics

Panel data contain at least two dimensions: a cross-sectional dimension of size $N$ and a time series dimension of size $T$. The observed data can take the form of either $N$ is fixed and $T$ is large; or $T$ is fixed and $N$ is large; or both $N$ and $T$ are finite or large. When the individual time-varying parameters $\boldsymbol{\gamma}_{it}$ are treated as fixed constants (the fixed effects model), and either $N$ or $T$ is fixed, it raises the *incidental parameters* issue because when sample size increases, so do the unknown $\boldsymbol{\gamma}_{it}$. The classical law of large numbers or central limit theorems rely on the assumption that the number of unknowns stay constant when sample size increases. If $\boldsymbol{\gamma}_{it}$ affects the observables $y_{it}$ linearly, simple linear transformation can eliminate $\boldsymbol{\gamma}_{it}$ from the transformed model (e.g., Anderson and Hsiao 1981, 1982; Kuh 1963). However, if $\boldsymbol{\gamma}_{it}$ affects $y_{it}$ nonlinearly, no general rule of transformation to eliminate the incidental parameters exists. Specific structure of a nonlinear model needs to be explored to find appropriate transformation to eliminate the incidental parameters (e.g., Chamberlain 1980; Honoré 1992; Honoré and Kyriazidou 2000a; Manski 1985).

When $N$ and $T$ are of similar magnitude or $N$ and $T$ increase at the same or arbitrary rate, Phillips and Moon (2000) show that naively by first applying one-dimensional asymptotics, followed by expanding the sample size in another dimension could lead to misleading inferences. The multidimensional asymptotics are quite complex (e.g., Alvarez and Arellano 2003; Hahn and Kuersteiner 2002 or some general remarks in Hsiao 2012). Moreover, when $N$ and $T$ are large, the cross-sectional dependence (e.g., Anselin 1988; Bai 2009; Lee 2004; Pesaran 2004) or time series properties of a variable (e.g., unit root or cointegration test) could impact inference significantly.

### 1.3.3 Sample Attrition

Another frequently observed source of bias in both cross-sectional and panel data is that the sample may not be randomly drawn from the population. Panel data follow a given individual over time. One of the notable feature of the NLS in Table 1.1 is the attrition over time. For instance, there were 5,020 individuals for the older men group in the NLS when the annual interview started in 1966. By 1990, when the annual interview of this group was stopped, only 2,092 individuals were left. When attrition is behaviorally related, the observed sample could no longer be viewed as a random sample.

Another example that the observed sample may not be viewed as a random sample is that the New Jersey negative income tax experiment excluded all families in the geographic areas of the experiment who had incomes above 1.5 times the officially defined poverty level. When the truncation is based on earnings, uses of the data that treat components of earnings (specifically, wages or hours) as dependent variables will often create what is commonly referred to as selection bias (e.g., Hausman and Wise 1977; Heckman 1976a, 1979; Hsiao 1974b).

For ease of exposition, we shall consider a cross-sectional example to get some idea of how using a nonrandom sample may bias the least-square estimates. We assume that in the population the relationship between earnings ($y$) and exogenous variables ($\mathbf{x}$), including education, intelligence, and so forth, is of the form

$$y_i = \boldsymbol{\beta}'\mathbf{x}_i + u_i, \quad i = 1, \ldots, N, \tag{1.3.3}$$

where the disturbance term $u_i$ is independently distributed with mean zero and variance $\sigma_u^2$. If the participants of an experiment are restricted to have earnings less than $L$, the selection criterion for families considered for inclusion in the experiment can be stated as follows:

$$\begin{aligned} y_i &= \boldsymbol{\beta}'\mathbf{x}_i + u_i \leq L, \quad \text{included}, \\ y_i &= \boldsymbol{\beta}'\mathbf{x}_i + u_i > L, \quad \text{excluded}. \end{aligned} \tag{1.3.4}$$

For simplicity, we assume that the values of exogenous variables, except for the education variable, are the same for each observation. In Figure 1.6 we let the upward-sloping solid line indicate the "average" relation between education and earnings and the dots represent the distribution of earnings around this mean for selected values of education. All individuals with earnings above a given level $L$, indicated by the horizontal line, would be eliminated from this experiment. In estimating the effect of education on earnings, we would observe only the points below the limit (circled) and thus would tend to underestimate the effect of education using ordinary least squares.[4] In other words, the sample selection procedure introduces correlation between right-hand variables and the error term, which leads to a downward-biased regression line, as the dashed line in Figure 1.6 indicates.

## 1.4 OUTLINE OF THE MONOGRAPH

The source of sample variation critically affects the formulation and inferences of many economic models. This monograph takes a pedagogical approach. We focus on controlling for the impact of unobserved heterogeneity in cross-sectional unit $i$ at time $t$ to draw inferences about certain characteristics of the population that are of interest to an investigator or policymaker. Instead of

[4] For a formal treatment of this, see Chapter 8.

presenting all the issues simultaneously in a general-to-specific manner, we take a pedagogical approach, introducing the various complications successively. We first discuss linear models because they remain widely used. We first briefly review the classic test of homogeneity for a linear regression model (analysis of covariance procedures) in Chapter 2. We then relax the assumption that the parameters that characterize all temporal cross-sectional sample observations are identical and examine a number of specifications that allow for differences in behavior across individuals as well as over time. For instance, a single equation model with observations of $y$ depending on a vector of characteristics $\mathbf{x}$ can be written in the following form:

1. Slope coefficients are constant, and the intercept varies over individuals:
$$y_{it} = \alpha_i^* + \sum_{k=1}^{K} \beta_k x_{kit} + u_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T. \end{matrix} \tag{1.4.1}$$
2. Slope coefficients are constant, and the intercept varies over individuals and time:
$$y_{it} = \alpha_{it}^* + \sum_{k=1}^{K} \beta_k x_{kit} + u_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T. \end{matrix} \tag{1.4.2}$$
3. All coefficients vary over individuals:
$$y_{it} = \alpha_i^* + \sum_{k=1}^{K} \beta_{ki} x_{kit} + u_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T. \end{matrix} \tag{1.4.3}$$
4. All coefficients vary over time and individuals:
$$y_{it} = \alpha_{it}^* + \sum_{k=1}^{K} \beta_{kit} x_{kit} + u_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T. \end{matrix} \tag{1.4.4}$$

In each of these cases the model can be classified further depending on whether the coefficients are assumed to be random or fixed.

We first focus on models in which the unobserved individual or time heterogeneity is invariant with respect to variations in explanatory variables, the constant slopes, and variable intercepts models (1.4.1) and (1.4.2) because they provide simple yet reasonably general alternatives to the assumption that parameters take values common to all agents at all times. Static models with variable intercepts are discussed in Chapter 3, dynamic models in Chapter 4, and simultaneous-equations models in Chapter 5. Chapter 6 relaxes the assumption that the time or individual invariance of unobserved heterogeneity with explanatory variables by allowing the unobserved heterogeneity to interact with them. Chapters 7 and 8 discuss the difficulties of controlling unobserved

heterogeneity in nonlinear models by focusing on two types of widely used models, the discrete data and sample selection models, respectively. Chapter 9 considers the issues of modeling cross-sectional dependence. Chapter 10 considers models for dynamic systems. The incomplete panel data issues such as rotating sample, pooling of a series of independent cross sections (pseudo-panel), pooling of a single cross section and a single time-series data, and estimating distributed-lag models in short panels are discussed in Chapter 11. Chapter 12 discusses miscellaneous topics such as duration data and count data models, panel quantile regression, simulation methods, data with multilevel structures, measurement errors, and the nonparametric approach. A summary view is provided in Chapter 13.

CHAPTER 2

# Homogeneity Tests for Linear Regression Models (Analysis of Covariance)

## 2.1 INTRODUCTION

Suppose we have sample observations of characteristics of $N$ individuals over $T$ time periods denoted by $y_{it}, x_{kit}$, $i = 1, \ldots N, t = 1, \ldots, T, k = 1, \ldots, K$. Conventionally, observations of $y$ are assumed to be the random outcomes of some experiment with a probability distribution conditional on vectors of the characteristics $\mathbf{x}$ and a fixed number of parameters $\boldsymbol{\theta}$, $f(y \mid \mathbf{x}, \boldsymbol{\theta})$. When panel data are used, one of the ultimate goals is to use all available information to make inferences on $\boldsymbol{\theta}$. For instance, a simple model commonly postulated is that $y$ is a linear function of $\mathbf{x}$. Yet to run a least-squares regression with all $NT$ observations, we need to assume that the regression parameters take value common to all cross-sectional units for all time periods. If this assumption is not valid, as shown in Chapter 1, Section 1.2, the pooled least-squares estimates may lead to false inferences. Thus, as a first step toward full exploitation of the data, we often test whether or not parameters characterizing the random outcome variable $y$ stay constant across all $i$ and $t$.

In the case of linear regression model, a widely used procedure to identify the source of sample variation preliminarily is the analysis of covariance (ANCOVA) test. The name "analysis of variance" (ANOVA) is often reserved for a particular category of linear hypotheses that stipulate that the expected value of a random variable $y$ depends only on the class (defined by one or more factors) to which the individual considered belongs, but excludes tests relating to regressions. On the other hand, ANCOVA models are of a mixed character involving genuine exogenous variables, as do regression models, and at the same time allowing the true relation for each individual to depend on the class to which the individual belongs, as do the usual ANOVA models.

A linear model commonly used to assess the effects of both quantitative and qualitative factors is postulated as

$$y_{it} = \alpha_{it}^{*} + \boldsymbol{\beta}_{it}'\mathbf{x}_{it} + u_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{matrix} \tag{2.1.1}$$

where $\alpha_{it}^{*}$ and $\boldsymbol{\beta}_{it}' = (\beta_{1it}, \beta_{2it}, \ldots, \beta_{Kit})$ are $1 \times 1$ and $1 \times K$ vectors of constants that vary across $i$ and $t$, respectively, $\mathbf{x}_{it}' = (x_{1it}, \ldots, x_{Kit})$ is a $1 \times K$

vector of exogenous variables, and $u_{it}$ is the error term. Two aspects of the estimated regression coefficients can be tested: first, the homogeneity of regression slope coefficients; second, the homogeneity of regression intercept coefficients. The test procedure as suggested by Kuh (1963) has three main steps:

1. Test whether or not slopes and intercepts simultaneously are homogeneous among different individuals at different times.
2. Test whether or not the regression slopes collectively are the same.
3. Test whether or not the regression intercepts are the same.

It is obvious that if the hypothesis of overall homogeneity (step 1) is accepted, the testing procedure will go no further. However, should the overall homogeneity hypothesis be rejected, the second step of the analysis is to decide if the regression slopes are the same. If this hypothesis of homogeneity is not rejected, one then proceeds to the third and final test to determine the equality of regression intercepts. In principle, step 1 is separable from steps 2 and 3.[1]

Although this type of analysis can be performed on several dimensions, as described by Scheffé (1959) or Searle (1971), only one-way ANCOVA has been widely used. Therefore, here we present only the procedures for performing one-way ANCOVA.

## 2.2 ANALYSIS OF COVARIANCE

Model (2.1.1) only has descriptive value. It can neither be estimated nor used to generate prediction because the available degrees of freedom, $NT$, is less than the number of parameters, $NT(K+1)$ number of parameters, characterizing the distribution of $y_{it}$. A structure has to be imposed on (2.1.1) before any inference can be made. To start with, we assume that parameters are constant over time, but can vary across individuals. Thus, we can postulate a separate regression for each individual:

$$y_{it} = \alpha_i^* + \boldsymbol{\beta}_i' \mathbf{x}_{it} + u_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T. \end{matrix} \tag{2.2.1}$$

Three types of restrictions can be imposed on (2.2.1), namely:

$H_1$: Regression slope coefficients are identical, and intercepts are not. That is,

$$y_{it} = \alpha_i^* + \boldsymbol{\beta}' \mathbf{x}_{it} + u_{it}. \tag{2.2.2}$$

$H_2$: Regression intercepts are the same, and slope coefficients are not. That is,

$$y_{it} = \alpha^* + \boldsymbol{\beta}_i' \mathbf{x}_{it} + u_{it}. \tag{2.2.3}$$

[1] Note that even if the homogeneity hypothesis is rejected, some useful information can be found in pooling the data, as long as the source of sample variability can be identified. For details, see later chapters.

$H_3$: Both slope and intercept coefficients are the same. That is,

$$y_{it} = \alpha^* + \boldsymbol{\beta}'\mathbf{x}_{it} + u_{it}. \tag{2.2.4}$$

Because it is seldom a meaningful question to ask if the intercepts are the same when the slopes are unequal, we shall ignore the type of restrictions postulated by (2.2.3). We shall refer to (2.2.1) as the unrestricted model, (2.2.2) as the individual-mean or cell-mean corrected regression model, and (2.2.4) as the pooled regression.

Let

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}, \tag{2.2.5}$$

$$\bar{\mathbf{x}}_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{it}, \tag{2.2.6}$$

be the means of $y$ and $\mathbf{x}$, respectively, for the $i$th individual. The least-squares estimates of $\boldsymbol{\beta}_i$ and $\alpha_i^*$ in the unrestricted model (2.2.1) are given by[2]

$$\hat{\boldsymbol{\beta}}_i = W_{xx,i}^{-1} W_{xy,i}, \quad \hat{\alpha}_i = \bar{y}_i - \hat{\boldsymbol{\beta}}_i'\bar{\mathbf{x}}_i, \quad i = 1, \ldots, N, \tag{2.2.7}$$

where

$$\begin{aligned} W_{xx,i} &= \sum_{t=1}^{T} (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)', \\ W_{xy,i} &= \sum_{t=1}^{T} (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(y_{it} - \bar{y}_i), \\ W_{yy,i} &= \sum_{t=1}^{T} (y_{it} - \bar{y}_i)^2. \end{aligned} \tag{2.2.8}$$

In the ANCOVA terminology, equations (2.2.7) are called within-group estimates. The $i$th-group residual sum of squares is $\mathrm{RSS}_i = W_{yy,i} - W_{xy,i}' W_{xx,i}^{-1} W_{xy,i}$. The unrestricted residual sum of squares is

$$S_1 = \sum_{i=1}^{N} \mathrm{RSS}_i. \tag{2.2.9}$$

The least-squares regression of the individual mean corrected model yields parameter estimates

$$\begin{aligned} \hat{\boldsymbol{\beta}}_w &= W_{xx}^{-1} W_{xy}, \\ \hat{\alpha}_i^* &= \bar{y}_i - \hat{\boldsymbol{\beta}}_w'\bar{\mathbf{x}}_i, \qquad i = 1, \ldots, N, \end{aligned} \tag{2.2.10}$$

[2] We assume that $T > K + 1$. For details of this, see Chapter 3, Section 3.2.

where

$$W_{xx} = \sum_{i=1}^{N} W_{xx,i} \quad \text{and} \quad W_{xy} = \sum_{i=1}^{N} W_{xy,i}.$$

Let $W_{yy} = \sum_{i=1}^{N} W_{yy,i}$; the residual sum of squares of (2.2.2) is

$$S_2 = W_{yy} - W'_{xy} W_{xx}^{-1} W_{xy}. \tag{2.2.11}$$

The least-squares regression of the pooled model (2.2.4) yields parameter estimates

$$\hat{\boldsymbol{\beta}} = T_{xx}^{-1} T_{xy}, \qquad \hat{\alpha}^* = \bar{y} - \hat{\boldsymbol{\beta}}' \bar{\mathbf{x}}, \tag{2.2.12}$$

where

$$T_{xx} = \sum_{i=1}^{N} \sum_{t=1}^{T} (\mathbf{x}_{it} - \bar{\mathbf{x}})(\mathbf{x}_{it} - \bar{\mathbf{x}})',$$

$$T_{xy} = \sum_{i=1}^{N} \sum_{t=1}^{T} (\mathbf{x}_{it} - \bar{\mathbf{x}})(y_{it} - \bar{y}),$$

$$T_{yy} = \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \bar{y})^2,$$

$$\bar{y} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} y_{it}, \qquad \bar{\mathbf{x}} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{x}_{it}.$$

The (overall) residual sum of squares is

$$S_3 = T_{yy} - T'_{xy} T_{xx}^{-1} T_{xy}. \tag{2.2.13}$$

Under the assumption that the $u_{it}$ are independently normally distributed over $i$ and $t$ with mean 0 and variance $\sigma_u^2$, $F$ tests can be used to test the restrictions postulated by (2.2.2) and (2.2.4). In effect, (2.2.2) and (2.2.4) can be viewed as (2.2.1) subject to various types of linear restrictions. For instance, the hypothesis of heterogeneous intercepts but homogeneous slopes [equation (2.2.2)] can be reformulated as (2.2.1) subject to $(N-1)K$ linear restrictions:

$$H_1 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_N.$$

The hypothesis of common intercept and slopes can be viewed as (2.2.1) subject to $(K+1)(N-1)$ linear restrictions:

$$H_3 : \alpha_1^* = \alpha_2^* = \cdots = \alpha_N^*,$$
$$\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_N.$$

Thus, application of the ANCOVA test is equivalent to the ordinary hypothesis test based on the sums of squared residuals from linear regression outputs.

The unrestricted residual sum of squares $S_1$ divided by $\sigma_u^2$ has a chi square ($\chi^2$) distribution with $NT - N(K+1)$ degrees of freedom. The increment in the explained sum of squares due to allowing for the parameters to vary across $i$ is measured by $(S_3 - S_1)$. Under $H_3$, the restricted residual sum of squares $S_3$ divided by $\sigma_u^2$ has a $\chi^2$ distribution with $NT - (K+1)$ degrees of freedom, and $(S_3 - S_1)/\sigma_u^2$ has a $\chi^2$ distribution with $(N-1)(K+1)$ degrees of freedom. Because $(S_3 - S_1)/\sigma_u^2$ is independent of $S_1/\sigma_u^2$, the $F$ statistic,

$$F_3 = \frac{(S_3 - S_1)/[(N-1)(K+1)]}{S_1/[NT - N(K+1)]}, \tag{2.2.14}$$

can be used to test $H_3$. If $F_3$ with $(N-1)(K+1)$ and $N(T-K-1)$ degrees of freedom is not significant, we pool the data and estimate a single equation of (2.2.4). If the $F$ ratio is significant, a further attempt is usually made to find out if the nonhomogeneity can be attributed to heterogeneous slopes or heterogeneous intercepts.

Under the hypothesis of heterogeneous intercepts but homogeneous slopes ($H_1$), the residual sum of squares of (2.2.2), $S_2 = W_{yy} - W'_{xy} W_{xx}^{-1} W_{xy}$, divided by $\sigma_u^2$ has a $\chi^2$ distribution with $N(T-1) - K$ degrees of freedom. The $F$ test of $H_1$ is thus given by

$$F_1 = \frac{(S_2 - S_1)/[(N-1)K]}{S_1/[NT - N(K+1)]}. \tag{2.2.15}$$

If $F_1$ with $(N-1)K$ and $NT - N(K+1)$ degrees of freedom is significant, the test sequence is naturally halted and model (2.2.1) is treated as the maintained hypothesis. If $F_1$ is not significant, we can then determine the extent to which nonhomogeneities can arise in the intercepts.

If $H_1$ is accepted, one can also apply a conditional test for homogeneous intercepts, namely,

$$H_4 : \alpha_1^* = \alpha_2^* = \cdots = \alpha_N^* \quad \text{given} \quad \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_N.$$

The unrestricted residual sum of squares now is $S_2$, and the restricted residual sum of squares is $S_3$. The reduction in the residual sum of squares in moving from (2.2.4) to (2.2.2) is $(S_3 - S_2)$. Under $H_4$, $S_3$ divided by $\sigma_u^2$ is $\chi^2$ distributed with $NT - (K+1)$ degrees of freedom, and $S_2$ divided by $\sigma_u^2$ is $\chi^2$ distributed with $N(T-1) - K$ degrees of freedom. Because $S_2/\sigma_u^2$ is independent of $(S_3 - S_2)/\sigma_u^2$, which is $\chi^2$ distributed with $N-1$ degrees of freedom, the $F$ test for $H_4$ is

$$F_4 = \frac{(S_3 - S_2)/(N-1)}{S_2/[N(T-1) - K]} \tag{2.2.16}$$

We can summarize these tests in an ANCOVA table (Table 2.1).

Alternatively, we can assume that coefficients are constant across individuals at a given time, but can vary over time. Hence, a separate regression can be

Table 2.1. *Covariance tests for homogeneity*

| Source of variation | Residual sum of squares | Degrees of freedom | Mean squares |
|---|---|---|---|
| Within group with heterogeneous intercept and slope | $S_1 = \sum_{i=1}^{N} \left( W_{yy,i} - W'_{xy,i} W_{xx,i}^{-1} W_{xy,i} \right)$ | $N(T-K-1)$ | $S_1/N(T-K-1)$ |
| Constant slope: heterogeneous intercept | $S_2 = W_{yy} - W'_{xy} W_{xx}^{-1} W_{xy}$ | $N(T-1)-K$ | $S_2/[N(T-1)-K]$ |
| Common intercept and slope | $S_3 = T_{yy} - T'_{xy} T_{xx}^{-1} T_{xy}$ | $NT-(K+1)$ | $S_3/[NT-(K+1)]$ |

*Notation:*

| | |
|---|---|
| Cells or groups (or individuals) | $i = 1, \ldots, N$ |
| Observations within cell | $t = 1, \ldots, T$ |
| Total sample size | $NT$ |
| Within-cell (group) mean | $\bar{y}_i, \bar{\mathbf{x}}_i$ |
| Overall mean | $\bar{y}, \bar{\mathbf{x}}$ |
| Within-group covariance | $W_{yy,i}, W_{yx,i}, W_{xx,i}$ |
| Total variation | $T_{yy}, T_{yx}, T_{xx}$ |

postulated for each cross section:

$$y_{it} = \alpha_t^* + \boldsymbol{\beta}_t' \mathbf{x}_{it} + u_{it}, \quad \begin{array}{l} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{array} \tag{2.2.17}$$

where we again assume that $u_{it}$ is independently normally distributed with mean 0 and constant variance $\sigma_u^2$. Analogous ANCOVA can then be performed to test the homogeneities of the cross-sectional parameters over time. For instance, we can test for overall homogeneity ($H_3' : \alpha_1^* = \alpha_2^* = \cdots = \alpha_T^*, \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_T$) by using the $F$ statistic

$$F_3' = \frac{(S_3 - S_1')/[(T-1)(K+1)]}{S_1'/[NT - T(K+1)]}. \tag{2.2.18}$$

with $(T-1)(K+1)$ and $NT - T(K+1)$ degrees of freedom, where

$$\begin{aligned} S_1' &= \sum_{t=1}^{T} (W_{yy,t} - W_{xy,t}' W_{xx,t}^{-1} W_{xy,t}), \\ W_{yy,t} &= \sum_{i=1}^{N} (y_{it} - \bar{y}_t)^2, \qquad \bar{y}_t = \frac{1}{N} \sum_{i=1}^{N} y_{it}, \\ W_{xx,t} &= \sum_{i=1}^{N} (\mathbf{x}_{it} - \bar{\mathbf{x}}_t)(\mathbf{x}_{it} - \bar{\mathbf{x}}_t)', \qquad \bar{\mathbf{x}}_t = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_{it}, \\ W_{xy,t} &= \sum_{i=1}^{N} (\mathbf{x}_{it} - \bar{\mathbf{x}}_t)(y_{it} - \bar{y}_t). \end{aligned} \tag{2.2.19}$$

Similarly, we can test the hypothesis of heterogeneous intercepts, but homogeneous slopes ($H_1' : \alpha_1^* \neq \alpha_2^* \neq \cdots \neq \alpha_T^*, \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_T$), by using the $F$ statistic

$$F_1' = \frac{(S_2' - S_1')/[(T-1)K]}{S_1'/[NT - T(K+1)]}, \tag{2.2.20}$$

with $(T-1)K$ and $NT - T(K+1)$ degrees of freedom, where

$$S_2' = \sum_{t=1}^{T} W_{yy,t} - \left( \sum_{t=1}^{T} W_{xy,t}' \right) \left( \sum_{t=1}^{T} W_{xx,t} \right)^{-1} \left( \sum_{t=1}^{T} W_{xy,t} \right), \tag{2.2.21}$$

or test the hypothesis of homogeneous intercepts conditional on homogeneous slopes $\beta_1 = \beta_2 = \cdots = \beta_T$ ($H_4'$) by using the $F$ statistic

$$F_4' = \frac{(S_3 - S_2')/(T-1)}{S_2'/[T(N-1) - K]}, \tag{2.2.22}$$

with $(T-1)$ and $T(N-1)-K$ degrees of freedom. In general, unless both cross section and time series ANCOVAs indicate the acceptance of homogeneity of regression coefficients, unconditional pooling (i.e., a single least-squares regression using all observations of cross-sectional units through time) may lead to serious bias.

It should be noted that the foregoing tests are not independent. For example, the uncomfortable possibility exists that according to $F_3$ (or $F_3'$) we might find homogeneous slopes and intercepts, and yet this finding could be compatible with opposite results according to $F_1(F_1')$ and $F_4(F_4')$, because the alternative or null hypotheses are somewhat different in the two cases. Worse still, we might reject the hypothesis of overall homogeneity using the test ratio $F_3(F_3')$, but then find according to $F_1(F_1')$ and $F_4(F_4')$ that we cannot reject the null hypothesis, so that the existence of heterogeneity indicated by $F_3$ (or $F_3'$) cannot be traced. This outcome is quite proper at a formal statistical level, although at the less formal but important level of interpreting test statistics it is an annoyance.

It should also be noted that the validity of the $F$-tests are based on the assumption that the errors of the equation, $u_{it}$, are independently, identically distributed (i.i.d.) and are independent of the regressors, $\mathbf{x}_{it}$ (i.e., the conditional variables, $\mathbf{x}_{it}$, are strictly exogenous (or are fixed constants). In empirical analysis, the errors of the equation could be heteroscedastic or serially correlated, or even correlated with the regressors due to simultaneity or joint dependence. Interpreting $F$-test statistics ignoring these issues could be seriously misleading. Nevertheless, the idea of $F$-tests continue to serve as the basis for developing more robust interference procedures (e.g., the robust standard errors of Stock and Watson 2008). Moreover, given the availability of $F$-test statistics in practically all statistical software packages, it could be considered as a useful first and preliminary step to explore the source of sample variability. We shall discuss some more sophisticated exploratory diagnostic statistics in later chapters when we relax the assumption of "classical" regression model one by one.

## 2.3 AN EXAMPLE

With the aim of suggesting certain modifications to existing theories of investment behavior and providing estimates of the coefficients of principal interest, Kuh (1963) used data on 60 small and middle-sized firms in capital-goods-producing industries from 1935 to 1955, excluding the war years (1942–1945), to probe the proper specification for the investment function. He explored various models based on capacity accelerator behavior or internal funds flows, with various lags. For ease of illustration, we report here only functional specifications and results based on profit theories, capacity-utilization theories, financial restrictions, and long-run growth theories in arithmetic form (Table 2.2, part A), their logarithmic transformations (part B), and several ratio models (part C). The equations are summarized in Table 2.2.

Table 2.2. *Investment equation forms estimated by Kuh (1963)*

*Part A*

$$\Delta I_{it} = \alpha_0 + \beta_1 C_i + \beta_2 \Delta K_{it} + \beta_3 \Delta S_{it} \tag{2.3.1}$$
$$\Delta I_{it} = \alpha_0 + \beta_1 C_i + \beta_2 \Delta K_{it} + \beta_4 \Delta P_{it} \tag{2.3.2}$$
$$\Delta I_{it} = \alpha_0 + \beta_1 C_i + \beta_2 \Delta K_{it} + \beta_3 \Delta S_{it} + \beta_4 \Delta P_{it} \tag{2.3.3}$$
$$I_{it} = \alpha_0 + \beta_1 C_i + \beta_2 K_{it} + \beta_3 S_{it} \tag{2.3.4}$$
$$I_{it} = \alpha_0 + \beta_1 C_i + \beta_2 K_{it} + \beta_4 P_{it} \tag{2.3.5}$$
$$I_{it} = \alpha_0 + \beta_1 C_i + \beta_2 K_{it} + \beta_3 S_{it} + \beta_4 P_{it} \tag{2.3.6}$$
$$I_{it} = \alpha_0 + \beta_1 C_i + \beta_2 K_{it} + \beta_3 S_{i,t-1} \tag{2.3.7}$$
$$I_{it} = \alpha_0 + \beta_1 C_i + \beta_2 K_{it} + \beta_4 P_{i,t-1} \tag{2.3.8}$$
$$I_{it} = \alpha_0 + \beta_1 C_i + \beta_2 K_{it} + \beta_3 S_{i,t-1} + \beta_4 P_{i,t-1} \tag{2.3.9}$$
$$I_{it} = \alpha_0 + \beta_1 C_i + \beta_2 K_{it} + \beta_3 [(S_{it} + S_{i,t-1}) \div 2] \tag{2.3.10}$$
$$I_{it} = \alpha_0 + \beta_1 C_i + \beta_2 K_{it} + \beta_4 [(P_{it} + P_{i,t-1}) \div 2] \tag{2.3.11}$$
$$I_{it} = \alpha_0 + \beta_1 C_i + \beta_2 K_{it} + \beta_3 [(S_{it} + S_{i,t-1}) \div 2] + \beta_4 [(P_{it} + P_{i,t-1}) \div 2] \tag{2.3.12}$$
$$[(I_{it} + I_{i,t-1}) \div 2] = \alpha_0 + \beta_1 C_i + \beta_2 K_{it} + \beta_3 [(S_{it} + S_{i,t-1}) \div 2] \tag{2.3.13}$$
$$[(I_{it} + I_{i,t-1}) \div 2] = \alpha_0 + \beta_1 C_i + \beta_2 K_{it} + \beta_4 [(P_{it} + P_{i,t-1}) \div 2] \tag{2.3.14}$$
$$[(I_{it} + I_{i,t-1}) \div 2] = \alpha_0 + \beta_1 C_i + \beta_2 K_{it} + \beta_3 [(S_{it} + S_{i,t-1}) \div 2] + \beta_4 [(P_{it} + P_{i,t-1}) \div 2] \tag{2.3.15}$$

*Part B*

$$\Delta \log I_{it} = \alpha_0 + \beta_1 \log C_i + \beta_2 \Delta \log K_{it} + \beta_3 \Delta \log S_{it} \tag{2.3.16}$$
$$\log I_{it} = \alpha_0 + \beta_1 \log C_i + \beta_2 \log K_{it} + \beta_3 \log S_{it} \tag{2.3.17}$$
$$\log I_{it} = \alpha_0 + \beta_1 \log C_i + \beta_2 \log K_{it} + \beta_3 \log S_{i,t-1} \tag{2.3.18}$$
$$\log I_{it} = \alpha_0 + \beta_1 \log C_i + \beta_2 \log [(K_{it} + K_{i,t-1}) \div 2] + \beta_3 \log [(S_{it} + S_{i,t-1}) \div 2] \tag{2.3.19}$$

*Part C*

$$\frac{I_{it}}{K_{it}} = \alpha_0 + \beta_1 \frac{P_{it}}{K_{it}} + \beta_2 \frac{S_{i,t-1}}{C_i \cdot K_{i,t-1}} \tag{2.3.20}$$
$$\frac{I_{it}}{K_{it}} = \alpha_0 + \beta_1 \frac{P_{it}}{K_{it}} + \beta_2 \frac{S_{i,t-1}}{C_i \cdot K_{i,t-1}} + \beta_3 \frac{S_{it}}{C_i \cdot K_{it}} \tag{2.3.21}$$
$$\frac{I_{it}}{K_{it}} = \alpha_0 + \beta_1 \frac{P_{it} + P_{i,t-1}}{K_{it} \cdot 2} + \beta_2 \frac{S_{i,t-1}}{C_i \cdot K_{i,t-1}} \tag{2.3.22}$$
$$\frac{I_{it}}{K_{it}} = \alpha_0 + \beta_1 \frac{P_{it} + P_{i,t-1}}{K_{it} \cdot 2} + \beta_2 \frac{S_{i,t-1}}{C_i \cdot K_{i,t-1}} + \beta_3 \frac{S_{it}}{C_i \cdot K_{it}} \tag{2.3.23}$$

*Note:* $I$ = gross investment; $C$ = capital-intensity index; $K$ = capital stock; $S$ = sales; $P$ = gross retained profits.

There were two main reasons that Kuh resorted to using individual-firm data rather than economic aggregates. One was the expressed doubt about the quality of the aggregate data, together with the problems associated with estimating an aggregate time series model when the explanatory variables are highly correlated. The other was the desire to construct and test more complicated behavioral models that require many degrees of freedom. However, as stated in Section 1.2, a single regression using all observations through time

makes sense only when individual observations conditional on the explanatory variables can be viewed as random draws from the same universe. Kuh (1963) used the ANCOVA techniques discussed in Section 2.2. to test for overall homogeneity ($F_3$ or $F_3'$), slope homogeneity ($F_1$ or $F_1'$), and homogeneous intercept conditional on acceptance of homogeneous slopes ($F_4$ or $F_4'$) for both cross-sectional units and time series units.[3] The results for testing homogeneity of time series estimates across cross-sectional units and homogeneity of cross-sectional estimates over time are reproduced in Tables 2.3 and 2.4, respectively.

A striking fact recorded from these statistics is that except for the time series results for equations (2.3.1) and (2.3.3) (which are in first-difference form), all other specifications failed the overall homogeneity tests.[4] Furthermore, in most cases, with the exception of cross-sectional estimates of (2.3.17) and (2.3.18) (Table 2.4), the intercept and slope variabilities cannot be rigorously separated. Nor do the time series results correspond closely to cross-sectional results for the same equation. Although ANCOVA, like other statistics, is not a mill that will grind out results automatically, these results do suggest that the effects of excluded variables in both time series and cross sections may be very different. It would be quite careless not to explore the possible causes of discrepancies that give rise to the systematic interrelationships between different individuals at different periods of time.[5]

Kuh explored the sources of estimation discrepancies through decomposition of the error variances, comparison of individual coefficient behavior, assessment of the statistical influence of various lag structures, and so forth. He concluded that sales seem to include critical time-correlated elements common to a large number of firms and thus have a much greater capability of annihilating systematic, cyclical factors. In general, his results are more favorable to the acceleration sales model than to the internal liquidity/profit hypothesis supported by the results obtained using cross-sectional data (e.g., Meyer and Kuh 1957). He found that the cash flow effect is more important some time before the actual capital outlays are made than it is in actually restricting the outlays during the expenditure period. It appears more appropriate to view internal liquidity flows as a critical part of the budgeting process that later is modified, primarily in light of variations in levels of output and capacity utilization.

The policy implications of Kuh's conclusions are clear. Other things being equal, a small percentage increase in sales will have a greater effect on

[3] See Johnston (1972, Chapter 6) for an illustration of the computation of analysis of covariance.

[4] If the firm differences stay constant over time, heterogeneity among firms can be absorbed into the intercept term. Because intercepts are eliminated by first-differencing, the first-difference model (such as (2.3.1) or (2.3.3)) will be more likely to display homogeneous responses. See Chapter 3 and Chapter 4.

[5] For further discussion of this issue, see Chapter 11, Section 11.3 and Mairesse (1990).

Table 2.3. *Covariance tests for regression-coefficient homogeneity across cross-sectional units*[a]

| | $F_3$ overall test | | | $F_1$ slope homogeneity | | | $F_4$ cell mean significance | | |
|---|---|---|---|---|---|---|---|---|---|
| | Degrees of freedom | | | Degrees of freedom | | | Degrees of freedom | | |
| Equation | Numerator | Denominator | Actual $F$s | Numerator | Denominator | Actual $F$s | Numerator | Denominator | Actual $F$s |
| (2.3.1) | 177 | 660 | 1.25 | 118 | 660 | 1.75[c] | 57 | 660 | 0.12 |
| (2.3.2) | 177 | 660 | 1.40[b] | 118 | 660 | 1.94[c] | 57 | 660 | 0.11 |
| (2.3.3) | 236 | 600 | 1.13 | 177 | 600 | 1.42[b] | 56 | 600 | 0.10 |
| (2.3.4) | 177 | 840 | 2.28[c] | 118 | 840 | 1.58[c] | 57 | 840 | 3.64[c] |
| (2.3.5) | 177 | 840 | 2.34[c] | 118 | 840 | 1.75[c] | 57 | 840 | 3.23[c] |
| (2.3.6) | 236 | 780 | 2.24[c] | 177 | 780 | 1.76[c] | 56 | 780 | 3.57[c] |
| (2.3.7) | 177 | 720 | 2.46[c] | 118 | 720 | 1.95[c] | 57 | 720 | 3.57[c] |
| (2.3.8) | 177 | 720 | 2.50[c] | 118 | 720 | 1.97[c] | 57 | 720 | 3.31[c] |
| (2.3.9) | 236 | 660 | 2.49[c] | 177 | 660 | 2.11[c] | 56 | 660 | 3.69[c] |
| (2.3.10) | 177 | 720 | 2.46[c] | 118 | 720 | 1.75[c] | 57 | 720 | 3.66[c] |
| (2.3.11) | 177 | 720 | 2.60[c] | 118 | 720 | 2.14[c] | 57 | 720 | 3.57[c] |
| (2.3.12) | 236 | 660 | 2.94[c] | 177 | 660 | 2.49[c] | 56 | 660 | 4.18[c] |
| (2.3.16) | 177 | 720 | 1.92[c] | 118 | 720 | 2.59[c] | 57 | 720 | 0.55 |
| (2.3.17) | 177 | 840 | 4.04[c] | 118 | 840 | 2.70[c] | 57 | 840 | 0.39 |
| (2.3.18) | 177 | 720 | 5.45[c] | 118 | 720 | 4.20[c] | 57 | 720 | 6.32[c] |
| (2.3.19) | 177 | 720 | 4.68[c] | 118 | 720 | 3.17[c] | 57 | 720 | 7.36[c] |
| (2.3.20) | 177 | 720 | 3.64[c] | 118 | 720 | 3.14[c] | 57 | 720 | 3.66[c] |
| (2.3.21) | 236 | 660 | 3.38[c] | 177 | 660 | 2.71[c] | 56 | 660 | 4.07[c] |
| (2.3.22) | 177 | 600 | 3.11[c] | 118 | 600 | 2.72[c] | 57 | 600 | 3.22[c] |
| (2.3.23) | 236 | 540 | 2.90[c] | 177 | 540 | 2.40[c] | 56 | 540 | 3.60[c] |

[a] Critical $F$ values were obtained from A.M. Mood, *Introduction to Statistics*, Table V, pp. 426–427. Linear interpolation was employed except for degrees of freedom exceeding 120. The critical $F$ values in every case have been recorded for 120 degrees of freedom for each denominator sum of squares even though the actual degrees of freedom were at least four times as great. The approximation error in this case is negligible.

[b] Significant at the 5 percent level.

[c] Significant at the 1 percent level.

*Source:* Kuh (1963, pp. 141–142).

Table 2.4. *Covariance tests for homogeneity of cross-sectional estimates over time*[a]

| | $F_3'$ overall test | | | $F_1'$ slope homogeneity | | | $F_4'$ cell mean significance | | |
|---|---|---|---|---|---|---|---|---|---|
| | Degrees of freedom | | | Degrees of freedom | | | Degrees of freedom | | |
| Equation | Numerator | Denominator | Actual $F$s | Numerator | Denominator | Actual $F$s | Numerator | Denominator | Actual $F$s |
| (2.3.1) | 52 | 784 | 2.45[b] | 39 | 784 | 2.36[b] | 10 | 784 | 2.89[b] |
| (2.3.2) | 52 | 784 | 3.04[b] | 39 | 784 | 2.64[b] | 10 | 784 | 4.97[b] |
| (2.3.3) | 65 | 770 | 2.55[b] | 52 | 770 | 2.49[b] | 9 | 770 | 3.23[b] |
| (2.3.4) | 64 | 952 | 2.01[b] | 48 | 952 | 1.97[b] | 13 | 952 | 2.43[b] |
| (2.3.5) | 64 | 952 | 2.75[b] | 48 | 952 | 2.45[b] | 13 | 952 | 3.41[b] |
| (2.3.6) | 80 | 935 | 1.91[b] | 64 | 935 | 1.82[b] | 12 | 935 | 2.66[b] |
| (2.3.7) | 56 | 840 | 2.30[b] | 42 | 840 | 2.11[b] | 11 | 840 | 3.66[b] |
| (2.3.8) | 56 | 840 | 2.83[b] | 42 | 840 | 2.75[b] | 11 | 840 | 3.13[b] |
| (2.3.9) | 70 | 825 | 2.25[b] | 56 | 825 | 2.13[b] | 10 | 825 | 3.53[b] |
| (2.3.10) | 56 | 840 | 1.80[b] | 42 | 840 | 1.80[b] | 11 | 840 | 1.72[d] |
| (2.3.11) | 56 | 840 | 2.30[b] | 42 | 840 | 2.30[b] | 11 | 840 | 1.79[d] |
| (2.3.12) | 70 | 825 | 1.70[b] | 56 | 825 | 1.74[b] | 10 | 825 | 1.42 |
| (2.3.13) | 56 | 840 | 2.08[b] | 42 | 840 | 2.11[b] | 11 | 840 | 2.21[c] |
| (2.3.14) | 56 | 840 | 2.66[b] | 42 | 840 | 2.37[b] | 11 | 840 | 2.87[b] |
| (2.3.15) | 70 | 825 | 1.81[b] | 56 | 825 | 1.76[b] | 10 | 825 | 2.35[c] |
| (2.3.16) | 56 | 840 | 3.67[b] | 42 | 840 | 2.85[b] | 11 | 840 | 3.10[b] |
| (2.3.17) | 64 | 952 | 1.51[c] | 48 | 952 | 1.14 | 13 | 952 | 0.80 |
| (2.3.18) | 56 | 840 | 2.34[b] | 42 | 840 | 1.04 | 11 | 840 | 1.99[c] |
| (2.3.19) | 56 | 840 | 2.29[b] | 42 | 840 | 2.03[b] | 11 | 840 | 2.05[c] |
| (2.3.20) | 42 | 855 | 4.13[b] | 28 | 855 | 5.01[b] | 12 | 855 | 2.47[b] |
| (2.3.21) | 56 | 840 | 2.88[b] | 42 | 840 | 3.12[b] | 11 | 840 | 2.56[b] |
| (2.3.22) | 42 | 855 | 3.80[b] | 28 | 855 | 4.62[b] | 12 | 855 | 1.61[b] |
| (2.3.23) | 56 | 840 | 3.51[b] | 42 | 840 | 4.00[b] | 11 | 840 | 1.71[b] |

[a] Critical $F$ values were obtained from A.M. Mood, *Introduction to Statistics*, Table V, pp. 426–427. Linear interpolation was employed except for degrees of freedom exceeding 120. The critical $F$ values in every case have been recorded for 120 degrees of freedom for each denominator sum of squares even though the actual degrees of freedom were at least four times as great. The approximation error in this case is negligible.

[b] Significant at the 1 percent level.

[c] Significant at the 5 percent level.

[d] Significant at the 10 percent level.

*Source:* Kuh (1963, pp. 137–138).

investment than will a small percentage increase in internal funds. If the government seeks to stimulate investment and the objective is magnitude, not qualitative composition, it inexorably follows that the greatest investment effect will come from measures that increase demand rather than from measures that increase internal funds.[6]

[6] For further discussion of investment expenditure behavior, see Chapter 6 or Hsiao and Tahmiscioglu (1997).

CHAPTER 3

# Simple Regression with Variable Intercepts

## 3.1 INTRODUCTION

When the overall homogeneity hypothesis is rejected by the panel data while the specification of a model appears proper, a simple way to take account of the unobserved heterogeneity across individuals and/or through time is to use the variable-intercept models (1.3.1) and (1.3.2). The basic assumption of such models is that, conditional on the observed explanatory variables, the effects of all omitted (or excluded) variables are driven by three types of variables: individual time-invariant, period individual-invariant, and individual time-varying variables.[1] The individual time-invariant variables are variables that are the same for a given cross-sectional unit through time but that vary across cross-sectional units. Examples of these are attributes of individual firm management, ability, sex, and socioeconomic background variables. The period individual-invariant variables are variables that are the same for all cross-sectional units at a given point in time but that vary through time. Examples of these variable are prices, interest rates, and widespread optimism or pessimism. The individual time-varying variables are variables that vary across cross-sectional units at a given point in time and also exhibit variations through time. Examples of these variables are firm profits, sales, and capital stock.

The variable-intercept models assume that the effects of the numerous omitted individual time-varying variables are each individually unimportant but are collectively significant and possess the property of a random variable that is uncorrelated with (or independent of) all other included and excluded variables. On the other hand, because the effects of remaining omitted variables either stay constant through time for a given cross-sectional unit or are the same for all cross-sectional units at a given point in time, or a combination of both, they can be absorbed into the intercept term of a regression model as a means to allow explicitly for the individual and/or time heterogeneity contained in the

[1] These three different sorts of variations apply, of course, to both included and excluded variables. Throughout this monograph we concentrate on relations between excluded variables and included variables.

temporal cross-sectional data. Moreover, when the individual- or time-specific effects are absorbed into the intercept term, there is no need to assume that the individual- or time-specific effects are uncorrelated with **x**, although sometimes they are.

The variable-intercept models can provide a fairly useful specification for fitting regression models using panel data. For example, consider fitting a Cobb–Douglas production function

$$y_{it} = \mu + \beta_1 x_{1it} + \cdots + \beta_K x_{Kit} + v_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{matrix} \tag{3.1.1}$$

where $y$ is the logarithm of output and $x_1, \ldots, x_K$ are the logarithms of respective inputs. The classic procedure is to assume that the effects of omitted variables are independent of **x** and are independently identically distributed. Thus, conditioning on **x** all observations are random variations of a representative firm. However, (3.1.1) has often been criticized for ignoring variables reflecting managerial and other technical differences between firms or variables that reflect general conditions affecting the productivity of all firms but that are fluctuating over time (such as weather factors in agriculture production) (e.g., Hoch 1962; Mundlak 1961; Nerlove 1965). Ideally, such firm- and time-effects variables, say $M_i$ and $P_t$, should be introduced explicitly into (3.1.1). Thus, $v_{it}$ can be written as

$$v_{it} = \alpha M_i + \lambda P_t + u_{it}, \tag{3.1.2}$$

with $u_{it}$ representing the effects of all remaining omitted variables. However, if there are no observations on $M_i$ and $P_t$, it is impossible to estimate $\alpha$ and $\lambda$ directly. A natural alternative would then be to consider the effects of the product, $\alpha_i = \alpha M_i$ and $\lambda_t = \lambda P_t$, which then leads to a variable-intercept model: (1.3.1) or (1.3.2).

Such a procedure was used by Hoch (1962) to estimate parameters of a Cobb–Douglas production function based on annual data for 63 Minnesota farms from 1946 to 1951. He treated output, $y$, as a function of labor, $x_1$; real estate, $x_2$; machinery, $x_3$; and feed, fertilizer, and related expenses, $x_4$. However, because of the difficulties of measuring real estate and machinery variables, he also tried an alternative specification that treated $y$ as a function of $x_1$, $x_4$, a current-expenditures item, $x_5$, and fixed capital, $x_6$. Regression results for both specifications rejected the overall homogeneity hypothesis at the 5 percent significance level. The least-squares estimates under three assumptions ($\alpha_i = \lambda_t = 0$; $\alpha_i = 0$, $\lambda_t \neq 0$; and $\alpha_i \neq 0$, $\lambda_t \neq 0$) are summarized in Table 3.1. They exhibit an increase in the adjusted $R^2$ from 0.75 to about 0.88 when $\alpha_i$ and $\lambda_t$ are introduced. There are also some important changes in parameter estimates when we move from the assumption of identical $\alpha_i$'s to the assumption that both $\alpha_i$ and $\lambda_t$ differ from zero. There is a significant drop in the sum of the elasticities, with the drop concentrated mainly in the labor variable. If one interprets $\alpha_i$ as the firm scale effect, then this indicates that efficiency increases

Table 3.1. *Least-squares estimates of elasticity of Minnesota farm production function based on alternative assumptions*

| Estimate of Elasticity: $\beta_k$ | Assumption | | |
|---|---|---|---|
| | $\alpha_i$ and $\lambda_t$ are identically zero for all $i$ and $t$ | $\alpha_i$ only is identically zero for all $i$ | $\alpha_i$ and $\lambda_t$ different from zero |
| *Variable set 1*[a] | | | |
| $\hat{\beta}_1$, labor | 0.256 | 0.166 | 0.043 |
| $\hat{\beta}_2$, real estate | 0.135 | 0.230 | 0.199 |
| $\hat{\beta}_3$, machinery | 0.163 | 0.261 | 0.194 |
| $\hat{\beta}_4$, feed & fertilizer | 0.349 | 0.311 | 0.289 |
| Sum of $\hat{\beta}$'s | 0.904 | 0.967 | 0.726 |
| Adjusted $R^2$ | 0.721 | 0.813 | 0.884 |
| *Variable set 2* | | | |
| $\hat{\beta}_1$, labor | 0.241 | 0.218 | 0.057 |
| $\hat{\beta}_5$, current expenses | 0.121 | 0.185 | 0.170 |
| $\hat{\beta}_6$, fixed capital | 0.278 | 0.304 | 0.317 |
| $\hat{\beta}_4$, feed & fertilizer | 0.315 | 0.285 | 0.288 |
| Sum of $\hat{\beta}$'s | 0.954 | 0.991 | 0.832 |
| Adjusted $R^2$ | 0.752 | 0.823 | 0.879 |

[a] All output and input variables are in service units, measured in dollars.
*Source:* Hoch (1962).

with scale. As demonstrated in Figure 1.1, when the production hyperplane of larger firms lies above the average production plane and the production plane of smaller firm below the average plane, the pooled estimates, neglecting firm differences, will have greater slope than the average plane. Some confirmation of this argument was provided by Hoch (1962). Table 3.2 lists the characteristics of firms grouped on the basis of firm-specific effects $\alpha_i$. The table suggests a fairly pronounced association between scale and efficiency.

This example demonstrates that by introducing the unit- and/or time-specific variables into the specification for panel data, it is possible to reduce or avoid the omitted-variable bias. In this chapter we focus on the estimation and hypothesis testing of models (1.3.1) and (1.3.2) under the assumption that all explanatory variables, $\mathbf{x}_{kit}$, are nonstochastic (or exogenous). For ease of seeing the relations between fixed and random effects inference, we shall assume there are no time-specific effects in Sections 3.2–3.5. In Section 3.2 we discuss estimation methods when the specific effects are treated as fixed constants (FE). Section 3.3 discusses estimation methods when they are treated as random variables (effects) (RE). Section 3.4 discusses the pros and cons of treating the specific effects as fixed or random. Tests for misspecification are discussed in Section 3.5. Section 3.6 discusses models with both individual- and time-specific effects and models with specific variables. Section 3.7 discusses

Table 3.2. *Characteristics of firms grouped on the basis of the firm constant*

| | | Firms classified by value of $\exp(\alpha_i)$[a] | | | | |
|---|---|---|---|---|---|---|
| Characteristics | All firms | <0.85 | 0.85–0.95 | 0.95–1.05 | 1.05–1.15 | >1.15 |
| Numbers of firms in group | 63 | 6 | 17 | 19 | 14 | 7 |
| Average value of: | | | | | | |
| $e^{\alpha_i}$, firm constant | 1.00 | 0.81 | 0.92 | 1.00 | 1.11 | 1.26 |
| Output (dollars) | 15,602 | 10,000 | 15,570 | 14,690 | 16,500 | 24,140 |
| Labor (dollars) | 3,468 | 2,662 | 3,570 | 3,346 | 3,538 | 4,280 |
| Feed & fertilizer (dollars) | 3,217 | 2,457 | 3,681 | 3,064 | 2,621 | 5,014 |
| Current expenses (dollars) | 2,425 | 1,538 | 2,704 | 2,359 | 2,533 | 2,715 |
| Fixed capital (dollars) | 3,398 | 2,852 | 3,712 | 3,067 | 3,484 | 3,996 |
| Profit (dollars) | 3,094 | 491 | 1,903 | 2,854 | 4,324 | 8,135 |
| Profit/output | 0.20 | 0.05 | 0.12 | 0.19 | 0.26 | 0.33 |

[a] The mean of firm effects, $\alpha_i$, is zero is invoked.
*Source:* Hoch (1962).

heteroscedasticity and autocorrelation adjustment. In Section 3.8 we use a multivariate setup of a single-equation model to provide a synthesis of the issues involved and to provide a link between the single equation model and the linear simultaneous equations model (see Chapter 5).

## 3.2 FIXED-EFFECTS MODELS: LEAST-SQUARES DUMMY VARIABLE APPROACH

The obvious generalization of the constant-intercept-and-slope model for panel data is to introduce dummy variables to account for the effects of those omitted variables that are specific to individual cross-sectional units but stay constant over time, and the effects that are specific to each time period but are the same for all cross-sectional units. For ease of highlighting the difference between the FE and RE specifications in this section and the next three sections we assume no time-specific effects and focus only on individual-specific effects. Thus, the value of the dependent variable for the $i$th unit at time $t$, $y_{it}$, depends on $K$ exogenous variables, $(x_{1it}, \ldots, x_{Kit}) = \mathbf{x}'_{it}$, that differ among individuals in a cross section at a given point in time and also exhibit variation through time, as well as on variables that are specific to the $i$th unit and that stay (more or less) constant over time. This is model (1.3.1), which we can rewrite as

$$y_{it} = \alpha_i^* + \underset{1\times K}{\mathbf{x}'_{it}} \quad \underset{K\times 1}{\boldsymbol{\beta}} + u_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{matrix} \tag{3.2.1}$$

where $\boldsymbol{\beta}$ is a $K \times 1$ vector of constants and $\alpha_i^*$ is a $1 \times 1$ scalar constant representing the effects of those variables peculiar to the $i$th individual in more

or less the same fashion over time. The error term, $u_{it}$, represents the effects of the omitted variables that are peculiar to both the individual units and time periods. We assume that $u_{it}$ is uncorrelated with $(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$ and can be characterized by an independently identically distributed random variable with mean 0 and variance $\sigma_u^2$.

The model (3.2.1) is also called the ANCOVA model. Without attempting to make the boundaries between regression analysis, ANOVA, and ANCOVA precise, we can say that regression model assumes that the expected value of $y$ is a function of exogenous factors, $\mathbf{x}$, while the conventional ANOVA model stipulates that the expected value of $y_{it}$ depends only on the class, $i$, to which the observation considered belongs and that the value of the measured quantity, $y$, assumes the relation that $y_{it} = \alpha_i^* + u_{it}$, where the effects of all other characteristics, $u_{it}$, are random and are in no way dependent on the individual-specific effects, $\alpha_i^*$. But if $y$ is also affected by other variables that we are not able to control and standardize within classes, the simple within-class sum of squares will be an overestimate of the stochastic component in $y$, and the differences between class means will reflect not only any class effect but also the effects of any differences in the values assumed by the uncontrolled variables in different classes. It was for this kind of problem that the ANCOVA model of the form (3.2.1) was first developed. The models are of a mixed character, involving genuine exogenous variables, $\mathbf{x}_{it}$, as do regression models, and at the same time allowing the true relation for each individual to depend on the class to which the individual belongs, $\alpha_i^*$, as do the usual ANOVA models. The regression model enables us to assess the effects of quantitative factors and the ANOVA model those of qualitative factors; the ANCOVA model covers both quantitative and qualitative factors.

Stacking all $NT$ observations of $y_{it}$ ((3.2.1)) in vector form, we have

$$Y = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{bmatrix} = \begin{bmatrix} \mathbf{e} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} \alpha_1^* + \begin{bmatrix} \mathbf{0} \\ \mathbf{e} \\ \vdots \\ \mathbf{0} \end{bmatrix} \alpha_2^* + \cdots + \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{e} \end{bmatrix} \alpha_N^* + \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_N \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_N \end{bmatrix}, \tag{3.2.2}$$

where

$$\underset{T\times 1}{\mathbf{y}_i} = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{bmatrix}, \qquad \underset{T\times K}{\mathbf{X}_i} = \begin{bmatrix} x_{1i1} & x_{2i1} & \cdots & x_{Ki1} \\ x_{1i2} & x_{2i2} & \cdots & x_{Ki2} \\ \vdots & \vdots & & \vdots \\ x_{1iT} & x_{2iT} & & x_{KiT} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_N \end{bmatrix}$$

$$\underset{1\times T}{\mathbf{e}'} = (1, 1, \ldots, 1), \qquad \underset{1\times T}{\mathbf{u}_i'} = (u_{i1}, \ldots, u_{iT}),$$

$$E\mathbf{u}_i = \mathbf{0}, \qquad E\mathbf{u}_i\mathbf{u}_i' = \sigma_u^2 I_T, \qquad E\mathbf{u}_i\mathbf{u}_j' = \mathbf{0} \quad \text{if } i \neq j,$$

$I_T$ denotes the $T \times T$ identity matrix. Let $\tilde{X} = (\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_N, X)$, where $\mathbf{d}_i$ is an $NT \times 1$ vector dummy variable with the first $(i-1) \times T$ elements equal to 0, $(i-1)T + 1$ to $iT$ elements equal to 1, and 0 from $iT + 1, \ldots, NT$, $i = 1, \ldots, N$. Then $\mathbf{y} = \tilde{X}\boldsymbol{\theta} + \mathbf{u}$, where $\boldsymbol{\theta} = (\alpha_1^*, \ldots, \alpha_N^*, \boldsymbol{\beta}')'$.

Given the assumed properties of $u_{it}$, we know that the ordinary least-squares (OLS) estimator of (3.2.2) is the best linear unbiased estimator (BLUE). The OLS estimators of $\alpha_i^*$ and $\boldsymbol{\beta}$ are obtained by minimizing

$$S = (\mathbf{y} - \tilde{X}\boldsymbol{\theta})'(\mathbf{y} - \tilde{X}\boldsymbol{\theta}) = \sum_{i=1}^{N} \mathbf{u}_i'\mathbf{u}_i$$

$$= \sum_{i=1}^{N} (\mathbf{y}_i - \mathbf{e}\alpha_i^* - X_i\boldsymbol{\beta})'(\mathbf{y}_i - \mathbf{e}\alpha_i^* - X_i\boldsymbol{\beta}). \tag{3.2.3}$$

Taking partial derivatives of $S$ with respect to $\alpha_i^*$ and setting them equal to 0, we have

$$\hat{\alpha}_i^* = \bar{y}_i - \bar{\mathbf{x}}_i'\boldsymbol{\beta}, \quad i = 1, \ldots, N, \tag{3.2.4}$$

where

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}, \quad \bar{\mathbf{x}}_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{it}.$$

Substituting (3.2.4) into (3.2.3) and taking the partial derivative of $S$ with respect to $\boldsymbol{\beta}$, we have[2]

$$\hat{\boldsymbol{\beta}}_{cv} = \left[\sum_{i=1}^{N}\sum_{t=1}^{T} (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\right]^{-1} \left[\sum_{i=1}^{N}\sum_{t=1}^{T} (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(y_{it} - \bar{y}_i)\right]. \tag{3.2.5}$$

The OLS estimator (3.2.5) is called the least-squares dummy variable (LSDV) estimator because the observed values to the coefficients $\alpha_i^*$ takes the form of dummy variables. However, the computational procedure for estimating the slope parameters in this model does not require the dummy variables for the individual (and/or time) effects actually be included in the matrix of explanatory variables. We need only find the means of time series observations separately for each cross-sectional unit, transform the observed variables by subtracting out the appropriate time series means, and then apply the least-squares method to the transformed data. Hence, we need only invert a matrix of order $K \times K$.

[2] Although the notations are different, (3.2.5) is identical with (2.2.10).

The foregoing procedure is equivalent to premultiplying the $i$th equation

$$\mathbf{y}_i = \mathbf{e}\alpha_i^* + X_i\boldsymbol{\beta} + \mathbf{u}_i$$

by a $T \times T$ idempotent (covariance) transformation matrix

$$Q = I_T - \frac{1}{T}\mathbf{e}\mathbf{e}' \tag{3.2.6}$$

to "sweep out" the individual effect $\alpha_i^*$ so that individual observations are measured as deviations from individual means (over time):

$$\begin{aligned} Q\mathbf{y}_i &= Q\mathbf{e}\alpha_i^* + QX_i\boldsymbol{\beta} + Q\mathbf{u}_i \\ &= QX_i\boldsymbol{\beta} + Q\mathbf{u}_i, \quad i = 1, \ldots, N. \end{aligned} \tag{3.2.7}$$

Applying the OLS procedure to (3.2.7) we have[3]

$$\hat{\boldsymbol{\beta}}_{cv} = \left[\sum_{i=1}^{N} X_i' Q X_i\right]^{-1} \left[\sum_{i=1}^{N} X_i' Q \mathbf{y}_i\right], \tag{3.2.8}$$

which is identical to (3.2.5). Because (3.2.2) is called the ANCOVA model, the LSDV estimator of $\boldsymbol{\beta}$ is sometimes called the covariance (CV) estimator. It is also called the within-group estimator, because only the variation within each group is utilized in forming this estimator.[4]

The CV estimator of $\boldsymbol{\beta}$ can also be derived as a method of moment estimator. The strict exogeneity of $\mathbf{x}_{it}$ implies that

$$E(\mathbf{u}_i \mid X_i, \alpha_i^*) = E(\mathbf{u}_i \mid X_i) = \mathbf{0}. \tag{3.2.9}$$

It follows that

$$E[(\mathbf{u}_i - \mathbf{e}\bar{u}_i) = (\mathbf{y}_i - \mathbf{e}\bar{y}_i) - (X_i - \mathbf{e}\bar{\mathbf{x}}_i')\boldsymbol{\beta} \mid X_i] = \mathbf{0}. \tag{3.2.10}$$

[3] Equation (3.2.7) can be viewed as a linear-regression model with singular-disturbance covariance matrix $\sigma_u^2 Q$. A generalization of Aitken's theorem leads to the generalized least-squares estimator

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{cv} &= \left(\sum_{i=1}^{N} X_i' Q' Q^- Q X_i\right)^{-1} \left(\sum_{i=1}^{N} X_i' Q' Q^- Q \mathbf{y}_i\right) \\ &= \left[\sum_{i=1}^{N} X_i' Q X_i\right]^{-1} \left[\sum_{i=1}^{N} X_t' Q \mathbf{y}_i\right], \end{aligned}$$

where $Q^-$ is the generalized inverse of $Q$ satisfying the conditions $QQ^-Q = Q$ (Theil (1971), their Sections 6.6 and 6.7).

[4] Because the slope coefficients are assumed the same for all $i$ and $t$, for simplicity we shall not distinguish the individual mean corrected estimator and the within-group estimator as we did in Chapter 2. We shall simply refer to (3.2.8) or its equivalent as the within-group estimator.

Approximating the moment conditions (3.2.10) by their sample moments yields

$$\frac{1}{N}\sum_{i=1}^{N} X_i'[(\mathbf{y}_i - \mathbf{e}\bar{y}_i) - (X_i - \mathbf{e}\bar{\mathbf{x}}_i')\hat{\boldsymbol{\beta}}] = \frac{1}{N}\sum_{i=1}^{N} X_i'[Q\mathbf{y}_i - QX_i\hat{\boldsymbol{\beta}}] = \mathbf{0}. \tag{3.2.10′}$$

Solving (3.2.10′) yields the CV estimator (3.2.8).

The CV estimator $\hat{\beta}_{cv}$ is unbiased. It is also consistent when either $N$ or $T$ or both tend to infinity. Its variance–covariance matrix is

$$\text{Var}\,(\hat{\boldsymbol{\beta}}_{cv}) = \sigma_u^2\left[\sum_{t=1}^{N} X_i' Q X_i\right]^{-1}. \tag{3.2.11}$$

However, the estimator for the intercept, (3.2.4), although unbiased, is consistent only when $T \to \infty$.

It should be noted that an alternative and equivalent formulation of (3.2.1) is to introduce a "mean intercept," $\mu$, so that

$$y_{it} = \mu + \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + u_{it}. \tag{3.2.12}$$

Because both $\mu$ and $\alpha_i$ are fixed constants, without additional restriction, they are not separately identifiable or estimable. One way to identify $\mu$ and $\alpha_i$ is to introduce the restriction that $\sum_{i=1}^{N}\alpha_i = 0$. Then the individual effect $\alpha_i$ represents the deviation of the $i$th individual from the common mean $\mu$.

Equations (3.2.12) and (3.2.1) lead to the same least-squares estimator for $\boldsymbol{\beta}$ [equation (3.2.5)]. This easily can be seen by noting that the BLUEs for $\mu, \alpha_i$, and $\boldsymbol{\beta}$ are obtained by minimizing

$$\sum_{i=1}^{N}\mathbf{u}_i'\mathbf{u}_i = \sum_{i=1}^{N}\sum_{t=1}^{T} u_{it}^2$$

subject to the restriction $\sum_{i=1}^{N}\alpha_i = 0$. Utilizing the restriction $\sum_{i=1}^{N}\alpha_i = 0$ in solving the marginal conditions, we have

$$\hat{\mu} = \bar{y} - \bar{\mathbf{x}}'\boldsymbol{\beta}, \quad \text{where } \bar{y} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}, \quad \bar{\mathbf{x}} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{x}_{it}, \tag{3.2.13}$$

$$\hat{\alpha}_i = \bar{y}_i - \hat{\mu} - \bar{\mathbf{x}}_i'\boldsymbol{\beta}. \tag{3.2.14}$$

Substituting (3.2.13) and (3.2.14) into (3.2.12) and solving the marginal condition for $\boldsymbol{\beta}$, we obtain (3.2.5).

When var $(u_{it}) = \sigma_i^2$, the LSDV estimator is no longer BLUE. However, it remains consistent. An efficient estimator is to apply the weighted least-squares estimator where each $(y_{it}, \mathbf{x}_{it}', 1)$ is weighted by the inverse of $\sigma_i$ before applying the LSDV estimator. An initial estimator of $\sigma_i$ can be obtained from

$$\hat{\sigma}_i^2 = \frac{1}{T}\sum_{t=1}^{T}(y_{it} - \hat{\alpha}_i^* - \mathbf{x}_{it}'\hat{\boldsymbol{\beta}}_{cv})^2. \tag{3.2.15}$$

## 3.3 RANDOM EFFECTS MODELS: ESTIMATION OF VARIANCE-COMPONENTS MODELS

In Section 3.2 we discussed the estimation of linear regression models when the effects of omitted individual-specific variables $(\alpha_i)$ are treated as fixed constants over time. In this section we treat the individual-specific effects, $\alpha_i$, like $u_{it}$, as random variables.

It is a standard practice in the regression analysis to assume that the large number of factors that affect the value of the dependent variable, but that have not been explicitly included as explanatory variables, can be appropriately summarized by a random disturbance. When numerous individual units are observed over time, it is sometimes assumed that some of the omitted variables will represent factors peculiar to both the individual units and time periods for which observations are obtained, whereas other variables will reflect individual differences that tend to affect the observations for a given individual in more or less the same fashion over time. Still other variables may reflect factors peculiar to specific time periods, but affecting individual units more or less equally. Thus, the residual, $v_{it}$, is often assumed to consist of three components:[5]

$$v_{it} = \alpha_i + \lambda_t + u_{it}, \tag{3.3.1}$$

However, the sample provides information only about the joint density of $(y_{it}, \mathbf{x}_{it}')$, $f(\mathbf{y}_i, \mathbf{x}_i)$, not the joint density of $f(y_i, \mathbf{x}_i, \alpha_i, \boldsymbol{\lambda})$, where $\mathbf{x}_i$ denotes the $TK \times 1$ observed $\mathbf{x}_{it}$, and $\boldsymbol{\lambda}$ denotes the $T \times 1$ vector $(\lambda_1, \ldots, \lambda_T)$. Since

$$\begin{aligned} f(\mathbf{y}_i, \mathbf{x}_i) &= f(\mathbf{y}_i \mid \mathbf{x}_i) f(\mathbf{x}_i) \\ &= \left[\int f(\mathbf{y}_i \mid \mathbf{x}_i, \alpha_i, \boldsymbol{\lambda}) f(\alpha_i, \boldsymbol{\lambda} \mid \mathbf{x}_i) d\alpha_i d\boldsymbol{\lambda}\right] \cdot f(\mathbf{x}_i), \end{aligned} \tag{3.3.2}$$

we need to know $f(\alpha_i, \boldsymbol{\lambda}_t \mid \mathbf{x}_i)$ to derive the random-effects estimator. However, $\alpha_i$ and $\lambda_t$ are unobserved. A common assumption for the random-effects model

[5] Note that we follow the formulation of (3.2.10) by treating $\alpha_i$ and $\lambda_t$ as deviations from the population mean. For ease of exposition we also restrict our attention to the homoscedastic variances of $\alpha_i$ and $\lambda_t$. For the heteroscedasticity generalization of the error-component model, see Chapter 3, Section 3.7 or Mazodier and Trognon (1978) and Wansbeek and Kapteyn (1982). For a test of individual heteroscedasticity, see Holly and Gardiol (2000).

is to assume

$$f(\alpha_i, \lambda_t \mid \mathbf{x}_i) = f(\alpha_i, \lambda_t) = f(\alpha_i) f(\lambda_t). \tag{3.3.3}$$

In other words, we assume that

$$\begin{aligned} E\alpha_i = E\lambda_t = Eu_{it} = 0, \quad E\alpha_i\lambda_t = E\alpha_i u_{it} = E\lambda_t u_{it} = 0, \\ E\alpha_i\alpha_j = \begin{cases} \sigma_\alpha^2 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases} \\ E\lambda_t\lambda_s = \begin{cases} \sigma_\lambda^2 & \text{if } t = s, \\ 0 & \text{if } t \neq s, \end{cases} \\ Eu_{it}u_{js} = \begin{cases} \sigma_u^2 & \text{if } i = j,\ t = s, \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{3.3.4}$$

and

$$E\alpha_i \mathbf{x}_{it}' = E\lambda_t \mathbf{x}_{it}' = Eu_{it}\mathbf{x}_{it}' = \mathbf{0}'.$$

The variance of $y_{it}$ conditional on $\mathbf{x}_{it}$ is, from (3.3.1) and (3.3.4), $\sigma_y^2 = \sigma_\alpha^2 + \sigma_\lambda^2 + \sigma_u^2$. The variances $\sigma_\alpha^2, \sigma_\lambda^2$, and $\sigma_u^2$ are accordingly called variance components; each is a variance in its own right and is a component of $\sigma_y^2$. Therefore, this kind of model is sometimes referred to as a variance-components (or error-components) model.

For ease of exposition we assume $\lambda_t = 0$ for all $t$ in this and the following three sections. That is, we concentrate on models of the form (3.2.12).

Rewriting (3.2.12) in vector form, we have

$$\underset{T\times 1}{\mathbf{y}_i} = \underset{T\times(K+1)}{\tilde{X}_i}\ \underset{(K+1)\times 1}{\boldsymbol{\delta}} + \underset{T\times 1}{\mathbf{v}_i}, \quad i = 1, 2, \ldots, N, \tag{3.3.5}$$

where $\tilde{X}_i = (\mathbf{e}, X_i)$, $\boldsymbol{\delta}' = (\mu, \boldsymbol{\beta}')$, $\mathbf{v}_i' = (v_{i1}, \ldots, v_{iT})$, and $v_{it} = \alpha_i + u_{it}$. The presence of $\alpha_i$ creates correlations of $v_{it}$ over time for a given individual, although $v_{it}$ remains uncorrelated across individuals. The variance–covariance matrix of $\mathbf{v}_i$ takes the form,

$$E\mathbf{v}_i\mathbf{v}_i' = \sigma_u^2 I_T + \sigma_\alpha^2 \mathbf{e}\mathbf{e}' = V. \tag{3.3.6}$$

Its inverse is (see Graybill 1969; Nerlove 1971b; Wallace and Hussain 1969)

$$V^{-1} = \frac{1}{\sigma_u^2}\left[I_T - \frac{\sigma_\alpha^2}{\sigma_u^2 + T\sigma_\alpha^2}\mathbf{e}\mathbf{e}'\right]. \tag{3.3.7}$$

### 3.3.1 Covariance Estimation

Regardless of whether the $\alpha_i$'s are treated as fixed or as random, the individual-specific effects for a given sample can be swept out by the idempotent (covariance) transformation matrix $Q$ [equation (3.2.6)], because $Q\mathbf{e} = \mathbf{0}$, and hence

$Q\mathbf{v}_i = Q\mathbf{u}_i$. Thus, premultiplying (3.3.5) by $Q$, we have

$$\begin{aligned} Q\mathbf{y}_i &= Q\mathbf{e}\mu + QX_i\boldsymbol{\beta} + Q\mathbf{e}\alpha_i + Q\mathbf{u}_i \\ &= QX_i\boldsymbol{\beta} + Q\mathbf{u}_i. \end{aligned} \tag{3.3.8}$$

Applying the least-squares method to (3.3.8), we obtain the CV estimator (3.2.8) of $\boldsymbol{\beta}$. We estimate $\mu$ by $\hat{\mu} = \bar{y} - \bar{\mathbf{x}}'\hat{\boldsymbol{\beta}}_{cv}$.

Whether $\alpha_i$ are treated as fixed or random, the CV estimator of $\boldsymbol{\beta}$ is unbiased and consistent either $N$ or $T$ or both tend to infinity. However, whereas the CV estimator is the BLUE under the assumption that $\alpha_i$ are fixed constants, the CV estimator is not the BLUE in finite samples when $\alpha_i$ are assumed random. The BLUE in the latter case is the generalized least-squares (GLS) estimator.[6] Moreover, if the explanatory variables contain some time-invariant variables, $\mathbf{z}_i$, then $\mathbf{e}\mathbf{z}_i'$ and $\mathbf{e}$ are perfectly correlated. Their coefficients cannot be estimated by CV because the CV transformation eliminates $\mathbf{z}_i$ from (3.3.8).

### 3.3.2 Generalized Least-Squares (GLS) Estimation

Under (3.3.4), $E(\mathbf{v}_i \mid \mathbf{x}_i) = 0$. The least-squares method can be applied. However, because $v_{it}$ and $v_{is}$ both contain $\alpha_i$, the residuals of (3.3.5) are serially correlated. To get efficient estimates of $\boldsymbol{\delta}' = (\mu, \boldsymbol{\beta}')$ we have to use the GLS method. The normal equations for the GLS estimators are

$$\left[\sum_{i=1}^{N} \tilde{X}_i' V^{-1} \tilde{X}_i\right] \hat{\boldsymbol{\delta}}_{\text{GLS}} = \left[\sum_{i=1}^{N} \tilde{X}_i' V^{-1} \mathbf{y}_i\right]. \tag{3.3.9}$$

Following Maddala (1971a), we write $V^{-1}$ [equation (3.3.7)] as

$$V^{-1} = \frac{1}{\sigma_u^2}\left[\left(I_T - \frac{1}{T}\mathbf{e}\mathbf{e}'\right) + \psi \cdot \frac{1}{T}\mathbf{e}\mathbf{e}'\right] = \frac{1}{\sigma_u^2}\left[Q + \psi \cdot \frac{1}{T}\mathbf{e}\mathbf{e}'\right], \tag{3.3.10}$$

where

$$\psi = \frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}. \tag{3.3.11}$$

Hence, (3.3.9) can conveniently be written as

$$[W_{\tilde{x}\tilde{x}} + \psi B_{\tilde{x}\tilde{x}}] \begin{bmatrix} \hat{\mu} \\ \hat{\boldsymbol{\beta}} \end{bmatrix}_{\text{GLS}} = [W_{\tilde{x}y} + \psi B_{\tilde{x}y}], \tag{3.3.12}$$

[6] For details, see Section 3.3.2.

where

$$T_{\tilde{x}\tilde{x}} = \sum_{i=1}^{N} \tilde{X}_i' \tilde{X}_i, \qquad T_{\tilde{x}y} = \sum_{i=1}^{N} \tilde{X}_i' \mathbf{y}_i,$$

$$B_{\tilde{x}\tilde{x}} = \frac{1}{T} \sum_{i=1}^{N} (\tilde{X}_i' \mathbf{e}\mathbf{e}' \tilde{X}_i), \qquad B_{\tilde{x}y} = \frac{1}{T} \sum_{i=1}^{N} (\tilde{X}_i' \mathbf{e}\mathbf{e}' y_i),$$

$$W_{\tilde{x}\tilde{x}} = T_{\tilde{x}\tilde{x}} - B_{\tilde{x}\tilde{x}}, \qquad W_{\tilde{x}y} = T_{\tilde{x}y} - B_{\tilde{x}y}.$$

The matrices $B_{\tilde{x}\tilde{x}}$ and $B_{\tilde{x}y}$ contain the sums of squares and sums of cross products between groups, $W_{\tilde{x}\tilde{x}}$ and $W_{\tilde{x}y}$ are the corresponding matrices within groups, and $T_{\tilde{x}\tilde{x}}$ and $T_{\tilde{x}y}$ are the corresponding matrices for total variation.

Solving (3.3.12), we have

$$\begin{bmatrix} \psi NT & \psi T \sum_{i=1}^{N} \bar{\mathbf{x}}_i' \\ \psi T \sum_{i=1}^{N} \bar{\mathbf{x}}_i & \sum_{i=1}^{N} X_i' Q X_i + \psi T \sum_{i=1}^{N} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i' \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\boldsymbol{\beta}} \end{bmatrix}_{\text{GLS}} = \begin{bmatrix} \psi NT \bar{y} \\ \sum_{i=1}^{N} X_i' Q \mathbf{y}_i + \psi T \sum_{i=1}^{N} \bar{\mathbf{x}}_i \bar{y}_i \end{bmatrix}. \tag{3.3.13}$$

Using the formula of the partitioned inverse, we obtain

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{GLS}} &= \left[ \frac{1}{T} \sum_{i=1}^{N} X_i' Q X_i + \psi \sum_{i=1}^{N} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})' \right]^{-1} \\ &\quad \cdot \left[ \frac{1}{T} \sum_{i=1}^{N} X_i' Q \mathbf{y}_i + \psi \sum_{i=1}^{N} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{y}_i - \bar{y}) \right] \\ &= \Delta \hat{\boldsymbol{\beta}}_b + (I_K - \Delta) \hat{\boldsymbol{\beta}}_{cv}, \\ \hat{\mu}_{\text{GLS}} &= \bar{y} - \bar{\mathbf{x}}' \hat{\boldsymbol{\beta}}_{\text{GLS}}, \end{aligned} \tag{3.3.14}$$

where

$$\begin{aligned} \Delta &= \psi T \left[ \sum_{i=1}^{N} X_i' Q X_i + \psi T \sum_{i=1}^{N} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})' \right]^{-1} \\ &\quad \cdot \left[ \sum_{i=1}^{N} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})' \right], \\ \hat{\boldsymbol{\beta}}_b &= \left[ \sum_{i=1}^{N} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})' \right]^{-1} \left[ \sum_{i=1}^{N} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{y}_i - \bar{y}) \right]. \end{aligned}$$

The estimator $\hat{\boldsymbol{\beta}}_b$ is called the between-group estimator because it ignores variation within the group.

The GLS estimator (3.3.14) is a weighted average of the between-group and within-group estimators. If $\psi \to 1$, $\hat{\tilde{\boldsymbol{\delta}}}_{\text{GLS}}$ converges to the OLS estimator $T_{\tilde{x}\tilde{x}}^{-1} T_{\tilde{x}y}$. If $\psi \to 0$, the GLS estimator for $\boldsymbol{\beta}$ becomes the CV estimator (LSDV) [equation (3.2.5)]. In essence, $\psi$ measures the weight given to the between-group variation. In the LSDV (or fixed-effects model) procedure, this source of variation is completely ignored. The OLS procedure corresponds to $\psi = 1$. The between-group and within-group variations are just added up. Thus, one can view the OLS and LSDV as somewhat all-or-nothing ways of utilizing the between-group variation. The procedure of treating $\alpha_i$ as random provides a solution intermediate between treating them all as different and treating them all as equal, as implied by the GLS estimator given in (3.3.14).

If $[W_{\tilde{x}\tilde{x}} + \psi B_{\tilde{x}\tilde{x}}]$ is nonsingular, the covariance matrix of GLS estimators of $\boldsymbol{\delta}$ can be written as

$$\text{Var} \begin{bmatrix} \hat{\mu} \\ \hat{\boldsymbol{\beta}} \end{bmatrix}_{\text{GLS}} = \sigma_u^2 [W_{\tilde{x}\tilde{x}} + \psi B_{\tilde{x}\tilde{x}}]^{-1} \tag{3.3.15}$$

$$= \sigma_u^2 \left[ \begin{pmatrix} 0 & \mathbf{0}' \\ \mathbf{0} & \sum_{i=1}^{N} X_i' Q X_i \end{pmatrix} + T\psi \begin{pmatrix} N & \sum_{i=1}^{N} \bar{\mathbf{x}}_i' \\ \sum_{i=1}^{N} \bar{\mathbf{x}}_i & \sum_{i=1}^{N} \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i' \end{pmatrix} \right]^{-1}.$$

Using the formula for partitioned inversion (e.g., Rao 1973, Chapter 2; Theil 1971, Chapter 1), we obtain

$$\text{Var}\,(\hat{\boldsymbol{\beta}}_{\text{GLS}}) = \sigma_u^2 \left[ \sum_{i=1}^{N} X_i' Q X_i + T\psi \sum_{i=1}^{N} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})' \right]^{-1}. \tag{3.3.16}$$

Because $\psi > 0$, we see immediately that the difference between the covariance matrices of $\hat{\boldsymbol{\beta}}_{cv}$ and $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ is a positive semidefinite matrix. However, for fixed $N$, as $T \to \infty$, $\psi \to 0$. Thus, under the assumption that $(1/NT)\Sigma_{i=1}^{N} X_i' X_i$ and $(1/NT)\Sigma_{i=1}^{N} X_i' Q X_i$, converge to finite positive definitive matrices, when $T \to \infty$, we have $\hat{\beta}_{\text{GLS}} \to \hat{\beta}_{cv}$ and $\text{Var}(\sqrt{T}\hat{\boldsymbol{\beta}}_{\text{GLS}}) \to$ Var $(\sqrt{T}\hat{\boldsymbol{\beta}}_{cv})$. This is because when $T \to \infty$, we have an infinite number of observations for each $i$. Therefore, we can consider each $\alpha_i$ as a random variable that has been drawn once and forever so that for each $i$ we can pretend that they are just like fixed parameters.

Computation of the GLS estimator can be simplified by noting the special form of $V^{-1}$ (3.3.10). Let $P = [I_T - (1 - \psi^{1/2})(1/T)\mathbf{ee}']$; we have $V^{-1} = \frac{1}{\sigma_u^2} P'P$. Premultiplying (3.3.5) by the transformation matrix, $P$, we obtain the GLS estimator (3.3.12) by applying the least-squares method to the transformed model (Theil 1971, Chapter 6). This is equivalent to first transforming the data by subtracting a fraction $(1 - \psi^{1/2})$ of individual means $\bar{y}_i$, and $\bar{\mathbf{x}}_i$ from their corresponding $y_{it}$ and $\mathbf{x}_{it}$, then regressing $[y_{it} - (1 - \psi^{1/2})\bar{y}_i]$ on a constant and $[\mathbf{x}_{it} - (1 - \psi^{1/2})\bar{\mathbf{x}}_i]$. Since $\psi^{1/2} \neq 0$, $\mathbf{x}_{it} - (1 - \psi^{1/2})\bar{\mathbf{x}}_i$ is different from 0 even $\mathbf{x}_{it}$ is time-invariant. In other words, the random-effects model allows one to estimate the coefficients of both time-varying and time-invariant variables while the fixed-effects model only allows us to estimate the coefficients of time-varying explanatory variables.

The GLS requires that of $\sigma_u^2$ and $\sigma_\alpha^2$ be known. If the variance components, $\sigma_u^2$ and $\sigma_\alpha^2$, are unknown, we can use two-step GLS estimation (feasible GLS, FGLS). In the first step we estimate the variance components using some consistent estimators. In the second step we substitute their estimated values into (3.3.10) or its equivalent form. Noting that $\bar{y}_i = \mu + \boldsymbol{\beta}'\bar{\mathbf{x}}_i + \alpha_i + \bar{u}_i$ and $(y_{it} - \bar{y}_i) = \boldsymbol{\beta}'(\mathbf{x}_{it} - \bar{\mathbf{x}}_i) + (u_{it} - \bar{u}_i)$, we can use the within- and between-group residuals to estimate $\sigma_u^2$ and $\sigma_\alpha^2$ respectively, by[7]

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}[(y_{it} - \bar{y}_i) - \hat{\boldsymbol{\beta}}_{cv}'(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)]^2}{N(T-1) - K}, \tag{3.3.17}$$

and

$$\hat{\sigma}_\alpha^2 = \frac{\sum_{i=1}^{N}(\bar{y}_i - \tilde{\mu} - \tilde{\boldsymbol{\beta}}'\bar{\mathbf{x}}_i)^2}{N - (K+1)} - \frac{1}{T}\hat{\sigma}_u^2, \tag{3.3.18}$$

where $(\tilde{\mu}, \tilde{\boldsymbol{\beta}}')' = B_{\tilde{x}\tilde{x}}^{-1} B_{\tilde{x}\tilde{y}}$. When the sample size is large (in the sense of $N \to \infty$, $T \to \infty$), the two-step GLS estimator will have the same asymptotic efficiency as the GLS procedure with known variance components (Fuller and Battese 1974). Even for moderate sample size [for $T \geq 3$, $N - (K+1) \geq 9$; for $T = 2$, $N - (K+1) \geq 10$], the two-step procedure is still more efficient than the CV (or within-group) estimator in the sense that the difference between the covariance matrices of the CV estimator and the two-step estimator is non-negative definite (Taylor 1980).

Amemiya (1971) has discussed efficient estimation of the variance components. However, substituting more efficiently estimated variance components into (3.3.12) need not lead to more efficient estimates of $\mu$ and $\boldsymbol{\beta}$ (Maddala and Mount 1973; Taylor 1980).

[7] Equation (3.3.18) may yield a negative estimate of $\sigma_\alpha^2$. For additional discussion on this issue, see Section 3.3.3.

### 3.3.3 Maximum-Likelihood Estimation

When $\alpha_i$ and $u_{it}$ are random and normally distributed, the logarithm of the likelihood function is

$$\begin{aligned}\log L &= -\frac{NT}{2}\log 2\pi - \frac{N}{2}\log \mid V \mid \\ &\quad -\frac{1}{2}\sum_{i=1}^{N}(\mathbf{y}_i - \mathbf{e}\mu - X_i\boldsymbol{\beta})'V^{-1}(\mathbf{y}_i - \mathbf{e}\mu - X_i\boldsymbol{\beta}) \\ &= -\frac{NT}{2}\log 2\pi - \frac{N(T-1)}{2}\log\sigma_u^2 - \frac{N}{2}\log\left(\sigma_u^2 + T\sigma_\alpha^2\right) \\ &\quad -\frac{1}{2\sigma_u^2}\sum_{i=1}^{N}(\mathbf{y}_i - \mathbf{e}\mu - X_i\boldsymbol{\beta})'Q(\mathbf{y}_i - \mathbf{e}\mu - X_i\boldsymbol{\beta}) \\ &\quad -\frac{T}{2(\sigma_u^2 + T\sigma_\alpha^2)}\sum_{i=1}^{N}(\bar{y}_i - \mu - \boldsymbol{\beta}'\bar{\mathbf{x}}_i)^2, \end{aligned} \tag{3.3.19}$$

where the second equality follows from (3.3.10) and

$$\mid V \mid = \sigma_u^{2(T-1)}\left(\sigma_u^2 + T\sigma_\alpha^2\right). \tag{3.3.20}$$

The maximum-likelihood estimator (MLE) of $(\mu, \boldsymbol{\beta}', \sigma_u^2, \sigma_\alpha^2) = \tilde{\boldsymbol{\theta}}'$ is obtained by solving the following first-order conditions simultaneously:

$$\frac{\partial \log L}{\partial \mu} = \frac{T}{(\sigma_u^2 + T\sigma_\alpha^2)}\sum_{i=1}^{N}\left(\bar{y}_i - \mu - \bar{\mathbf{x}}_i'\boldsymbol{\beta}\right) = 0, \tag{3.3.21}$$

$$\begin{aligned}\frac{\partial \log L}{\partial \boldsymbol{\beta}'} &= \frac{1}{\sigma_u^2}\Bigg[\sum_{i=1}^{N}(\mathbf{y}_i - \mathbf{e}\mu - X_i\boldsymbol{\beta})'QX_i \\ &\quad + \frac{T\sigma_u^2}{\left(\sigma_u^2 + T\sigma_\alpha^2\right)}\sum_{i=1}^{N}(\bar{y}_i - \mu - \bar{\mathbf{x}}_i'\boldsymbol{\beta})\bar{\mathbf{x}}_i'\Bigg] = \mathbf{0}', \end{aligned} \tag{3.3.22}$$

$$\begin{aligned}\frac{\partial \log L}{\partial \sigma_u^2} &= -\frac{N(T-1)}{2\sigma_u^2} - \frac{N}{2(\sigma_u^2 + T\sigma_\alpha^2)} + \frac{1}{2\sigma_u^4}\sum_{i=1}^{N}(\mathbf{y}_i - \mathbf{e}\mu \\ &\quad - X_i\boldsymbol{\beta})'Q(\mathbf{y}_i - \mathbf{e}\mu - X_i\boldsymbol{\beta}) \\ &\quad + \frac{T}{2\left(\sigma_u^2 + T\sigma_\alpha^2\right)^2}\sum_{i=1}^{N}(\bar{y}_i - \mu - \bar{\mathbf{x}}_i'\boldsymbol{\beta})^2 = 0, \end{aligned} \tag{3.3.23}$$

$$\frac{\partial \log L}{\partial \sigma_\alpha^2} = -\frac{NT}{2\left(\sigma_u^2 + T\sigma_\alpha^2\right)} + \frac{T^2}{2\left(\sigma_u^2 + T\sigma_\alpha^2\right)^2}\sum_{i=1}^{N}(\bar{y}_i - \mu - \bar{\mathbf{x}}_i'\boldsymbol{\beta})^2 = 0. \tag{3.3.24}$$

Simultaneous solution of (3.3.21)–(3.3.24) is complicated. The Newton–Raphson iterative procedure can be used to solve for the MLE. The procedure uses an initial trial value of $\tilde{\boldsymbol{\theta}}$, $\hat{\tilde{\boldsymbol{\theta}}}^{(1)}$, to start the iteration by substituting it into the formula

$$\hat{\tilde{\boldsymbol{\theta}}}^{(j)} = \hat{\tilde{\boldsymbol{\theta}}}^{(j-1)} - \left[\frac{\partial^2 \log L}{\partial \tilde{\boldsymbol{\theta}} \partial \tilde{\boldsymbol{\theta}}'}\right]^{-1}_{\tilde{\boldsymbol{\theta}}=\hat{\tilde{\boldsymbol{\theta}}}^{(j-1)}} \left.\frac{\partial \log L}{\partial \tilde{\boldsymbol{\theta}}}\right|_{\tilde{\boldsymbol{\theta}}=\hat{\tilde{\boldsymbol{\theta}}}^{(j-1)}} \tag{3.3.25}$$

to obtain a revised estimate of $\tilde{\boldsymbol{\theta}}$, $\hat{\tilde{\boldsymbol{\theta}}}^{(2)}$. The process is repeated until the $j$th iterative solution $\hat{\tilde{\boldsymbol{\theta}}}^{(j)}$ is close to the $(j-1)$th iterative solution $\hat{\tilde{\boldsymbol{\theta}}}^{(j-1)}$.

Alternatively, we can use a sequential iterative procedure to obtain the MLE. We note that from (3.3.21) and (3.3.22) we have

$$\begin{aligned}
\begin{bmatrix} \hat{\mu} \\ \hat{\boldsymbol{\beta}} \end{bmatrix} &= \left[\sum_{i=1}^{N} \tilde{X}_i' V^{-1} \tilde{X}_i\right]^{-1} \left[\sum_{i=1}^{N} \tilde{X}_i' V^{-1} \mathbf{y}_i\right] \\
&= \left\{\sum_{i=1}^{N} \begin{bmatrix} \mathbf{e}' \\ X_i' \end{bmatrix} \left[I_T - \frac{\sigma_\alpha^2}{\sigma_u^2 + T\sigma_\alpha^2} \mathbf{e}\mathbf{e}'\right] (\mathbf{e}, X_i)\right\}^{-1} \\
&\quad \cdot \left\{\sum_{i=1}^{N} \begin{bmatrix} \mathbf{e}' \\ X_i' \end{bmatrix} \left[I_T - \frac{\sigma_\alpha^2}{\sigma_u^2 + T\sigma_\alpha^2} \mathbf{e}\mathbf{e}'\right] \mathbf{y}_i\right\}.
\end{aligned} \tag{3.3.26}$$

Substituting (3.3.24) into (3.3.23), we have

$$\hat{\sigma}_u^2 = \frac{1}{N(T-1)} \sum_{i=1}^{N} (\mathbf{y}_i - \mathbf{e}\mu - X_i \boldsymbol{\beta})' Q (\mathbf{y}_i - \mathbf{e}\mu - X_i \boldsymbol{\beta}). \tag{3.3.27}$$

From (3.3.24) we have

$$\hat{\sigma}_\alpha^2 = \frac{1}{N} \sum_{i=1}^{N} (\bar{y}_i - \hat{\mu} - \bar{\mathbf{x}}_i' \hat{\boldsymbol{\beta}})^2 - \frac{1}{T} \hat{\sigma}_u^2. \tag{3.3.28}$$

Thus, we can obtain the MLE by first substituting an initial trial value of $\sigma_\alpha^2/(\sigma_u^2 + T\sigma_\alpha^2)$ into (3.3.26) to estimate $\mu$ and $\boldsymbol{\beta}$, and then estimate $\sigma_u^2$ by (3.3.27) using the solution of (3.3.26). Substituting the solutions of (3.3.26) and (3.3.27) into (3.3.28), we obtain an estimate of $\sigma_\alpha^2$. Then we repeat the process by substituting the new values of $\sigma_u^2$ and $\sigma_\alpha^2$ into (3.3.26) to obtain new estimates of $\mu$ and $\boldsymbol{\beta}$, and so on until the solution converges.

When $T$ is fixed and $N$ goes to infinity, the MLE is consistent and asymptotically normally distributed with variance–covariance matrix

$$\text{Var}\left(\sqrt{N}\hat{\tilde{\boldsymbol{\theta}}}_{\text{MLE}}\right) = NE\left[-\frac{\partial^2 \log L}{\partial\tilde{\boldsymbol{\theta}}\partial\tilde{\boldsymbol{\theta}}'}\right]^{-1}$$

$$= \begin{bmatrix} \frac{T}{\sigma^2} & \frac{T}{\sigma^2}\frac{1}{N}\sum_{i=1}^{N}\bar{\mathbf{x}}_i' & 0 & 0 \\ \frac{1}{\sigma_u^2}\frac{1}{N}\sum_{i=1}^{N}X_i'\left(I_T - \frac{\sigma_\alpha^2}{\sigma^2}\mathbf{e}\mathbf{e}'\right)X_i & & \mathbf{0} & \mathbf{0} \\ & & \frac{T-1}{2\sigma_u^2} + \frac{1}{2\sigma^4} & \frac{T}{2\sigma^4} \\ & & & \frac{T^2}{2\sigma^4} \end{bmatrix}^{-1} \tag{3.3.29}$$

where $\sigma^2 = \sigma_u^2 + T\sigma_\alpha^2$. When $N$ is fixed and $T$ tends to infinity, the MLEs of $\mu$, $\boldsymbol{\beta}$ and $\sigma_u^2$ converge to the CV estimator, and are consistent, but the MLE of $\sigma_\alpha^2$ is inconsistent. This is because when $N$ is fixed, there is not sufficient variation in $\alpha_i$ no matter how large $T$ is; for details, see Anderson and Hsiao (1981, 1982).

Although the MLE is asymptotically efficient, sometimes simultaneous solution of (3.3.21)–(3.3.24) yields an estimated value of $\sigma_\alpha^2$ that is negative.[8] When there is a unique solution to the partial derivative equations (3.3.21)–(3.3.24), with $\sigma_u^2 > 0, \sigma_\alpha^2 > 0$, the solution is the MLE. However, when we constrain $\sigma_u^2 \geq 0$ and $\sigma_\alpha^2 \geq 0$, a boundary solution may occur. The solution, then, no longer satisfies all the derivative equations (3.3.21)–(3.3.24). Maddala (1971a) has shown that the boundary solution of $\sigma_u^2 = 0$ cannot occur, but the boundary solution of $\sigma_\alpha^2 = 0$ will occur when $T_{yy} - T_{\tilde{x}y}'T_{\tilde{x}\tilde{x}}^{-1}T_{\tilde{x}y} > T[B_{yy} - 2T_{\tilde{x}y}'T_{\tilde{x}\tilde{x}}^{-1}T_{\tilde{x}y} + T_{\tilde{x}y}'T_{\tilde{x}\tilde{x}}^{-1}B_{\tilde{x}\tilde{x}}T_{\tilde{x}\tilde{x}}^{-1}T_{\tilde{x}y}]$. However, the probability of a boundary solution tends to 0 when either $T$ or $N$ tends to infinity.

## 3.4 FIXED EFFECTS OR RANDOM EFFECTS

### 3.4.1 An Example

In previous sections we discussed the estimation of a linear regression model (3.2.1) when the effects, $\alpha_i$, are treated either as fixed or as random. Whether to treat the effects as fixed or random makes no difference when $N$ is fixed and

[8] The negative-variance-components problem also arises in the two-step GLS method. As one can see from (3.3.17) and (3.3.18) that there is no guarantee that (3.3.18) necessarily yields a positive estimate of $\sigma_\alpha^2$. A practical guide in this situation is to replace a negative estimated variance component by its boundary value, zero. See Baltagi (1981b) and Maddala and Mount (1973) for a Monte Carlo studies of the desirable results of using this procedure in terms of the mean square error of the estimate. For additional discussion of the MLE of random effects model, see Breusch (1987).

$T$ is large because both the LSDV estimator (3.2.8) and the generalized least-squares estimator (3.3.14) become the same estimator. When $T$ is finite and $N$ is large, whether to treat the effects as fixed or random is not an easy question to answer. It can make a surprising amount of difference in the estimates of the parameters. In fact, when only a few observations are available for different individuals over time, it is exceptionally important to make the best use of the lesser amount of information over time for the efficient estimation of the common behavioral relationship.

For example, Hausman (1978) found that using a fixed-effects specification produced significantly different results from a random-effects specification when estimating a wage equation using a sample of 629 high school graduates followed over six years by the Michigan income dynamics study. The explanatory variables in the Hausman wage equation include a piecewise-linear representation of age, the presence of unemployment or poor health in the previous year, and dummy variables for self-employment, living in the South, or living in a rural area. The fixed-effects specification was estimated using (3.2.5).[9] The random-effects specification was estimated using (3.3.14). The results are reproduced in Table 3.3. In comparing these two estimates, it is apparent that the effects of unemployment, self-employment, and geographical location differ widely (relative to their standard errors) in the two models.

### 3.4.2 Conditional Inference or Unconditional (Marginal) Inference

If the effects of omitted variables can be appropriately summarized by a random variable and the individual (or time) effects represent the ignorance of the investigator, it does not seem reasonable to treat one source of ignorance ($\alpha_i$) as fixed and the other source of ignorance ($u_{it}$) as random. It appears that one way to unify the fixed-effects and random-effects models is to assume from the outset that the effects are random. The fixed-effects model is viewed as one in which investigators make inferences conditional on the effects that are in the sample. The random-effects model is viewed as one in which investigators make unconditional or marginal inferences with respect to the population of all effects. There is really no distinction in the "nature (of the effect)." It is up to the investigator to decide whether to make inference with respect to the population characteristics or only with respect to the effects that are in the sample.

In general, whether one wishes to consider the conditional likelihood function or the marginal likelihood function depends on the context of the data, the manner in which they were gathered, and the environment from which they came. For instance, consider an example in which several technicians provide maintenance for machines. The effects of technicians can be assumed random if the technicians are all randomly drawn from a common population. However, if the situation were one of analyzing just a few individuals, say five or six,

[9] We note that the fixed-effects estimator, although not efficient, is consistent under the random-effects formulation (Chapter 3, Section 3.3.1).

Table 3.3. *Wage equations (dependent variable: log wage[a])*

| Variable | Fixed effects | Random effects |
|---|---|---|
| 1. Age 1 (20–35) | 0.0557 | 0.0393 |
| | (0.0042) | (0.0033) |
| 2. Age 2 (35–45) | 0.0351 | 0.0092 |
| | (0.0051) | (0.0036) |
| 3. Age 3 (45–55) | 0.0209 | −0.0007 |
| | (0.0055) | (0.0042) |
| 4. Age 4 (55–65) | 0.0209 | −0.0097 |
| | (0.0078) | (0.0060) |
| 5. Age 5 (65–) | −0.0171 | −0.0423 |
| | (0.0155) | (0.0121) |
| 6. Unemployed previous year | −0.0042 | −0.0277 |
| | (0.0153) | (0.0151) |
| 7. Poor health previous year | −0.0204 | −0.0250 |
| | (0.0221) | (0.0215) |
| 8. Self-employment | −0.2190 | −0.2670 |
| | (0.0297) | (0.0263) |
| 9. South | −0.1569 | −0.0324 |
| | (0.0656) | (0.0333) |
| 10. Rural | −0.0101 | −0.1215 |
| | (0.0317) | (0.0237) |
| 11. Constant | — | 0.8499 |
| | — | (0.0433) |
| $s^2$ | 0.0567 | 0.0694 |
| Degrees of freedom | 3,135 | 3,763 |

[a] 3,774 observations; standard errors are in parentheses.
*Source:* Hausman (1978).

and the sole interest lay in just these individuals, and if we want to assess differences between those specific technicians, then the fixed-effects model is more appropriate. On the other hand, if an experiment involves hundreds of individuals who are considered a random sample from some larger population, random effects would be more appropriate. The situation to which a model applies and the inferences based on it are the deciding factors in determining whether we should treat effects as random or fixed. When inferences are going to be confined to the effects in the model, the effects are more appropriately considered fixed. When inferences will be made about a population of effects from which those in the data are considered to be a random sample, then the effects should be considered random.[10]

If one accepts this view, then why do the fixed-effects and random-effects approaches sometimes yield vastly different estimates of the common slope coefficients that are not supposed to vary across individuals? It appears that

[10] In this sense, if $N$ becomes large, one would not be interested in the specific effect of each individual but rather in the characteristics of the population. A random-effects framework would be more appropriate.

in addition to the efficiency issue discussed earlier, there is also a different but important issue of whether or not the model is properly specified, that is, whether the differences in individual effects can be attributed to the chance mechanism.

In the random effects framework of (3.3.3)–(3.3.5), there are two fundamental assumptions. One is that the unobserved individual effects, $\alpha_i$, are random draws from a common population. The other is that the explanatory variables are strictly exogenous. That is, the error terms are uncorrelated with (or orthogonal to) the past, current, and future values of the regressors,

$$\begin{aligned} E(u_{it} \mid \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}) &= E(\alpha_i \mid \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}) \\ &= E(v_{it} \mid \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}) = 0 \quad \text{for } t = 1, \ldots, T. \end{aligned} \tag{3.4.1}$$

In the aforementioned example if there are fundamental differences in the technicians, for instance, in the ability, age, years of experiences, etc., then the difference in technician cannot be attributed to a pure chance mechanism. It is more appropriate to view the technicians as drawn from heterogeneous populations and the individual effects $\alpha_i^* = \alpha_i + \mu$ representing the fundamental difference among the heterogeneous populations. If the difference in technicians, captured by $\alpha_i^*$ is ignored, the least-squares estimator of (3.3.5) yields

$$\hat{\boldsymbol{\beta}}_{LS} = \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it} - \bar{\mathbf{x}})(\mathbf{x}_{it} - \bar{\mathbf{x}})'\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it} - \bar{\mathbf{x}})(y_{it} - \bar{y})\right] \tag{3.4.2}$$

$$= \boldsymbol{\beta} + \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it} - \bar{\mathbf{x}})(\mathbf{x}_{it} - \bar{\mathbf{x}})'\right]^{-1}\left\{T\sum_{i=1}^{N}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\alpha_i^* - \bar{\alpha})\right\} + \mathrm{o}(1)$$

where $\bar{\alpha} = \frac{1}{N}\sum_{i=1}^{N}\alpha_i^*$. However, if the fundamental characteristics that drive $\alpha_i^*$, say, ability, age, and years of experience in the example of technicians, are correlated with $\mathbf{x}_i$, then it is clear that $\frac{1}{N}\sum_{i=1}^{N}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\alpha_i^* - \bar{\alpha})$ will not converge to 0 as $N \to \infty$. The least-squares estimator of $\boldsymbol{\beta}$ is inconsistent. The bias of $\hat{\boldsymbol{\beta}}_{LS}$ depends on the correlation between $\mathbf{x}_{it}$ and $\alpha_i^*$.

On the other hand, if $\alpha_i^*$ (or $\alpha_i$) are treated as fixed constants, then the regressors for $y_{it}$ are $(\mathbf{x}_{it}', 1)$. As long as $(\mathbf{x}_{it}', 1)$ are uncorrelated with $u_{it}$, the least-squares estimators for $\boldsymbol{\beta}$ and $\alpha_i^*$ (or $\alpha_i$) are unbiased. The issue of whether $\alpha_i^*$ are correlated with $\mathbf{x}_{it}$ is no longer relevant under the fixed-effects formulation. Thus, unless the distribution of $\alpha_i^*$ conditional on $\mathbf{x}_i$ can be appropriately formulated, it would be more appropriate to treat $\alpha_i^*$ as fixed and different (Hsiao and Sun 2000).

### *3.4.2.1 Mundlak's Formulation*

Mundlak (1978a) criticized the random-effects formulation (3.3.4) on the grounds that it neglects the correlation that may exist between the effects,

$\alpha_i$, and the explanatory variables, $\mathbf{x}_{it}$. There are reasons to believe that in many circumstances $\alpha_i$ and $\mathbf{x}_{it}$ are indeed correlated. For instance, consider the estimation of production function using firm data. The output of each firm, $y_{it}$, may be affected by unobservable managerial ability, $\alpha_i$. Firms with more efficient management tend to produce more and use more inputs, $X_i$. Less efficient firms tend to produce less and use fewer inputs. In this situation, $\alpha_i$ and $X_i$ cannot be independent. Ignoring this correlation can lead to biased estimation.

The properties of various estimators we have discussed thus far depend on the existence and extent of the relations between the $X$'s and the effects. Therefore, we have to consider the joint distribution of these variables. However, $\alpha_i$ are unobservable. Mundlak (1978a) suggested that we approximate $E(\alpha_i \mid X_i)$ by a linear function. He introduced the auxiliary regression

$$\alpha_i = \sum_t \mathbf{x}_{it}'\mathbf{a}_t + \omega_i, \quad \omega_i \sim N\left(0, \sigma_\omega^2\right). \tag{3.4.3a}$$

A simple approximation to (3.4.3a) is to let

$$\alpha_i = \bar{\mathbf{x}}_i'\mathbf{a} + \omega_i, \quad \omega_i \sim N\left(0, \sigma_\omega^2\right). \tag{3.4.3b}$$

Clearly, $\mathbf{a}$ will be equal to 0 (and $\sigma_\omega^2 = \sigma_\alpha^2$) if (and only if) the explanatory variables are uncorrelated with the effects.

Substituting (3.4.3b) into (3.3.5), and stacking equations over $t$ and $i$, we have

$$\begin{aligned}\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\\ \vdots\\ \mathbf{y}_N\end{bmatrix} &= \begin{bmatrix}\tilde{X}_1\\ \tilde{X}_2\\ \vdots\\ \tilde{X}_N\end{bmatrix}\boldsymbol{\delta} + \begin{bmatrix}\mathbf{e}\bar{\mathbf{x}}_1'\\ \mathbf{e}\bar{\mathbf{x}}_2'\\ \vdots\\ \mathbf{e}\bar{\mathbf{x}}_N'\end{bmatrix}\mathbf{a} + \begin{bmatrix}\mathbf{e}\\ \mathbf{0}\\ \vdots\\ \mathbf{0}\end{bmatrix}\omega_1\\ &\quad + \begin{bmatrix}\mathbf{0}\\ \mathbf{e}\\ \vdots\\ \mathbf{0}\end{bmatrix}\omega_2 + \cdots + \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\\ \vdots\\ \mathbf{e}_N\end{bmatrix}\omega_N + \begin{bmatrix}\mathbf{u}_1\\ \mathbf{u}_2\\ \vdots\\ \mathbf{u}_N\end{bmatrix},\end{aligned} \tag{3.4.4}$$

where

$$E(\mathbf{u}_i + \mathbf{e}\omega_i) = \mathbf{0},$$

$$E(\mathbf{u}_i + \mathbf{e}\omega_i)(\mathbf{u}_j + \mathbf{e}\omega_j)' = \begin{cases}\sigma_u^2 I_T + \sigma_\omega^2 \mathbf{e}\mathbf{e}' = \tilde{V}, & \text{if } i = j,\\ \mathbf{0}, & \text{if } i \neq j,\end{cases}$$

$$\tilde{V}^{-1} = \frac{1}{\sigma_u^2}\left[I_T - \frac{\sigma_\omega^2}{\sigma_u^2 + T\sigma_\omega^2}\mathbf{e}\mathbf{e}'\right].$$

Utilizing the expression for the inverse of a partitioned matrix (Theil 1971, Chapter 1), we obtain the GLS of $(\mu, \boldsymbol{\beta}', \mathbf{a}')$ as

$$\hat{\mu}^*_{\text{GLS}} = \bar{y} - \bar{\mathbf{x}}'\hat{\boldsymbol{\beta}}_b, \tag{3.4.5}$$

$$\hat{\boldsymbol{\beta}}^*_{\text{GLS}} = \hat{\boldsymbol{\beta}}_{cv}, \tag{3.4.6}$$

$$\hat{\mathbf{a}}^*_{\text{GLS}} = \hat{\boldsymbol{\beta}}_b - \hat{\boldsymbol{\beta}}_{cv}. \tag{3.4.7}$$

Thus, in the present framework, the BLUE of $\boldsymbol{\beta}$ is the CV estimator of (3.2.1) or (3.2.10′). It does not depend on knowledge of the variance components. Therefore, Mundlak (1978a) maintained that the imaginary difference between the fixed-effects and random-effects approaches is based on an incorrect specification. In fact, applying GLS to (3.2.12) yields a biased estimator. This can be seen by noting that the GLS estimate of $\boldsymbol{\beta}$ for (3.3.5), that is, (3.3.12), can be viewed as the GLS estimate of (3.4.4) after imposing the restriction $\mathbf{a} = \mathbf{0}$. As shown in (3.3.12),

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = \Delta\hat{\boldsymbol{\beta}}_b + (I_K - \Delta)\hat{\boldsymbol{\beta}}_{CV}. \tag{3.4.8}$$

If (3.4.4) is the correct specification, $E\hat{\boldsymbol{\beta}}_b$ is equal to $\boldsymbol{\beta} + \mathbf{a}$, and $E\hat{\boldsymbol{\beta}}_{cv} = \boldsymbol{\beta}$, so that

$$E\hat{\boldsymbol{\beta}}_{\text{GLS}} = \boldsymbol{\beta} + \Delta\mathbf{a}. \tag{3.4.9}$$

This is a biased estimator if $\mathbf{a} \neq \mathbf{0}$. However, when $T$ tends to infinity, $\Delta$ tends to 0, and $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ tends to $\hat{\boldsymbol{\beta}}_{cv}$ and is asymptotically unbiased. But in the more relevant situation in which $T$ is fixed and $N$ tends to infinity, $\text{plim}_{N\to\infty}\hat{\boldsymbol{\beta}}_{\text{GLS}} \neq \boldsymbol{\beta}$ in Mundlak's formulation.

Though it is important to recognize the possible correlation between the effects and the explanatory variables, Mundlak's (1978a) claim that there is only one estimator and that efficiency is not a consideration in distinguishing between the random-effects and fixed-effects approaches is perhaps a bit strong. Mandlak derived (3.4.6) from the assumption that $f(\alpha_i \mid \mathbf{x}_i)$ has mean $\bar{\mathbf{x}}'\mathbf{a}$ and variance $\sigma^2_\omega$ for the linear model (3.2.12) only. In the dynamic, random-coefficient, and discrete-choice models to be discussed later, one can show that the two approaches do not lead to the same estimator even when one allows for the correlation between $\alpha_i$ and $X_i$ following the formulation of Mundlak (1978a). Moreover, in the linear static model, if $\mathbf{a} = \mathbf{0}$, the efficient estimator is (3.3.14), not the CV estimator (3.2.8).

### *3.4.2.2 Conditional and Unconditional Inferences in the Presence or Absence of Correlation between Individual Effects and Attributes*

To gain further intuitive notions about the differences between models (3.3.5) and (3.4.4) within the conditional and unconditional inference frameworks, we consider the following two experiments. Let a population be made up of a certain composition of red and black balls. The first experiment consists of $N$ individuals, each picking a fixed number of balls randomly from this population

to form his person-specific jar. Each individual then makes $T$ independent trials of drawing a ball from his specific jar and putting it back. The second experiment assumes that individuals have different preferences for the compositions of red and black balls for their specific jars and allows personal attributes to affect the compositions. Specifically, before making $T$ independent trials with replacement from their respective jars, individuals are allowed to take any number of balls from the population until their compositions reach the desired proportions.

If one is interested in making inferences regarding an individual jar's composition of red and black balls, a fixed-effects model should be used, whether the sample comes from the first or the second experiment. On the other hand, if one is interested in the population composition, a marginal or unconditional inference should be used. However, the marginal distributions are different for these two cases. In the first experiment, differences in individual jars are outcomes of random sampling. The subscript $i$ is purely a labeling device, with no substantive content. A conventional random-effects model assuming independence between $\alpha_i$ and $\mathbf{x}_{it}$ would be appropriate. In the second experiment, the differences in individual jars reflect differences in personal attributes. A proper marginal inference has to allow for these nonrandom effects. In other words, individuals are not random draws from a common population, but from heterogeneous populations. In Mundlek's formulation, this heterogeneity is captured by the observed attributes $\mathbf{x}_i$. For the Mundlak's formulation a marginal inference that properly allows for the correlation between individual effects ($\alpha_i$) and the attributes ($\mathbf{x}_i$) in the data-generating process gives rise to the same estimator as when the individual effects are treated as fixed. It is not that in making inferences about population characteristics, we should assume a fixed-effects model.

Formally, let $u_{it}$ and $\alpha_i$ be independent normal processes that are mutually independent. In the case of the first experiment, $\alpha_i$ are independently distributed and independent of individual attributes, $\mathbf{x}_i$, so the distribution of $\alpha_i$ must be expressible as random sampling from a univerate distribution (Box and Tiao 1968; Chamberlain 1980). Thus, the conditional distribution of $\{(u_i + e\alpha_i)', \alpha_i \mid X_i\}$ is identical with the marginal distribution of $\{(\mathbf{u}_i + \mathbf{e}\alpha_i)', \alpha_i\}$,

$$\begin{bmatrix} u_{i1}+\alpha_i \\ \vdots \\ u_{iT}+\alpha_i \\ \cdots \\ \alpha_i \end{bmatrix} = \left[\begin{array}{cc} u_{i1}+\alpha_i & | \\ \vdots & | \\ u_{iT}+\alpha_i & | \quad X_i \\ \cdots & | \\ \alpha_i & | \end{array}\right] \tag{3.4.10a}$$

$$\sim N\left[\begin{bmatrix} \mathbf{0} \\ \cdots \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 I_T + \sigma_\alpha^2 \mathbf{e}\mathbf{e}' & \vdots\, \sigma_\alpha^2 \mathbf{e} \\ \cdots & \vdots \cdots \\ \sigma_\alpha^2 \mathbf{e}' & \vdots\, \sigma_\alpha^2 \end{bmatrix}\right].$$

In the second experiment, $\alpha_i$ may be viewed as a random draw from a heterogeneous population with mean $a_i^*$ and variance $\sigma_{\omega_i}^2$ (Mundlak's (1978a) formulation may be viewed as a special case of this in which $E(\alpha_i \mid X_i) = a_i^* = \mathbf{a}'\bar{\mathbf{x}}_i$ and $\sigma_{\omega_i}^2 = \sigma_\omega^2$ for all $i$). Then the conditional distribution of $\{(\mathbf{u}_i + \mathbf{e}\alpha_i)' \vdots \alpha_i \mid X_i\}$ is

$$\left[\begin{array}{c|c} u_{i1} + \alpha_i & \\ \vdots & \\ u_{iT} + \alpha_i & X_i \\ \cdots & \\ \alpha_i & \end{array}\right] \sim N\left[\left[\begin{array}{c} \mathbf{e}a_i^* \\ \cdots \\ a_i^* \end{array}\right], \left[\begin{array}{cc} \sigma_u^2 I_T + \sigma_{\omega i}^2 \mathbf{e}\mathbf{e}' & \vdots\, \sigma_{\omega i}^2 \mathbf{e} \\ \cdots & \\ \sigma_{\omega i}^2 \mathbf{e}' & \vdots\, \sigma_{\omega i}^2 \end{array}\right]\right]. \quad (3.4.10b)$$

In both cases, the conditional density of $\mathbf{u}_i + \mathbf{e}\alpha_i$, given $\alpha_i$, is[11]

$$\left(2\pi\sigma_u^2\right)^{T/2} \exp\left\{-\frac{1}{2\sigma_u^2}\mathbf{u}_i'\mathbf{u}_i\right\}. \quad (3.4.11)$$

But the marginal densities of $\mathbf{u}_i + \mathbf{e}\alpha_i$, given $X_i$, are different [(3.4.10a) and (3.4.10b), respectively]. Under the independence assumption, $\{\mathbf{u}_i + \mathbf{e}\alpha_i \mid X_i\}$ has a common mean of 0 for $i = 1, \ldots, N$. Under the assumption that $\alpha_i$ and $X_i$ are correlated or $\alpha_i$ is a draw from a heterogeneous population, $\{\mathbf{u}_i + \mathbf{e}\alpha_i \mid X_i\}$ has a different mean $\mathbf{e}a_i^*$ for different $i$.

In the linear regression model, conditional on $\alpha_i$ the Jacobian of transformation from $\mathbf{u}_i + \mathbf{e}\alpha_i$ to $\mathbf{y}_i$ is 1. Maximizing the conditional likelihood function of $(\mathbf{y}_1 \mid \alpha_1, X_1), \ldots, (\mathbf{y}_N \mid \alpha_N, X_N)$, treating $\alpha_i$ as unknown parameters, yields the CV (or within-group) estimators for both cases. Maximizing the marginal likelihood function of $(y_1, \ldots, y_N \mid X_1, \ldots, X_N)$ yields the GLS estimator for model (3.3.12) under (3.4.10a) if $\sigma_u^2$ and $\sigma_\alpha^2$ are known, and it happens to yield the CV estimator for model (3.2.12) under (3.4.10b). In other words, there is no loss of information using a conditional approach for the case of (3.4.10b). However, there is a loss in efficiency in maximizing the conditional likelihood function for the former case [i.e., (3.4.10a)] because of the loss of degrees of freedom in estimating additional $(\alpha_1, \ldots, \alpha_N)$ unknown parameters, which leads to ignoring the information contained in the between-group variation.

The advantage of the unconditional inference is that the likelihood function may depend on only a finite number of parameters, and hence can often lead to efficient inference. The disadvantage is that the correct specification of the

[11] If $(Y^{(1)'}, Y^{(2)'})'$ is normally distributed with mean $(\boldsymbol{\mu}^{(1)'}, \boldsymbol{\mu}^{(2)'})'$ and variance–covariance matrix

$$\left[\begin{array}{cc} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{array}\right],$$

the conditional distribution of $Y^{(1)}$ given $Y^{(2)} = \mathbf{y}^{(2)}$ is normal, with mean $\boldsymbol{\mu}^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{y}^{(2)} - \boldsymbol{\mu}^{(2)})$ and covariance matrix $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ (e.g., Anderson 1985, Section 2.5).

conditional density of $\mathbf{y}_i$ given $\mathbf{X}_i$,

$$f(\mathbf{y}_i \mid X_i) = \int f(\mathbf{y}_i \mid X_i, \alpha_i) f(\alpha_i \mid X_i)\, d\alpha_i \tag{3.4.12}$$

depends on the correct specification of $f(\alpha_i \mid X_i)$. A misspecified $f(\alpha_i \mid X_i)$ can lead to a misspecified $f(\mathbf{y}_i \mid X_i)$. Maximizing the wrong $f(\mathbf{y}_i \mid X_i)$ can lead to biased and inconsistent estimators. The bias of the GLS estimator (3.3.12) in the case that $\alpha_i \sim N(a_i^*, \sigma_{\omega i}^2)$ is not due to any fallacy of the unconditional inference, but due to the misspecification of $f(\alpha_i \mid X_i)$.

The advantage of the conditional inference is that there is no need to specify $f(\alpha_i \mid X_i)$. Therefore, if the distribution of effects cannot be represented by a simple parametric functional form (say bimodal), or one is not sure of the correlation pattern between the effects and $X_i$, there may be an advantage to base one's inference conditionally. For instance, in the situation that there are fundamental differences between the effects, if there are fundamental differences in the ability, years of experiences, etc. as in the previous example of technicians, then it is more appropriate to treat the technicians' effects as fixed.

The disadvantage of the conditional inference is that not only there is a loss of efficiency due to the loss of degrees of freedom of estimating the effects, but there is also an issue of incidental parameters if $T$ is finite (Neyman–Scott 1948). A typical panel contains a large number of individuals observed over a short time period, and the number of individual effects parameters ($\alpha_i^*$) increases with the number of cross-sectional dimension, $N$. Because an increase in $N$ provides no information on a particular $\alpha_i^*$ apart from those already contained in $\mathbf{y}_i$, $\alpha_i^*$ cannot be consistently estimated with finite $T$. The condition that

$$E(u_{it} \mid \mathbf{x}_{it}) = 0 \tag{3.4.13}$$

is not informative about the common parameters, $\boldsymbol{\beta}$, in the absence of any knowledge about $\alpha_i^*$. If the estimation of the incidental parameters, $\alpha_i^*$, is not asymptotically independent of the estimation of the common parameters (called structural parameters in statistical literature), the conditional inference of the common parameter, $\boldsymbol{\beta}$, conditional on the inconsistently estimated $\alpha_i^*$, in general, will be inconsistent.

In the case of linear static model (3.2.1) or (3.2.12), the strict exogeneity of $\mathbf{x}_{it}$ to $u_{it}$,

$$E(u_{it} \mid \mathbf{x}_i) = 0, \quad t = 1, 2, \ldots, T, \tag{3.4.14}$$

where $\mathbf{x}_i' = (\mathbf{x}_{i1}', \ldots, \mathbf{x}_{iT}')$, implies that

$$E(u_{it} - \bar{u}_i \mid \mathbf{x}_i) = 0, \quad \begin{matrix} t = 1, 2, \ldots, T, \\ i = 1, \ldots, N. \end{matrix} \tag{3.4.15}$$

Since $\boldsymbol{\beta}$ can be identified from the moment conditions of the form (3.4.15) in the linear static model and (3.4.15) no longer involves $\alpha_i^*$, consistent estimators of $\boldsymbol{\beta}$ can be proposed by making use of these moment conditions (e.g., (3.2.8)).

Unfortunately, for nonlinear panel data models, it is in general not possible to find moment conditions that are independent of $\alpha_i^*$ to provide consistent estimators of common parameters.

The advantage of fixed-effects inference is that there is no need to assume that the effects are independent of $\mathbf{x}_i$. The disadvantage is that it introduces the issue of incidental parameters. Moreover, in the case of linear regression model, condition (3.4.15) implies that the coefficients of time-invariant variables cannot be estimated and the estimation of $\boldsymbol{\beta}$ makes use of only within-group variation, which is usually much smaller than between-group variation. The advantage of random-effects inference is that the number of parameters is fixed when sample size increases. It also allows the derivation of efficient estimators that make use of both within- and between-group variation. The impact of time-invariant variables can also be estimated. The disadvantage is that one has to make specific assumption about the pattern of correlation (or no correlation) between the effects and the included explanatory variables. A common assumption is that $f(\alpha_i \mid \mathbf{x}_i)$ is identical to the marginal density $f(\alpha_i)$. However, if the effects are correlated with $\mathbf{x}_{it}$ or if there is a fundamental difference among individual units, that is, conditional on $\mathbf{x}_{it}$, $y_{it}$ cannot be viewed as a random draw from a common distribution, common random-effects model (3.3.4) is misspecified and the resulting estimator is biased. In short, the advantages of random-effects specification are the disadvantages of fixed-effects specification and the disadvantages of random-effects specification are the advantages of fixed-effects specification. Unfortunately, there is no universally accepted way to make explicit assumptions about the way in which observables and unobservables interact in all contexts.

Finally, it should be noted that the assumption of randomness does not carry with it the assumption of normality. Often this assumption is made for random effects, but it is a separate assumption made subsequent to the randomness assumption. Most estimation procedures do not require normality, although if distributional properties of the resulting estimators are to be investigated, then normality is often assumed.

## 3.5 TESTS FOR MISSPECIFICATION

As discussed in Section 3.4, the fundamental issue is not whether $\alpha_i$ should be treated fixed or random. The issue is whether or not $f(\alpha_i \mid \mathbf{x}_i) \equiv f(\alpha_i)$, or whether $\alpha_i$ can be viewed as random draws from a common population. In the linear regression framework, treating $\alpha_i$ as fixed in (3.2.12) leads to the identical estimator of $\boldsymbol{\beta}$ whether $\alpha_i$ is correlated with $\mathbf{x}_i$ as in (3.4.3a) or is from a heterogeneous population. Hence, for ease of reference, when $\alpha_i$ is correlated with $\mathbf{x}_i$, we shall follow the convention and call (3.2.12) a fixed-effects model, and when $\alpha_i$ is uncorrelated with $\mathbf{x}_i$, we shall call it a random-effects model.

Thus, one way to decide whether to use a fixed-effects or random-effects model is to test for misspecification of (3.3.4), where $\alpha_i$ is assumed random and uncorrelated with $\mathbf{x}_i$. Using Mundlak's formulation, (3.4.3a) or (3.4.3b),

this test can be reduced to a test of

$$H_0 : \mathbf{a} = \mathbf{0},$$

against

$$H_1 : \mathbf{a} \neq \mathbf{0}.$$

If the alternative hypothesis, $H_1$, holds, we use the fixed-effects model (3.2.1). If the null hypothesis, $H_0$, holds, we use the random-effects model (3.3.4). The ratio

$$F = \frac{\left[\Sigma_{i=1}^{N}(\mathbf{y}_i - \tilde{X}_i\hat{\boldsymbol{\delta}}_{\mathrm{GLS}})'V^{*-1}(\mathbf{y}_i - \tilde{X}_i\hat{\boldsymbol{\delta}}_{\mathrm{GLS}}) - \Sigma_{i=1}^{N}(\mathbf{y}_i - \tilde{X}_i\hat{\boldsymbol{\delta}}^{*}_{\mathrm{GLS}} - \mathbf{e}\bar{\mathbf{x}}_i'\hat{\mathbf{a}}^{*}_{\mathrm{GLS}})'V^{*-1}\cdot(\mathbf{y}_i - \tilde{X}_i\hat{\boldsymbol{\delta}}^{*}_{\mathrm{GLS}} - \mathbf{e}\bar{\mathbf{x}}_i'\hat{\mathbf{a}}^{*}_{\mathrm{GLS}})\right]/K}{\Sigma_{i=1}^{N}(\mathbf{y}_i - \tilde{X}_i\hat{\boldsymbol{\delta}}^{*}_{\mathrm{GLS}} - \mathbf{e}\bar{\mathbf{x}}_i'\hat{\mathbf{a}}^{*}_{\mathrm{GLS}})'V^{*-1}(\mathbf{y}_i - \tilde{X}_i\hat{\boldsymbol{\delta}}^{*}_{\mathrm{GLS}} - \mathbf{e}\bar{\mathbf{x}}_i'\hat{\mathbf{a}}^{*}_{\mathrm{GLS}})/[NT - (2K+1)]} \tag{3.5.1}$$

under $H_0$ has a central $F$ distribution with $K$ and $NT - (2K + 1)$ degrees of freedom, where $\hat{\boldsymbol{\delta}}^{*}_{\mathrm{GLS}} = (\hat{\mu}_{\mathrm{GLS}}, \hat{\boldsymbol{\beta}}^{'}_{\mathrm{GLS}})'$, and $\hat{\mathbf{a}}^{*}_{\mathrm{GLS}}$ are given by (3.4.5)–(3.4.7), $V^{*-1} = (1/\sigma_u^2)[Q + \psi^*(1/T)\mathbf{e}\mathbf{e}']$, and $\psi^* = \sigma_u^2/(\sigma_u^2 + T\sigma_\omega^2)$. Hence, (3.5.1) can be used to test $H_0$ against $H_1$.[12]

An alternative testing procedure suggested by Hausman (1978) notes that under $H_0$ the GLS for (3.3.5) achieves the Cramer–Rao lower bounds, but under $H_1$, the GLS is a biased estimator. In contrast, the CV estimator of $\boldsymbol{\beta}$ is consistent under both $H_0$ and $H_1$. Hence, the Hausman test basically asks if the CV and GLS estimates of $\boldsymbol{\beta}$ are significantly different.

To derive the asymptotic distribution of the differences of the two estimates, Hausman makes use of the following lemma:[13]

**Lemma 3.5.1:** Based on a sample of $N$ observations, consider two estimates $\hat{\boldsymbol{\beta}}_0$ and $\hat{\boldsymbol{\beta}}_1$ that are both consistent and asymptotically normally distributed, with $\hat{\beta}_0$ attaining the asymptotic Cramer–Rao bound so that $\sqrt{N}(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta})$ is asymptotically normally distributed with variance–covariance matrix $V_0$. $\sqrt{N}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta})$ is asymptotically normally distributed, with mean 0 and variance–covariance matrix $V_1$. Let $\hat{\mathbf{q}} = \hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0$. Then the limiting distribution of $\sqrt{N}(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta})$ and $\sqrt{N}\hat{\mathbf{q}}$ has 0 covariance: Cov $(\hat{\boldsymbol{\beta}}_0, \hat{\mathbf{q}}) = \mathbf{0}$, a zero matrix.

[12] When $\psi^*$ is unknown, we substitute it by an estimated value and treat (3.5.1) as having an approximate $F$ distribution.

[13] For proof, see Hausman (1978) or Rao (1973, p. 317).

From this lemma, it follows that $\text{Var}(\hat{\mathbf{q}}) = \text{Var}(\hat{\boldsymbol{\beta}}_1) - \text{Var}(\hat{\boldsymbol{\beta}}_0)$. Thus, Hausman suggests using the statistic[14]

$$m = \hat{\mathbf{q}}' \text{Var}(\hat{\mathbf{q}})^{-1} \hat{\mathbf{q}}, \tag{3.5.2}$$

where $\hat{\mathbf{q}} = \hat{\boldsymbol{\beta}}_{\text{CV}} - \hat{\boldsymbol{\beta}}_{\text{GLS}}$, $\text{Var}(\hat{\mathbf{q}}) = \text{Var}(\hat{\boldsymbol{\beta}}_{\text{CV}}) - \text{Var}(\hat{\boldsymbol{\beta}}_{\text{GLS}})$, to test the null hypothesis $E(\alpha_i \mid X_i) = 0$ against the alternative $E(\alpha_i \mid X_i) \neq 0$. Under the null hypothesis, this statistic is distributed asymptotically as central $\chi^2$, with $K$ degrees of freedom. Under the alternative, it has a noncentral $\chi^2$ distribution with noncentrality parameter $\bar{\mathbf{q}}' \text{ Var}(\hat{\mathbf{q}})^{-1} \bar{\mathbf{q}}$, where $\bar{\mathbf{q}} = \text{plim}(\hat{\boldsymbol{\beta}}_{\text{CV}} - \hat{\boldsymbol{\beta}}_{\text{GLS}})$.

When $N$ is fixed and $T$ tends to infinity, $\hat{\boldsymbol{\beta}}_{\text{CV}}$ and $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ become identical. However, it was shown by Ahn and Moon (2001) that the numerator and denominator of (3.5.2) approach 0 at the same speed. Therefore the ratio remains $\chi^2$ distributed, although in this situation the fixed-effects and random-effects models become indistinguishable for all practical purposes. The more typical case in practice is that $N$ is large relative to $T$, so that differences between the two estimators or two approaches are important problems.

We can use either (3.5.1) or (3.5.2) to test whether a fixed-effects or random-effects formulation is more appropriate for the wage equation cited at the beginning of Section 3.4 (Table 3.3). The advantage of the Hausman approach is that no $f(\alpha_i \mid \mathbf{x}_i)$ needs to be postulated. The $\chi^2$ statistic for (3.5.2) computed by Hausman (1978) is 129.9. The critical value for the 1 percent significance level at 10 degrees of freedom is 23.2, a very strong indication of misspecification in the conventional random-effects model (3.3.3). Similar conclusions are also obtained by using (3.5.1). The $F$ value computed by Hausman (1978) is 139.7, which well exceeds the 1 percent critical value. These tests imply that in the Michigan survey, important individual effects are present that are correlated with the right-hand variables. Because the random-effects estimates appear to be significantly biased with high probability, it may well be important to take account of permanent unobserved differences across individuals in estimating earnings equations using panel data.

## 3.6 MODELS WITH TIME- AND/OR INDIVIDUAL-INVARIANT EXPLANATORY VARIABLES AND BOTH INDIVIDUAL- AND TIME-SPECIFIC EFFECTS

### 3.6.1 Estimation of Models with Individual-Specific Variables

Model (3.2.12) can be generalized to a number of different directions with no fundamental change in the analysis. For instance, we can include a $1 \times p$ vector $\mathbf{z}_i'$ of individual-specific variables (such as sex, race, socioeconomic

[14] Strictly speaking, the Hausman test is more general than a test of $\Sigma_t \mathbf{x}_{it}' \mathbf{a}_t = 0$ versus $\Sigma_t \mathbf{x}_{it}' \mathbf{a}_t \neq 0$. The null of $f(\alpha_i \mid \mathbf{x}_i) = f(\alpha_i)$ implies that $\Sigma_t \mathbf{x}_{it}' \mathbf{a}_t = 0$, but not necessarily the converse. For a discussion of the general relationship between Hausman's specification testing and conventional testing procedures, see Holly (1982).

background variables, which vary across individual units but do not vary over time) in the specification of the equation for $y_{it}$ and consider

$$\begin{array}{ccccccccccc} y_i & = & \mathbf{e} & \mu & + & Z_i & \boldsymbol{\gamma} & + & X_i & \boldsymbol{\beta} & \\ T\times 1 & & T\times 1 & 1\times 1 & & T\times p & p\times 1 & & T\times K & K\times 1 & \\ & + & \mathbf{e} & \alpha_i & + & \mathbf{u}_i & & & & & \\ & & T\times 1 & 1\times 1 & & T\times 1 & & & & & \end{array}, \quad i=1,\ldots,N, \tag{3.6.1}$$

where

$$\begin{array}{ccc} Z_i = & \mathbf{e} & \mathbf{z}_i' \\ & T\times 1 & 1\times p. \end{array}$$

If we assume that the $\alpha_i$ are fixed constants, model (3.6.1) is subject to perfect multicollinearity because $Z=(Z_1',\ldots,Z_N')'$ and $(I_N\otimes\mathbf{e})$ are perfectly correlated.[15] Hence, $\boldsymbol{\gamma}$, $\mu$, and $\alpha_i$ are not separately estimable. However, $\boldsymbol{\beta}$ may still be estimated by the covariance method (provided $\Sigma_{i=1}^N X_i'QX_i$ is of full rank). Premultiplying (3.6.1) by the (covariance) transformation matrix $Q$ [(3.2.6)], we sweep out $Z_i$, $\mathbf{e}\mu$, and $\mathbf{e}\alpha_i$ from (3.6.1), so that

$$Q\mathbf{y}_i = QX_i\boldsymbol{\beta} + Q\mathbf{u}_i, \quad i=1,\ldots,N. \tag{3.6.2}$$

Applying OLS to (3.6.2), we obtain the CV estimate of $\boldsymbol{\beta}$, (3.2.8).

There is no way one can separately identify $\boldsymbol{\gamma}$ and $\alpha_i^*$ under a fixed-effects formulation. However, if $\mathbf{z}_i$ and $\alpha_i^*$ are uncorrelated across $i$, one can treat $\alpha_i=\alpha_i^*-\mu$ as a random variable, where $\mu=\lim\frac{1}{N}\sum_{i=1}^N\alpha_i^*$. When the $\alpha_i$ are assumed random and uncorrelated with $X_i$ and $Z_i$, CV uses the same method to estimate $\boldsymbol{\beta}$ (3.2.8). To estimate $\boldsymbol{\gamma}$, we note that the individual mean over time can be written in the form

$$\bar{y}_i - \bar{\mathbf{x}}_i'\boldsymbol{\beta} = \mu + \mathbf{z}_i'\boldsymbol{\gamma} + \alpha_i + \bar{u}_i, \quad i=1,\ldots,N. \tag{3.6.3}$$

Treating $(\alpha_i+\bar{u}_i)$ as the error term and minimizing $\Sigma_{i=1}^N(\alpha_i+\bar{u}_i)^2$, we obtain

$$\hat{\boldsymbol{\gamma}} = \left[\sum_{i=1}^N(\mathbf{z}_i-\bar{\mathbf{z}})(\mathbf{z}_i-\bar{\mathbf{z}})'\right]^{-1}\left\{\sum_{i=1}^N(\mathbf{z}_i-\bar{\mathbf{z}})[(\bar{y}_i-\bar{y})-(\bar{\mathbf{x}}_i-\bar{\mathbf{x}})'\boldsymbol{\beta}]\right\}, \tag{3.6.4}$$

$$\hat{\mu} = \bar{y} - \bar{\mathbf{x}}'\boldsymbol{\beta} - \bar{\mathbf{z}}'\hat{\boldsymbol{\gamma}}, \tag{3.6.5}$$

where

$$\bar{\mathbf{z}} = \frac{1}{N}\sum_{i=1}^N\mathbf{z}_i, \quad \bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^N\bar{\mathbf{x}}_i, \quad \bar{y} = \frac{1}{N}\sum_{i=1}^N\bar{y}_i.$$

[15] We use $\otimes$ to denote the Kronecker product of two matrices (Theil 1971, their Chapter 7). Suppose that $A=(a_{ij})$ is an $m\times n$ matrix and $B$ is a $p\times q$ matrix; $A\otimes B$ is defined as an $mp\times nq$ matrix of

$$\begin{bmatrix} a_{11}B & a_{12}B & \ldots & a_{1n}B \\ \vdots & & & \vdots \\ a_{m1}B & a_{m2}B & \ldots & a_{mn}B \end{bmatrix}.$$

Substituting the CV estimate of $\boldsymbol{\beta}$ into (3.6.4) and (3.6.5), we obtain estimators of $\boldsymbol{\gamma}$ and $\mu$. When $N$ tends to infinity, this two-step procedure is consistent. When $N$ is fixed and $T$ tends to infinity, $\boldsymbol{\beta}$ can still be consistently estimated by (3.2.8). But $\boldsymbol{\gamma}$ can no longer be consistently estimated, because when $N$ is fixed, we have a limited amount of information on $\alpha_i$ and $\mathbf{z}_i$. To see this, note that the OLS estimate of (3.6.3) after substituting $\operatorname{plim}_{T\to\infty}\hat{\boldsymbol{\beta}}_{cv} = \boldsymbol{\beta}$ converges to

$$\begin{aligned}\hat{\boldsymbol{\gamma}}_{\mathrm{OLS}} = \boldsymbol{\gamma} &+ \left[\sum_{i=1}^{N}(\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})'\right]^{-1}\left[\sum_{i=1}^{N}(\mathbf{z}_i - \bar{\mathbf{z}})(\alpha_i - \bar{\alpha})\right] \\ &+ \left[T\sum_{i=1}^{N}(\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})'\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{z}_i - \bar{\mathbf{z}})(u_{it} - \bar{u})\right],\end{aligned} \tag{3.6.6}$$

where

$$\bar{u} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}u_{it}, \quad \bar{\alpha} = \frac{1}{N}\sum_{i=1}^{N}\alpha_i.$$

It is clear that

$$\operatorname*{plim}_{T\to\infty} \frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i - \bar{\mathbf{z}})\frac{1}{T}\sum_{t=1}^{T}(u_{it} - \bar{u}) = 0,$$

but $(1/N)\Sigma_{i=1}^{N}(\mathbf{z}_i - \bar{\mathbf{z}})(\alpha_i - \bar{\alpha})$ is a random variable, with mean 0 and covariance $\sigma_\alpha^2[\Sigma_{i=1}^{N}(z_i - \bar{z})(z_i - \bar{z})'/N^2] \neq 0$ for finite $N$, so that the second term in (3.6.6) does not have zero plim.

When $\alpha_i$ are random and uncorrelated with $X_i$ and $Z_i$, the CV is not the BLUE. The BLUE of (3.6.1) is the GLS estimator

$$\begin{bmatrix}\hat{\mu}\\ \hat{\boldsymbol{\gamma}}\\ \hat{\boldsymbol{\beta}}\end{bmatrix} = \begin{bmatrix} NT\psi & NT\psi\bar{\mathbf{z}}' & NT\psi\bar{\mathbf{x}}' \\ NT\psi\bar{\mathbf{z}} & T\psi\sum_{i=1}^{N}\mathbf{z}_i\mathbf{z}_i' & T\psi\sum_{i=1}^{N}\mathbf{z}_i\bar{\mathbf{x}}_i' \\ NT\psi\bar{\mathbf{x}} & T\psi\sum_{i=1}^{N}\bar{\mathbf{x}}_i\mathbf{z}_i' & \sum_{i=1}^{N}X_i'QX_i + \psi T\sum_{i=1}^{N}\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i' \end{bmatrix}^{-1}$$

$$\cdot \begin{bmatrix} NT\psi\bar{y} \\ \psi T\sum_{i=1}^{N}\mathbf{z}_i\bar{y}_i \\ \sum_{i=1}^{N}X_i'Q\mathbf{y}_i + \psi T\sum_{i=1}^{N}\bar{\mathbf{x}}_i\bar{\mathbf{y}}_i \end{bmatrix} \tag{3.6.7}$$

If $\psi$ in (3.6.7) is unknown, we can substitute a consistent estimate for it. When $T$ is fixed, the GLS is more efficient than the CV. When $N$ is fixed and $T$ tends

to infinity, the GLS estimator of $\boldsymbol{\beta}$ converges to the CV estimator because $V^{-1}$ (3.3.7) converges to $\frac{1}{\sigma_u^2} Q$; for details, see Lee (1978b).

One way to view (3.6.1) is that by explicitly incorporating time-invariant explanatory variables, $\mathbf{z}_i$, we can eliminate or reduce the correlation between $\alpha_i$ and $\mathbf{x}_{it}$. However, if $\alpha_i$ remains correlated with $\mathbf{x}_{it}$ or $\mathbf{z}_i$, the GLS will be a biased estimator. The CV will produce an unbiased estimate of $\boldsymbol{\beta}$, but the OLS estimates of $\boldsymbol{\gamma}$ and $\mu$ in (3.6.3) are inconsistent even when $N$ tends to infinity if $\alpha_i$ is correlated with $\mathbf{z}_i$.[16] Thus, Hausman and Taylor (1981) suggested estimating $\boldsymbol{\gamma}$ in (3.6.3) by two-stage least squares, using those elements of $\bar{\mathbf{x}}_i$ that are uncorrelated with $\alpha_i$ as instruments for $\mathbf{z}_i$. A necessary condition to implement this method is that the number of elements of $\bar{\mathbf{x}}_i$ that are uncorrelated with $\alpha_i$ must be greater than the number of elements of $\mathbf{z}_i$ that are correlated with $\alpha_i$.

### 3.6.2 Estimation of Models with Both Individual and Time Effects

We can further generalize model (3.6.1) to include time-specific variables and effects. Let

$$\underset{1\times 1}{y_{it}} = \mu + \underset{1\times p}{\mathbf{z}_i'}\ \underset{p\times 1}{\boldsymbol{\gamma}} + \underset{1\times l}{\mathbf{r}_t'}\ \underset{l\times 1}{\boldsymbol{\rho}} + \underset{1\times K}{\mathbf{x}_{it}'}\ \underset{K\times 1}{\boldsymbol{\beta}} + \alpha_i + \lambda_t + u_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1 \ldots, T, \end{matrix} \tag{3.6.8}$$

where $\mathbf{r}_t$ and $\lambda_t$ denote $l \times 1$ and $1 \times 1$ time-specific variables and effects. Stacking (3.6.8) over $i$ and $t$, we have

$$\begin{aligned} \underset{NT\times 1}{Y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_N \end{bmatrix} &= \begin{bmatrix} \mathbf{e} & Z_1 & R & X_1 \\ \mathbf{e} & Z_2 & R & X_2 \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{e} & Z_N & R & X_N \end{bmatrix} \begin{bmatrix} \mu \\ \boldsymbol{\gamma} \\ \boldsymbol{\rho} \\ \boldsymbol{\beta} \end{bmatrix} \\ &\quad (I_N \otimes \mathbf{e})\boldsymbol{\alpha} + (\mathbf{e}_N \otimes I_T)\boldsymbol{\lambda} + \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_N \end{bmatrix}, \end{aligned} \tag{3.6.9}$$

where $\boldsymbol{\alpha}' = (\alpha_1, \ldots, \alpha_N)$, $\boldsymbol{\lambda}' = (\lambda_1 \ldots, \lambda_T)$, $R' = (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_T)$, $\mathbf{e}_N$ is an $N \times 1$ vector of ones, and $\otimes$ denotes the Kronecker product.

When both $\alpha_i$ and $\lambda_t$ are present, estimators ignoring the presence of $\lambda_t$ could be inconsistent no matter how large $N$ is if $T$ is finite. Take the simple case where $\mathbf{z}_i \equiv 0$ and $\mathbf{r}_t \equiv 0$, then the CV estimator of $\boldsymbol{\beta}$ ignoring the presence

[16] This is because $Q$ sweeps out $\alpha_i$ from (3.6.1).

of $\lambda_t$ (3.2.8) leads to

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{cv} &= \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)'\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)(y_{it}-\bar{y}_i)\right] \\ &= \boldsymbol{\beta}+\left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)'\right]^{-1} \\ &\quad\cdot\left\{\frac{1}{NT}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)(\lambda_t-\bar{\lambda})+\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)(u_{it}-\bar{u}_i)\right]\right\},\end{aligned} \tag{3.6.10}$$

where $\bar{\lambda}=\frac{1}{T}\sum_{t=1}^{T}\lambda_t$. Under the assumption that $x_{it}$ and $u_{it}$ are uncorrelated, the last term after the second equality converges to 0. But

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it}-\bar{\mathbf{x}}_i)(\lambda_t-\bar{\lambda})=\frac{1}{T}\sum_{t=1}^{T}(\bar{\mathbf{x}}_t-\bar{\mathbf{x}})(\lambda_t-\bar{\lambda}), \tag{3.6.11}$$

where $\bar{\mathbf{x}}_t=\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{it}$, $\bar{\mathbf{x}}=\frac{1}{N}\sum_{i=1}^{N}\bar{\mathbf{x}}_i=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{x}_{it}$, will converge to 0 only if $\lambda_t$ are uncorrelated with $\bar{\mathbf{x}}_t$ and $T\longrightarrow\infty$. If $\lambda_t$ is correlated with $\bar{\mathbf{x}}_t$ or even $E(\lambda_t\mathbf{x}_{it}')=0$, if $T$ is finite, (3.6.11) will not converge to 0 no matter how large $N$ is. To obtain a consistent estimator of $\boldsymbol{\beta}$, both $\alpha_i$ and $\lambda_t$ need to be considered.

If $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$ are treated as fixed constants, there is a multi-collinearity problem, for the same reasons stated for model (3.6.1). The coefficients $\boldsymbol{\alpha}$, $\boldsymbol{\lambda}$, $\boldsymbol{\gamma}$, $\boldsymbol{\rho}$, and $\mu$ cannot be separately estimated. The coefficient $\boldsymbol{\beta}$ can still be estimated by the covariance method. Using the $NT\times NT$ (covariance) transformation matrix

$$\tilde{Q}=I_{NT}-I_N\otimes\frac{1}{T}\mathbf{e}\mathbf{e}'-\frac{1}{N}\mathbf{e}_N\mathbf{e}_N'\otimes I_T+\frac{1}{NT}J, \tag{3.6.12}$$

where $J$ is an $NT\times NT$ matrix of ones, we can sweep out $\mu$, $\mathbf{z}_i$, $\mathbf{r}_t$, $\alpha_i$, and $\lambda_t$ and estimate $\boldsymbol{\beta}$ by

$$\tilde{\boldsymbol{\beta}}_{cv}=[(X_1',\ldots,X_N')\tilde{Q}(X_1',\ldots,X_N')']^{-1}[(X_1',\ldots,X_N')\tilde{Q}Y]. \tag{3.6.13}$$

In other words, $\tilde{\boldsymbol{\beta}}$ is obtained by applying the least-squares regression to the model

$$\begin{aligned}(y_{it}-\bar{y}_i-\bar{y}_t+\bar{y}) &= (\mathbf{x}_{it}-\bar{\mathbf{x}}_i-\bar{\mathbf{x}}_t+\bar{\mathbf{x}})'\boldsymbol{\beta} \\ &\quad+(u_{it}-\bar{u}_i-\bar{u}_t+\bar{u}),\end{aligned} \tag{3.6.14}$$

where $\bar{y}_t=\frac{1}{N}\sum_{i=1}^{N}y_{it}$, $\bar{y}=\frac{1}{N}\sum_{i=1}^{N}\bar{y}_i=\frac{1}{T}\sum_{t=1}^{T}\bar{y}_t=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}y_{it}$, and $\bar{u}_i$, $\bar{u}_t$, $\bar{u}$ are similarly defined.

When $u_{it}$ is independently, identically distributed with constant variance, the variance–covariance matrix of the CV estimator (3.6.13) is equal to

$$\text{Cov}\,(\hat{\boldsymbol{\beta}}_{cv})=\sigma_u^2\left[(X_1',\ldots,X_N')\tilde{Q}(X_1',\ldots,X_N')'\right]^{-1}. \tag{3.6.15}$$

To estimate $\mu$, $\boldsymbol{\gamma}$, and $\boldsymbol{\rho}$, we note that the individual-mean (over time) and time-mean (over individuals) equations are of the form

$$\bar{y}_i - \bar{\mathbf{x}}_i'\boldsymbol{\beta} = \mu_c^* + \mathbf{z}_i'\boldsymbol{\gamma} + \alpha_i + \bar{u}_i, \quad i = 1, \ldots, N, \tag{3.6.16}$$

$$\bar{y}_t - \bar{\mathbf{x}}_t'\boldsymbol{\beta} = \mu_T^* + \mathbf{r}_t'\boldsymbol{\rho} + \lambda_t + \bar{u}_t, \quad t = 1. \ldots, T, \tag{3.6.17}$$

where

$$\mu_c^* = \mu + \bar{\mathbf{r}}'\boldsymbol{\rho} + \bar{\lambda}, \tag{3.6.18}$$

$$\mu_T^* = \mu + \bar{\mathbf{z}}'\boldsymbol{\gamma} + \bar{\alpha}, \tag{3.6.19}$$

and

$$\bar{\mathbf{r}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{r}_t, \quad \bar{\mathbf{z}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{z}_i, \quad \bar{\lambda} = \frac{1}{T}\sum_{t=1}^{T}\lambda_t, \quad \bar{\alpha} = \frac{1}{N}\sum_{i=1}^{N}\alpha_i,$$

$$\bar{y}_t = \frac{1}{N}\sum_{i=1}^{N}y_{it}, \quad \bar{\mathbf{x}}_t = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{it}, \quad \bar{u}_t = \frac{1}{N}\sum_{i=1}^{N}u_{it}.$$

Replacing $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}_{cv}$, we can estimate $(\mu_c^*, \boldsymbol{\gamma}')$ and $(\mu_T^*, \boldsymbol{\rho}')$ by applying OLS to (3.6.16) and (3.6.17) over $i$ and $t$, respectively, if $\alpha_i$ and $\lambda_t$ are uncorrelated with $\mathbf{z}_i$, $\mathbf{r}_t$, and $\mathbf{x}_{it}$. To estimate $\mu$, we can substitute estimated values of $\boldsymbol{\gamma}$, $\boldsymbol{\rho}$, and $\boldsymbol{\beta}$ into any of

$$\hat{\mu} = \hat{\mu}_c^* - \bar{\mathbf{r}}'\hat{\boldsymbol{\rho}}, \tag{3.6.20}$$

$$\hat{\mu} = \hat{\mu}_T^* - \bar{\mathbf{z}}'\hat{\boldsymbol{\gamma}}, \tag{3.6.21}$$

$$\hat{\mu} = \bar{y} - \bar{\mathbf{z}}'\hat{\boldsymbol{\gamma}} - \bar{\mathbf{r}}'\hat{\boldsymbol{\rho}} - \bar{\mathbf{x}}'\hat{\boldsymbol{\beta}}, \tag{3.6.22}$$

or apply the least-squares method to the combined equations (3.6.20)–(3.6.22). When both $N$ and $T$ go to infinity, $\hat{\mu}$ is consistent.

If $\alpha_i$ and $\lambda_t$ are random, we can still estimate $\boldsymbol{\beta}$ by the CV estimator (3.6.13). However, if $\alpha_i$ and $\lambda_t$ are uncorrelated with $z_i$, $\mathbf{r}_t$, and $\mathbf{x}_{it}$, the BLUE is the GLS estimator. Assuming $\alpha_i$ and $\lambda_t$ satisfy (3.3.4), the $NT \times NT$ variance–covariance matrix of the error term, $\mathbf{u} + (I_N \otimes \mathbf{e})\boldsymbol{\alpha} + (\mathbf{e}_N \otimes I_T)\boldsymbol{\lambda}$, is

$$\tilde{V} = \sigma_u^2 I_{NT} + \sigma_\alpha^2 I_N \otimes \mathbf{e}\mathbf{e}' + \sigma_\lambda^2 \mathbf{e}_N\mathbf{e}_N' \otimes I_T. \tag{3.6.23}$$

Its inverse (Henderson 1971; Nerlove 1971b; Wallace and Hussain 1969) (see Appendix 3B) is

$$\tilde{V}^{-1} = \frac{1}{\sigma_u^2}[I_{NT} - \eta_1 I_N \otimes \mathbf{e}\mathbf{e}' - \eta_2 \mathbf{e}_N\mathbf{e}_N' \otimes I_T + \eta_3 J], \tag{3.6.24}$$

where

$$\eta_1 = \frac{\sigma_\alpha^2}{\sigma_u^2 + T\sigma_\alpha^2}, \quad \eta_2 = \frac{\sigma_\lambda^2}{\sigma_u^2 + N\sigma_\lambda^2},$$
$$\eta_3 = \frac{\sigma_\alpha^2\sigma_\lambda^2}{(\sigma_u^2 + T\sigma_\alpha^2)(\sigma_u^2 + N\sigma_\lambda^2)} \left( \frac{2\sigma_u^2 + T\sigma_\alpha^2 + N\sigma_\lambda^2}{\sigma_u^2 + T\sigma_\alpha^2 + N\sigma_\lambda^2} \right).$$

When $N \to \infty$, $T \to \infty$, and the ratio $N$ over $T$ tends to a nonzero constant, Wallace and Hussain (1969) have shown that the GLS estimator converges to the CV estimator. It should also be noted that, contrary to the conventional linear regression model without specific effects, the speed of convergence of $\boldsymbol{\beta}_{\text{GLS}}$ to $\boldsymbol{\beta}$ is $(NT)^{1/2}$, whereas the speed of convergence for $\hat{\mu}$ is $N^{1/2}$. This is because the effect of a random component can be averaged out only in the direction of that random component. For details, see Kelejian and Stephan (1983).

For the discussion of the MLE of the two-way error components models, see Baltagi (1995) and Baltagi and Li (1992).

## 3.7 HETEROSCEDASTICITY AND AUTOCORRELATION

### 3.7.1 Heteroscedasticity

So far we have confined our discussion to the assumption that the variances of the errors across individuals are identical. However, many panel studies involve cross-sectional units of varying size. In an error-components setup, heteroscedasticity can arise because the variance of $\alpha_i$, $\sigma_{\alpha i}^2$, varies with $i$ (e.g., Baltagi and Griffin 1983; Mazodier and Trognon 1978) or the variance of $u_{it}$, $\sigma_{ui}^2$, varies with $i$, or both $\sigma_{\alpha i}^2$ and $\sigma_{ui}^2$ vary with $i$. Then

$$E\mathbf{v}_i\mathbf{v}_i' = \sigma_{ui}^2 I_T + \sigma_{\alpha i}^2 \mathbf{e}\mathbf{e}' = V_i. \tag{3.7.1}$$

The $V_i^{-1}$ is of the same form as equation (3.3.5) with $\sigma_{ui}^2$ and $\sigma_{\alpha i}^2$ in place of $\sigma_u^2$ and $\sigma_\alpha^2$. The GLS estimator of $\boldsymbol{\delta}$ is obtained by replacing $V$ by $V_i$ in (3.3.7).

When $\sigma_{ui}^2$ and $\sigma_{\alpha i}^2$ are unknown, substituting the unknown true values by their estimates, a feasible (or two-step) GLS estimator can be implemented. Unfortunately, with a single realization of $\alpha_i$, there is no way one can get a consistent estimator for $\sigma_{\alpha i}^2$ even when $T \to \infty$. The conventional formula

$$\hat{\sigma}_{\alpha i}^2 = \hat{\bar{v}}_i^2 - \frac{1}{T}\hat{\sigma}_{ui}^2, \quad i = 1, \ldots, N, \tag{3.7.2}$$

where $\hat{v}_{it}$ is the initial estimate of $v_{it}$, say, the least-squares or CV estimated residual of (3.3.3), converges to $\alpha_i^2$, not $\sigma_{\alpha i}^2$. However, $\sigma_{ui}^2$ can be consistently

estimated by

$$\hat{\sigma}_{ui}^2 = \frac{1}{T-1}\sum_{t=1}^{T}(\hat{v}_{it} - \hat{\bar{v}}_i)^2, \tag{3.7.3}$$

as $T$ tends to infinity. In the event that $\sigma_{\alpha i}^2 = \sigma_\alpha^2$ for all $i$, we can estimate $\sigma_\alpha^2$ by taking the average of (3.7.2) across $i$ as their estimates.

It should be noted that when $T$ is finite, there is no way we can get consistent estimates of $\sigma_{ui}^2$ and $\sigma_{\alpha i}^2$ even when $N$ tends to infinity. This is the classical incidental parameter problem of Neyman and Scott (1948). However, if $\sigma_{\alpha i}^2 = \sigma_\alpha^2$ for all $i$, then we can get consistent estimates of $\sigma_{ui}^2$ and $\sigma_\alpha^2$ when both $N$ and $T$ tend to infinity. Substituting $\hat{\sigma}_{ui}^2$ and $\hat{\sigma}_\alpha^2$ for $\sigma_{ui}^2$ and $\sigma_\alpha^2$ in $V_i$, we obtain its estimation $\hat{V}_i$. Alternatively, one may assume that the conditional variance of $\alpha_i$ conditional on $\mathbf{x}_i$ has the same functional form across individuals, var $(\alpha_i \mid \mathbf{x}_i) = \sigma^2(\mathbf{x}_i)$, to allow for the consistent estimation of heteroscadastic variance, $\sigma_{\alpha i}^2$. The feasible GLS estimator of $\boldsymbol{\delta}$,

$$\hat{\hat{\boldsymbol{\delta}}}_{\text{FGLS}} = \left[\sum_{i=1}^{N} \tilde{X}_i' \hat{V}_i^{-1} \tilde{X}_i\right]^{-1} \left[\sum_{i=1}^{N} \tilde{X}_i' \hat{V}_i^{-1} \mathbf{y}_i\right] \tag{3.7.4}$$

is asymptotically equivalent to the GLS estimator when both $N$ and $T$ approach infinity. The asymptotic variance–covariance matrix of the $\hat{\hat{\boldsymbol{\delta}}}_{\text{FGLS}}$ can be approximated by $(\sum_{i=1}^{N} \tilde{X}_i' \hat{V}_i^{-1} \tilde{X}_i)^{-1}$.

In the case that both $\sigma_{\alpha i}^2$ and $\sigma_{ui}^2$ vary across $i$, another way to estimate the model is to treat $\alpha_i$ as fixed by taking the covariance transformation to eliminate the effect of $\alpha_i$, then apply the feasible weighted least-squares method. That is, we first weigh each individual observation by the inverse of $\sigma_{ui}$, $\mathbf{y}_i^* = \frac{1}{\sigma_{ui}}\mathbf{y}_i$, $X_i^* = \frac{1}{\sigma_{ui}}X_i$ and then apply the CV estimator to the transformed data

$$\hat{\boldsymbol{\beta}}_{cv} = \left[\sum_{i=1}^{N} X_i^{*'} Q X_i^*\right]^{-1} \left[\sum_{i=1}^{N} X_i^{*'} Q \mathbf{y}_i^*\right]. \tag{3.7.5}$$

### 3.7.2 Models with Serially Correlated Errors

The fundamental assumption we made with regard to the variable-intercept model was that the error term is serially uncorrelated conditional on the individual effects $\alpha_i$. But there are cases in which the effects of unobserved variables vary systematically over time, such as the effect of serially correlated omitted variables or the effects of transitory variables whose effects last more than one period. The existence of these variables is not well described by an error term that is either constant or independently distributed over time periods. To provide for a more general autocorrelation scheme, one can relax the restriction that $u_{it}$ are serially uncorrelated (e.g., Lillard and Weiss 1979; Lillard and Willis

1978).[17] Anderson and Hsiao (1982) have considered the MLE of the model (3.3.5) with $u_{it}$ following a first-order autoregressive process,

$$u_{it} = \rho u_{i,t-1} + \epsilon_{it}, \tag{3.7.6}$$

where $\epsilon_{it}$ are independently, identically distributed, with 0 mean and variance $\sigma_\epsilon^2$. However, computation of the MLE is complicated. But if we know $\rho$, we can transform the model into a standard variance–components model,

$$y_{it} - \rho y_{i,t-1} = \mu(1-\rho) + \boldsymbol{\beta}'(\mathbf{x}_{it} - \rho \mathbf{x}_{i,t-1}) + (1-\rho)\alpha_i + \epsilon_{it}. \tag{3.7.7}$$

Therefore, we can obtain an asymptotically efficient estimator of $\boldsymbol{\beta}$ by the following multistep procedure:

Step 1. Eliminate the individual effect $\alpha_i$ by subtracting the individual mean from (3.3.5). We have

$$y_{it} - \bar{y}_i = \boldsymbol{\beta}'(\mathbf{x}_{it} - \bar{\mathbf{x}}_i) + (u_{it} - \bar{u}_i). \tag{3.7.8}$$

Step 2. Use the least-squares residual of (3.7.8) to estimate the serial correlation coefficient $\rho$, or use the Durbin (1960) method by regressing $(y_{it} - \bar{y}_i)$ on $(y_{i,t-1} - \bar{y}_{i,-1})$, and $(\mathbf{x}_{i,t-1} - \bar{\mathbf{x}}_{i,-1})$, and treat the coefficient of $(y_{i,t-1} - \bar{y}_{i,-1})$ as the estimated value of $\rho$, where $\bar{y}_{i,-1} = (1/T)\Sigma_{t=1}^{T} y_{i,t-1}$ and $\bar{\mathbf{x}}_{i,-1} = (1/T)\Sigma_{t=1}^{T}\mathbf{x}_{i,t-1}$. (For simplicity, we assume that $y_{i0}$ and $x_{i0}$ are observable.)

Step 3. Estimate $\sigma_\epsilon^2$ and $\sigma_\alpha^2$ by

$$\begin{aligned}\hat{\sigma}_\epsilon^2 =& \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\{(y_{it} - \bar{y}_i) \\ & - \hat{\rho}(y_{i,t-1} - \bar{y}_{i,-1}) \\ & - \hat{\boldsymbol{\beta}}'[(\mathbf{x}_{it} - \bar{\mathbf{x}}_i) - (\mathbf{x}_{i,t-1} - \bar{\mathbf{x}}_{i,-1})\hat{\rho}]\}^2\end{aligned} \tag{3.7.9}$$

and

$$\begin{aligned}\hat{\sigma}_\alpha^2 =& \frac{1}{(1-\hat{\rho})^2}\left\{\frac{1}{N}\sum_{i=1}^{N}[\bar{y}_i - \hat{\mu}(1-\hat{\rho}) \right. \\ & \left. - \hat{\rho}\bar{y}_{i,-1} - \hat{\boldsymbol{\beta}}'(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_{i,-1}\hat{\rho})]^2 - \frac{1}{T}\hat{\sigma}_\epsilon^2\right\}.\end{aligned} \tag{3.7.10}$$

Step 4. Substituting $\hat{\rho}$, (3.7.9), and (3.7.10) for $\rho$, $\sigma_\epsilon^2$, and $\sigma_\alpha^2$ in the variance–covariance matrix of $\epsilon_{it} + (1-\rho)\alpha_i$, we estimate (3.7.7) by the feasible GLS method.

The above multistep or feasible generalized least-squares procedure treats the initial $u_{i1}$ as fixed constants. A more efficient, but computationally more

[17] See Li and Hsiao (1998) for a test of whether the serial correlation in the error is caused by an individual-specific time invariant component or by the inertia in the shock and Hong and Kao (2004) for testing of serial correlation of unknown form.

burdensome feasible GLS is to treat initial $u_{i1}$ as random variables with mean 0 and variance $\frac{\sigma_\epsilon^2}{1-\rho^2}$ (e.g., Baltagi and Li 1991). Premultiplying (3.3.5) by the $T \times T$ transformation matrix

$$R = \begin{pmatrix} (1-\rho^2)^{1/2} & 0 & 0 & \cdot & \cdot & 0 \\ -\rho & 1 & 0 & \cdot & \cdot & \cdot \\ 0 & -\rho & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & -\rho & 1 \end{pmatrix},$$

transforms $\mathbf{u}_i$ into serially uncorrelated homoscedastic error terms, but also transforms $\mathbf{e}_T\alpha_i$ into $(1-\rho)\boldsymbol{\ell}_T\alpha_i$, where $\boldsymbol{\ell}_T = [(\frac{1+\rho}{1-\rho})^{1/2}, 1, \ldots, 1]'$. Therefore, the transformed error terms will have covariance matrix

$$V^* = \sigma_\epsilon^2 I_T + (1-\rho)^2\sigma_\alpha^2 \boldsymbol{\ell}_T \boldsymbol{\ell}_T', \tag{3.7.11}$$

with inverse

$$V^{*-1} = \frac{1}{\sigma_\epsilon^2}[I_T - \frac{(1-\rho)^2\sigma_\alpha^2}{[T-(T-1)\rho-\rho^2]\sigma_\alpha^2 + \sigma_\epsilon^2}\boldsymbol{\ell}_T\boldsymbol{\ell}_T']. \tag{3.7.12}$$

Substituting initial estimates of $\rho$, $\sigma_\alpha^2$, and $\sigma_\epsilon^2$ into (3.7.12), one can apply the GLS procedure using (3.7.12) to estimate $\boldsymbol{\delta}$.

When $T$ tends to infinity, the GLS estimator of $\boldsymbol{\beta}$ converges to the covariance estimator of the transformed model (3.7.7). In other words, an asymptotically efficient estimator of $\boldsymbol{\beta}$ is obtained by finding a consistent estimate of $\rho$, transforming the model to eliminate the serial correlation, and then applying the covariance method to the transformed model (3.7.7).

MaCurdy (1982) has considered a similar estimation procedure for (3.3.5) with a more general time series process of $u_{it}$. His procedure essentially involves eliminating $\alpha_i$ by first differencing and treating $y_{it} - y_{i,t-1}$ as the dependent variable. He then modeled the variance–covariance matrix of $\mathbf{u}_i$ by using a standard Box–Jenkins (1970) type of procedure to model the least-squares predictor of $u_{it} - u_{i,t-1}$, and estimated the parameters by an efficient algorithm.

Kiefer (1980) considered estimation of fixed-effects models of (3.2.1) with arbitrary intertemporal correlations for $u_{it}$. When $T$ is fixed, the individual effects cannot be estimated consistently. He suggested that we first eliminate the individual effects by transforming the model to the form (3.7.8) using the transformation matrix $Q = I_T - (1/T)\mathbf{e}\mathbf{e}'$. Then estimate the intertemporal variance–covariance matrix of $Q\mathbf{u}_i$ by

$$\hat{\Sigma}^* = \frac{1}{N}\sum_{i=1}^{N}[Q(\mathbf{y}_i - X_i\hat{\boldsymbol{\beta}})][Q(\mathbf{y}_i - X_i\hat{\boldsymbol{\beta}})]', \tag{3.7.13}$$

where $\hat{\boldsymbol{\beta}}$ is any arbitrary consistent estimator of $\boldsymbol{\beta}$ (e.g., CV of $\boldsymbol{\beta}$). Given an estimate of $\hat{\Sigma}^*$ one can estimate $\boldsymbol{\beta}$ by the GLS method,

$$\boldsymbol{\beta}^* = \left[\sum_{i=1}^{N} X_i' Q \hat{\Sigma}^{*-} Q X_i\right]^{-1} \left[\sum_{i=1}^{N} X_i' Q \hat{\Sigma}^{*-} Q \mathbf{y}_i\right], \tag{3.7.14}$$

where $\hat{\Sigma}^{*-}$ is a generalized inverse of $\Sigma^*$, because $\Sigma^*$ has only rank $T-1$. The asymptotic variance–covariance matrix of $\hat{\boldsymbol{\beta}}^*$ is

$$\text{Var}\,(\hat{\boldsymbol{\beta}}^*) = \left[\sum_{i=1}^{N} X_i' Q \hat{\Sigma}^{*-} Q X_i\right]^{-1}. \tag{3.7.15}$$

Although any generalized inverse can be used for $\hat{\Sigma}^*$, a particularly attractive choice is

$$\hat{\Sigma}^{*-} = \begin{bmatrix} \hat{\Sigma}_{T-1}^{*-1} & \mathbf{0} \\ \mathbf{0}' & 0 \end{bmatrix}, \tag{3.7.16}$$

where $\hat{\Sigma}_{T-1}^*$ is the $(T-1)\times(T-1)$ full-rank submatrix of $\hat{\Sigma}^*$ obtained by deleting the last row and column from $\hat{\Sigma}^*$. Using this generalized inverse simply amounts to deleting the $T$th observation from the transformed observations $Q\mathbf{y}_i$ and $QX_i$, and then applying GLS to the remaining subsample. However, it should be noted that this is not the GLS estimator that would be used if the variance–covariance matrix of $\mathbf{u}_i$ were known.

### 3.7.3 Heteroscedasticity Autocorrelation Consistent Estimator for the Covariance Matrix of the CV Estimator

The previous two subsections discuss the estimation procedures when the patterns of heteroscedasticity or serial correlations are known. In the case that the errors $u_{it}$ have unknown heteroscedasticity (across individuals and over time) and/or autocorrelation patterns, one may still use the covariance estimator (3.2.5) or (3.6.13) to obtain a consistent estimate of $\boldsymbol{\beta}$. However, the covariance matrix of the CV estimator of $\boldsymbol{\beta}$ no longer has the form (3.2.11) or $\sigma_u^2(X'\tilde{Q}X)^{-1}$, where $X' = (X_1', \ldots, X_N')$. For instance, when $u_{it}$ has heteroscedasticity of unknown form, $\sqrt{NT}(\hat{\boldsymbol{\beta}}_{cv} - \boldsymbol{\beta})$ is asymptotically normally distributed with mean 0 and covariance matrix of the form (e.g., Arellano 2003)

$$\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}_{it}'\right)^{-1} \Omega \left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}_{it}'\right)^{-1}, \tag{3.7.17}$$

where

$$\tilde{\mathbf{x}}_{it} = \mathbf{x}_{it} - \bar{\mathbf{x}}_i, \tag{3.7.18}$$

for model (3.2.1) and

$$\tilde{\mathbf{x}}_{it} = \mathbf{x}_{it} - \bar{\mathbf{x}}_i - \bar{\mathbf{x}}_t + \bar{\mathbf{x}} \tag{3.7.19}$$

for model (3.6.8), and

$$\Omega = \frac{1}{T}\sum_{t=1}^{T} E\left(\tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}'_{it}u_{it}^2\right). \tag{3.7.20}$$

It is shown by Stock and Watson (2008) that

$$\hat{\Omega} = \left(\frac{T-1}{T-2}\right)\left\{\frac{1}{NT-N-K}\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}'_{it}\hat{\tilde{u}}_{it}^2 - \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}'_{it}\right)\left(\frac{1}{T-1}\sum_{s=1}^{T}\hat{\tilde{u}}_{is}^2\right)\right\}, \tag{3.7.21}$$

is a consistent estimator of $\Omega$ for any sequence of $N$ or $T \longrightarrow \infty$. Where $\hat{\tilde{u}}_{it} = \tilde{y}_{it} - \tilde{\mathbf{x}}'_{it}\hat{\boldsymbol{\beta}}_{cv}$, and $\tilde{y}_{it} - \bar{y}_i$ or $\tilde{y}_{it} = y_{it} - \bar{y}_i - \bar{y}_t + \bar{y}$.

When both $N$ and $T$ are large, Vogelsang (2012) suggests a robust estimator of the variance–covariance matrix of the CV estimator of $\boldsymbol{\beta}$ as

$$T\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}'_{it}\right)^{-1}\left(\sum_{i=1}^{N}\hat{\Omega}_i\right)\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}'_{it}\right)^{-1}, \tag{3.7.22}$$

where

$$\hat{\Omega}_i = \frac{1}{T}\left[\sum_{t=1}^{T}\hat{u}_{it}^2\tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}'_{it} + \sum_{t=2}^{T}\sum_{j=1}^{t-1}k\left(\frac{j}{m}\right)\hat{u}_{it}\hat{u}_{i,t-j}\left(\tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}'_{i,t-j} + \tilde{\mathbf{x}}_{i,t-j}\tilde{\mathbf{x}}'_{it}\right)\right], \tag{3.7.23}$$

where $k(\frac{j}{m})$ denotes the kernel such that $k(\frac{j}{m}) = 1 - \frac{j}{m}$ if $|\ \frac{j}{m}\ | \leq 1$ and $k(\frac{j}{m}) = 0$ if $|\ \frac{j}{m}\ | \geq 1$. If $M = T$, then all the sample autocorrelations are used for (3.7.23). If $M < T$, a truncated kernel is used.

## 3.8 MODELS WITH ARBITRARY ERROR STRUCTURE – CHAMBERLAIN $\pi$-APPROACH

The focus of this chapter is formulation and estimation of linear regression models when there exist time-invariant and/or individual-invariant omitted (latent) variables. In Sections 3.1–3.7 we have been assuming that the variance–covariance matrix of the error term possesses a known structure. In fact, when $N$ tends to infinity, the characteristics of short panels allow us to exploit the unknown structure of the error process. Chamberlain (1982, 1984) has proposed treating each period as an equation in a multivariate setup to transform the

problems of estimating a single-equation model involving two dimensions (cross sections and time series) into a one-dimensional problem of estimating a $T$-variate regression model with cross-sectional data. This formulation avoids imposing restrictions a priori on the variance–covariance matrix, so that serial correlation and certain forms of heteroscedasticity in the error process, which covers certain kinds of random-coefficient models (see Chapter 6), can be incorporated. The multivariate setup also provides a link between the single-equation and simultaneous-equations models (see Chapter 5). Moreover, the extended view of the Chamberlain method can also be reinterpreted in terms of the generalized method of moments (GMM) to be discussed in Chapter 4 (Crépon and Mairesse 1996).

For simplicity, consider the following model:

$$y_{it} = \alpha_i^* + \boldsymbol{\beta}'\mathbf{x}_{it} + u_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{matrix} \tag{3.8.1}$$

and

$$E(u_{it} \mid \mathbf{x}_{i1} \ldots, \mathbf{x}_{iT}, \alpha_i^*) = 0. \tag{3.8.2}$$

When $T$ is fixed and $N$ tends to infinity, we can stack the $T$ time period observations of the $i$th individual's characteristics into a vector $(\mathbf{y}_i', \mathbf{x}_i')$, where $\mathbf{y}_i' = (y_{i1}, \ldots, y_{iT})$ and $\mathbf{x}_i' = (\mathbf{x}_{i1}' \ldots, \mathbf{x}_{iT}')$ are $1 \times T$ and $1 \times KT$ vectors, respectively. We assume that $(\mathbf{y}_i', \mathbf{x}_i')$ is an independent draw from a common (unknown) multivariate distribution function with finite fourth-order moments and with $E\mathbf{x}_i\mathbf{x}_i' = \Sigma_{xx}$ positive definite. Then each individual observation vector corresponds to a $T$-variate regression

$$\underset{T \times 1}{\mathbf{y}_i} = \mathbf{e}\alpha_i^* + (I_T \otimes \boldsymbol{\beta}')\mathbf{x}_i + \mathbf{u}_i, \quad i = 1 \ldots, N. \tag{3.8.3}$$

To allow for the possible correlation between $\alpha_i^*$ and $\mathbf{x}_i$, Chamberlain, following the idea of Mundlak (1978), assumes that

$$E(\alpha_i^* \mid \mathbf{x}_i) = \mu + \sum_{t=1}^{T} \mathbf{a}_t'\mathbf{x}_{it} = \mu + \mathbf{a}'\mathbf{x}_i, \tag{3.8.4}$$

where $\mathbf{a}' = (\mathbf{a}_1' \ldots, \mathbf{a}_T')$. While $E(\mathbf{y}_i \mid \mathbf{x}_i, \alpha_i^*)$ is assumed linear, it is possible to relax the assumption of $E(\alpha_i^* \mid \mathbf{x}_i)$ being linear for the linear model. In the case in which $E(\alpha_i^* \mid \mathbf{x}_i)$ is not linear, Chamberlain (1984) replaces (3.8.4) by

$$E^*(\alpha_i^* \mid \mathbf{x}_i) = \mu + \mathbf{a}'\mathbf{x}_i, \tag{3.8.5}$$

where $E^*(\alpha_i^* \mid \mathbf{x}_i)$ refers to the (minimum mean square error) linear predictor (or the projection) of $\alpha_i^*$ onto $\mathbf{x}_i$. Then,[18]

$$\begin{aligned} E^*(\mathbf{y}_i \mid \mathbf{x}_i) &= E^*\{E^*(\mathbf{y}_i \mid \mathbf{x}_i, \alpha_i^*) \mid \mathbf{x}_i\} \\ &= E^*\{\mathbf{e}\alpha_i^* + (I_T \otimes \boldsymbol{\beta}')\mathbf{x}_i \mid \mathbf{x}_i\} \\ &= \mathbf{e}\mu + \Pi\mathbf{x}_i, \end{aligned} \tag{3.8.6}$$

where

$$\underset{T\times KT}{\Pi} \quad = \quad I_T \otimes \boldsymbol{\beta}' + \mathbf{e}\mathbf{a}'. \tag{3.8.7}$$

Rewrite equations (3.8.3) and (3.8.6) as

$$\mathbf{y}_i = \mathbf{e}\mu + [I_T \otimes \mathbf{x}_i']\boldsymbol{\pi} + \boldsymbol{\nu}_i, \quad i = 1\ldots, N, \tag{3.8.8}$$

where $\boldsymbol{\nu}_i = \mathbf{y}_i - E^*(\mathbf{y}_i \mid \mathbf{x}_i)$ and $\boldsymbol{\pi}' = \text{vec}\,(\Pi)' = [\boldsymbol{\pi}_1', \ldots, \boldsymbol{\pi}_T']$ is a $1 \times KT^2$ vector with $\boldsymbol{\pi}_t'$ denoting the $t$th row of $\Pi'$. Treating the coefficients of (3.8.8) as if they were unconstrained, we regress $(\mathbf{y}_i - \bar{\mathbf{y}}^*)$ on $[I_T \otimes (\mathbf{x}_i - \bar{\mathbf{x}}^*)']$ and obtain the least-squares estimate of $\boldsymbol{\pi}$ as[19]

$$\begin{aligned} \hat{\boldsymbol{\pi}} &= \left\{\sum_{i=1}^N [I_T \otimes (\mathbf{x}_i - \bar{\mathbf{x}}^*)][I_T \otimes (\mathbf{x}_i - \bar{\mathbf{x}}^*)']\right\}^{-1} \\ &\quad \cdot \left\{\sum_{i=1}^N [I_T \otimes (\mathbf{x}_i - \bar{\mathbf{x}}^*)](\mathbf{y}_i - \bar{\mathbf{y}}^*)\right\} \\ &= \boldsymbol{\pi} + \left\{\frac{1}{N}\sum_{i=1}^N [I_T \otimes (\mathbf{x}_i - \bar{\mathbf{x}}^*)][I_T \otimes (\mathbf{x}_i - \bar{\mathbf{x}}^*)']\right\}^{-1} \\ &\quad \cdot \left\{\frac{1}{N}\sum_{i=1}^N [I_T \otimes (\mathbf{x}_i - \bar{\mathbf{x}}^*)]\boldsymbol{\nu}_i\right\}, \end{aligned} \tag{3.8.9}$$

where $\bar{\mathbf{y}}^* = (1/N)\Sigma_{i=1}^N \mathbf{y}_i$ and $\bar{\mathbf{x}}^* = (1/N)\Sigma_{i=1}^N \mathbf{x}_i$.

By construction, $E(\boldsymbol{\nu}_i \mid \mathbf{x}_i) = 0$, and $E(\boldsymbol{\nu}_i \otimes \mathbf{x}_i) = 0$. The law of large numbers implies that $\hat{\boldsymbol{\pi}}$ is a consistent estimator of $\boldsymbol{\pi}$ when $T$ is fixed and $N$ tends to infinity (Rao 1973, Chapter 2). Moreover, because

$$\begin{aligned} \underset{N\to\infty}{\text{plim}} \quad \tfrac{1}{N}\sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}}^*)(\mathbf{x}_i - \bar{\mathbf{x}}^*)' &= E[\mathbf{x}_i - E\mathbf{x}_i][\mathbf{x}_i - E\mathbf{x}_i]' \\ &= \Sigma_{xx} - (E\mathbf{x})(E\mathbf{x})' = \Phi_{xx}, \end{aligned}$$

[18] If $E(\alpha_i^* \mid \mathbf{x}_i)$ is linear, $E^*(\mathbf{y}_i \mid \mathbf{x}_i) = E(\mathbf{y}_i \mid \mathbf{x}_i)$.

[19] Of course, we can obtain the least-squares estimate of $\pi$ by imposing the restriction that all $T$ equations have identical intercepts $\mu$. But this only complicates the algebraic equation of the least-squares estimate without a corresponding gain in insight.

we have $\sqrt{N}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})$ converging in distribution to (Rao 1973, Chapter 2)

$$\begin{aligned}&\left[I_T \otimes \Phi_{xx}^{-1}\right]\left\{\frac{1}{\sqrt{N}}\sum_{i=1}^{N}[I_T \otimes (\mathbf{x}_i - \bar{\mathbf{x}}^*)]\boldsymbol{\nu}_i\right\}\\ &\quad = \left[I_T \otimes \Phi_{xx}^{-1}\right]\left\{\frac{1}{\sqrt{N}}\sum_{i=1}^{N}[\boldsymbol{\nu}_i \otimes (\mathbf{x}_i - \bar{\mathbf{x}}^*)]\right\}.\end{aligned} \tag{3.8.10}$$

So the central-limit theorem implies that $\sqrt{N}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})$ is asymptotically normally distributed, with mean 0 and variance–covariance matrix $\Omega$, where[20]

$$\begin{aligned}\Omega =& E[(\mathbf{y}_i - \mathbf{e}\mu - \Pi\mathbf{x}_i)(\mathbf{y}_i - \mathbf{e}\mu - \Pi\mathbf{x}_i)'\\ &\otimes \Phi_{xx}^{-1}(\mathbf{x}_i - E\mathbf{x})(\mathbf{x}_i - E\mathbf{x})'\Phi_{xx}^{-1}].\end{aligned} \tag{3.8.11}$$

A consistent estimator of $\Omega$ is readily available from the corresponding sample moments,

$$\begin{aligned}\hat{\Omega} =& \frac{1}{N}\sum_{i=1}^{N}\left\{\left[(\mathbf{y}_i - \bar{\mathbf{y}}^*) - \hat{\Pi}(\mathbf{x}_i - \bar{\mathbf{x}}^*)\right][(\mathbf{y}_i - \bar{\mathbf{y}}^*)\right.\\ &\left. - \hat{\Pi}(\mathbf{x}_i - \bar{\mathbf{x}}^*)]' \otimes S_{xx}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}^*)(\mathbf{x}_i - \bar{\mathbf{x}}^*)'S_{xx}^{-1}\right\},\end{aligned} \tag{3.8.12}$$

where

$$S_{xx} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}}^*)(\mathbf{x}_i - \bar{\mathbf{x}}^*)'.$$

Equation (3.8.7) implies that $\Pi$ is subject to restrictions. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \mathbf{a}')$. We specify the restrictions on $\Pi$ [equation (3.8.7)] by the conditions that

$$\boldsymbol{\pi} = \mathbf{f}(\boldsymbol{\theta}). \tag{3.8.13}$$

We can impose these restrictions by using a minimum-distance estimator. Namely, choose $\boldsymbol{\theta}$ to minimize

$$[\hat{\boldsymbol{\pi}} - \mathbf{f}(\boldsymbol{\theta})]'\hat{\Omega}^{-1}[\hat{\boldsymbol{\pi}} - \mathbf{f}(\boldsymbol{\theta})]. \tag{3.8.14}$$

Under the assumptions that $\mathbf{f}$ possesses continuous second partial derivatives and the matrix of first partial derivatives

$$F = \frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}'} \tag{3.8.15}$$

has full column rank in an open neighborhood containing the true parameter $\boldsymbol{\theta}$, the minimum-distance estimator of (3.8.14), $\hat{\boldsymbol{\theta}}$, is consistent, and $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$,

[20] For details, see White (1980) or Chamberlain (1982).

is asymptotically normally distributed, with mean 0 and variance–covariance matrix

$$(F'\Omega^{-1}F)^{-1}. \tag{3.8.16}$$

The quadratic form

$$N[\hat{\boldsymbol{\pi}} - \mathbf{f}(\boldsymbol{\theta})]'\hat{\Omega}^{-1}[\hat{\boldsymbol{\pi}} - \mathbf{f}(\boldsymbol{\theta})] \tag{3.8.17}$$

converges to a $\chi^2$ distribution, with $KT^2 - K(1+T)$ degrees of freedom.[21]

The advantage of the multivariate setup is that we need only to assume that the $T$ period observations of the characteristics of the $i$th individual are independently distributed across cross-sectional units with finite fourth-order moments. We do not need to make specific assumptions about the error process. Nor do we need to assume that $E(\alpha_i^* \mid \mathbf{x}_i)$ is linear.[22] In the more restrictive case that $E(\alpha_i^* \mid \mathbf{x}_i)$ is indeed linear, [then the regression function is linear, that is, $E(\mathbf{y}_i \mid \mathbf{x}_i) = \mathbf{e}\mu + \Pi\mathbf{x}_i$], and $\text{Var}(\mathbf{y}_i \mid \mathbf{x}_i)$ is uncorrelated with $\mathbf{x}_i\mathbf{x}_i'$, (3.8.12) will converge to

$$E[\text{Var}(\mathbf{y}_i \mid \mathbf{x}_i)] \otimes \Phi_{xx}^{-1}. \tag{3.8.18}$$

If the conditional variance–covariance matrix is homoscedastic, so that $\text{Var}(\mathbf{y}_i \mid \mathbf{x}_i) = \Sigma$ does not depend on $\mathbf{x}_i$, then (3.8.12) will converge to

$$\Sigma \otimes \Phi_{xx}^{-1}. \tag{3.8.19}$$

The Chamberlain procedure of combining all $T$ equations for a single individual into one system, obtaining the matrix of unconstrained linear-predictor coefficients and then imposing restrictions by using a minimum-distance estimator, also has a direct analog in the linear simultaneous-equations model, in which an efficient estimator is provided by applying a minimum-distance procedure to the reduce form (Malinvaud 1970, Chapter 19). We demonstrate this by considering the standard simultaneous-equations model for the time series data,[23]

$$\Gamma\mathbf{y}_t + B\mathbf{x}_t = \mathbf{u}_t, \quad t = 1\ldots., T, \tag{3.8.20}$$

and its reduced form

$$\mathbf{y}_t = \Pi\mathbf{x}_t + \mathbf{v}_t, \quad \Pi = -\Gamma^{-1}B, \quad \mathbf{v}_t = \Gamma^{-1}\mathbf{u}_t, \tag{3.8.21}$$

where $\Gamma$, $B$, and $\Pi$ are $G \times G$, $G \times K$, and $G \times K$ matrices of coefficients, $\mathbf{y}_t$ and $\mathbf{u}_t$ are $G \times 1$ vectors of observed endogenous variables and unobserved disturbances, respectively, and $\mathbf{x}_t$ is a $K \times 1$ vector of observed exogenous variables. The $\mathbf{u}_t$ is assumed to be serially independent, with bounded variances and covariances.

[21] For proof, see Appendix 3A, Chamberlain (1982), Chiang (1956), or Malinvaud (1970).

[22] If $E(\alpha_i^* \mid \mathbf{x}_i) \neq E^*(\alpha_i^* \mid \mathbf{x}_i)$, then there will be heteroscedasticity, because the residual will contain $E(\alpha_i^* \mid \mathbf{x}_i) - E^*(\alpha_i^* \mid \mathbf{x}_i)$.

[23] For fitting model (3.8.20) to panel data, see Chapter 5.

In general, there are restrictions on $\Gamma$ and $B$. We assume that the model (3.8.20) is identified by zero restrictions (e.g., Hsiao 1983) so that the $g$th structural equation is of the form

$$y_{gt} = \mathbf{w}_{gt}'\boldsymbol{\theta}_g + v_{gt}, \tag{3.8.22}$$

where the components of $\mathbf{w}_{gt}$ are the variables in $\mathbf{y}_t$ and $\mathbf{x}_t$ that appear in the $g$th equation with unknown coefficients. Let $\Gamma(\boldsymbol{\theta})$ and $B(\boldsymbol{\theta})$ be parametric representations of $\Gamma$ and $B$ that satisfy the zero restrictions and the normalization rule, where $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_G')$. Then $\boldsymbol{\pi} = \mathbf{f}(\boldsymbol{\theta}) = \text{vec}\{[-\Gamma^{-1}(\boldsymbol{\theta})B(\boldsymbol{\theta})]'\}$.

Let $\hat{\Pi}$ be the least-squares estimator of $\Pi$, and

$$\tilde{\Omega} = \frac{1}{T}\sum_{t=1}^{T}\left[(\mathbf{y}_t - \hat{\Pi}\mathbf{x}_t)(\mathbf{y}_t - \hat{\Pi}\mathbf{x}_t)' \otimes S_x^{*-1}(\mathbf{x}_t\mathbf{x}_t')S_x^{*-1}\right], \tag{3.8.23}$$

where $S_x^* = (1/T)\Sigma_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'$. The generalization of the Malinvaud (1970) minimum-distance estimator is to choose $\hat{\boldsymbol{\theta}}$ to

$$\min\ [\hat{\boldsymbol{\pi}} - \mathbf{f}(\boldsymbol{\theta})]'\tilde{\Omega}^{-1}[\hat{\boldsymbol{\pi}} - \mathbf{f}(\boldsymbol{\theta})]. \tag{3.8.24}$$

Then we have $\sqrt{T}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ being asymptotically normally distributed, with mean 0 and variance–covariance matrix $(F'\tilde{\Omega}^{-1}F)^{-1}$, where $F = \partial\,\mathbf{f}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$.

The formula for $\partial\boldsymbol{\pi}/\partial\boldsymbol{\theta}'$ is given in Rothenberg (1973, p. 69):

$$F = \frac{\partial\boldsymbol{\pi}}{\partial\boldsymbol{\theta}'} = -(\Gamma^{-1} \otimes I_K)\left[\Sigma_{wx}\left(I_G \otimes \Sigma_{xx}^{-1}\right)\right]', \tag{3.8.25}$$

where $\Sigma_{wx}$ is block-diagonal: $\Sigma_{wx} = \text{diag}\{E(\mathbf{w}_{1t}\mathbf{x}_t'), \ldots, E(\mathbf{w}_{Gt}\mathbf{x}_t')\}$ and $\Sigma_{xx} = E(\mathbf{x}_t\mathbf{x}_t')$. So we have

$$(F'\tilde{\Omega}^{-1}F)^{-1} = \{\Sigma_{wx}[E(\mathbf{u}_t\mathbf{u}_t' \otimes \mathbf{x}_t\mathbf{x}_t')]^{-1}\Sigma_{wx}'\}^{-1}, \tag{3.8.26}$$

If $\mathbf{u}_t\mathbf{u}_t'$ is uncorrelated with $\mathbf{x}_t\mathbf{x}_t'$, then (3.8.26) reduces to

$$\{\Sigma_{wx}\left[[E(\mathbf{u}_t\mathbf{u}_t')]^{-1} \otimes \Sigma_{xx}^{-1}\right]\Sigma_{xw}'\}^{-1}, \tag{3.8.27}$$

which is the conventional asymptotic covariance matrix for the three-stage least-squares (3SLS) estimator (Zellner and Theil 1962). If $\mathbf{u}_t\mathbf{u}_t'$ is correlated with $\mathbf{x}_t\mathbf{x}_t'$, then the minimum-distance estimator of $\hat{\boldsymbol{\theta}}$ is asymptotically equivalent to the Chamberlain (1982) generalized 3SLS estimator,

$$\hat{\boldsymbol{\theta}}_{G3\text{SLS}} = (S_{wx}\hat{\Psi}^{-1}S_{wx}')^{-1}(S_{wx}\hat{\Psi}^{-1}\mathbf{s}_{xy}), \tag{3.8.28}$$

where

$$S_{wx} = \text{diag}\left\{\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_{1t}\mathbf{x}_t', \ldots, \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_{Gt}\mathbf{x}_t'\right\},$$

$$\hat{\Psi} = \frac{1}{T}\sum_{t=1}^{T}\{\hat{\mathbf{u}}_t\hat{\mathbf{u}}_t' \otimes \mathbf{x}_t\mathbf{x}_t'\}, \quad \mathbf{s}_{xy} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t \otimes \mathbf{x}_t,$$

and

$$\hat{\mathbf{u}}_t = \hat{\Gamma}\mathbf{y}_t + \hat{B}\mathbf{x}_t,$$

where $\hat{\Gamma}$ and $\hat{B}$ are any consistent estimators for $\Gamma$ and $B$. When certain equations are exactly identified, then just as in the conventional 3SLS case, applying the generalized 3SLS estimator to the system of equations, excluding the exactly identified equations, yields the same asymptotic covariance matrix as the estimator obtained by applying the generalized 3SLS estimator to the full set of $G$ equations.[24]

However, as with any generalization, there is a cost associated with it. The minimum-distance estimator is efficient only relative to the class of estimators that do not impose a priori restrictions on the variance–covariance matrix of the error process. If the error process is known to have an error-component structure, as assumed in previous sections, the least-squares estimate of $\Pi$ is not efficient (see Section 5.2), and hence the minimum-distance estimator, ignoring the specific structure of the error process, cannot be efficient, although it remains consistent.[25] The efficient estimator is the GLS estimator. Moreover, computation of the minimum-distance estimator can be quite tedious, whereas the two-step GLS estimation procedure is fairly easy to implement.

## APPENDIX 3A: CONSISTENCY AND ASYMPTOTIC NORMALITY OF THE MINIMUM-DISTANCE ESTIMATOR

In this appendix we briefly sketch the proof of consistency and asymptotic normality of the minimum-distance estimator.[26] For completeness we shall state the set of conditions and properties that they imply in general forms.

Let

$$S_N = [\hat{\boldsymbol{\pi}}_N - \mathbf{f}(\boldsymbol{\theta})]' A_N [\hat{\boldsymbol{\pi}}_N - \mathbf{f}(\boldsymbol{\theta})]. \tag{3A.1}$$

**Assumption 3A.1:** The vector $\hat{\boldsymbol{\pi}}_N$ converges to $\boldsymbol{\pi} = f(\boldsymbol{\theta})$ in probability.[27] The matrix $A_N$ converges to $\Psi$ in probability, where $\Psi$ is positive definite.

[24] This follows from examining the partitioned inverse of (3.8.26).

[25] If $\hat{\boldsymbol{\pi}}^*$ is another estimator of $\boldsymbol{\pi}$ with asymptotic variance–covariance matrix $\Omega^*$, then the minimum-distance estimator of $\boldsymbol{\theta}$ by choosing $\hat{\boldsymbol{\theta}}^*$ to minimize $[\hat{\boldsymbol{\pi}}^* - \mathbf{f}(\boldsymbol{\theta})]'\Omega^{*-1}[\hat{\boldsymbol{\pi}}^* - \mathbf{f}(\boldsymbol{\theta})]$ has asymptotic variance–covariance matrix $(F'\Omega^{*-1}F)^{-1}$. Suppose $\Omega - \Omega^*$ is positive semidefinite; then $F'\Omega^{*-1}F - F'\Omega^{-1}F = F'(\Omega^{*-1} - \Omega^{-1})F$ is positive semidefinite. Thus, the efficiency of the minimum-distance estimator depends crucially on the efficiency of the (unconstrained) estimator of $\boldsymbol{\pi}$.

[26] For a comprehensive discussion of the Chamberlain $\pi$-approach and the GMM method, see Crépon and Mairesse (1996).

[27] In fact, a stronger result can be established for the proposition that $\hat{\boldsymbol{\pi}}$ converges to $\boldsymbol{\pi}$ almost surely. In this monograph we do not attempt to distinguish the concept of convergence in probability and convergence almost surely (Rao 1973, their Section 2.c), because the stronger result requires a lot more rigor in assumptions and derivations without much gain in intuition.

**Assumption 3A.2:** The vector $\boldsymbol{\theta}$ belongs to a compact subset of $p$-dimensional space. The functions $\mathbf{f}(\boldsymbol{\theta})$ possess continuous second partial derivatives, and the matrix of the first partial derivatives [equation (3.8.15)] has full column rank $p$ in an open neighborhood containing the true parameter $\boldsymbol{\theta}$.

**Assumption 3A.3:** $\sqrt{N}[\hat{\boldsymbol{\pi}}_N - \mathbf{f}(\boldsymbol{\theta})]$ is asymptotically normally distributed with mean zero and variance–covariance matrix $\Delta$.

The minimum-distance estimator chooses $\hat{\boldsymbol{\theta}}$ to minimize $S_N$.

**Proposition 3A.1:** *If assumptions 3A.1 and 3A.2 are satisfied,* $\hat{\boldsymbol{\theta}}$ *converges to* $\boldsymbol{\theta}$ *in probability.*

*Proof:* Assumption 3.A.1 implies that $S_N$ converges to $S = [\mathbf{f}(\boldsymbol{\theta}) - \mathbf{f}(\hat{\boldsymbol{\theta}})]'\Psi[\mathbf{f}(\boldsymbol{\theta}) - \mathbf{f}(\hat{\boldsymbol{\theta}})] = h \geq 0$. Because $\min S = 0$ and the rank condition [assumption 3A.2 or (3.8.15)] implies that in the neighborhood of the true $\boldsymbol{\theta}$, $\mathbf{f}(\boldsymbol{\theta}) = \mathbf{f}(\boldsymbol{\theta}^*)$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ (Hsiao 1983, p. 256), $\hat{\boldsymbol{\theta}}$ must converge to $\boldsymbol{\theta}$ in probability. Q.E.D.

**Proposition 3A.2:** *If assumptions 3A.1–3A.3 are satisfied,* $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ *is asymptotically normally distributed, with mean 0 and variance–covariance matrix*

$$(F'\Psi F)^{-1}F'\Psi\Delta\Psi F(F'\Psi F)^{-1}. \tag{3A.2}$$

*Proof:* $\hat{\boldsymbol{\theta}}$ is the solution of

$$\mathbf{d}_N(\hat{\boldsymbol{\theta}}) = \frac{\partial S_N}{\partial \boldsymbol{\theta}}\Big|_{\hat{\boldsymbol{\theta}}} = -2\left(\frac{\partial \mathbf{f}'}{\partial \hat{\boldsymbol{\theta}}}\right)A_N[\hat{\boldsymbol{\pi}}_N - \mathbf{f}(\hat{\boldsymbol{\theta}})] = \mathbf{0}. \tag{3A.3}$$

The mean-value theorem implies that

$$\mathbf{d}_N(\hat{\boldsymbol{\theta}}) = \mathbf{d}_N(\boldsymbol{\theta}) + \left(\frac{\partial d_N(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}'}\right)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}), \tag{3A.4}$$

where $\boldsymbol{\theta}^*$ is on the line segment connecting $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$. Because $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}$, direct evaluation shows that $\partial d_N(\boldsymbol{\theta}^*)/\partial\boldsymbol{\theta}'$ converges to

$$\frac{\partial d_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = 2\left(\frac{\partial \mathbf{f}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right)'\Psi\left(\frac{\partial \mathbf{f}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right) = 2F'\Psi F.$$

Hence, $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ has the same limiting distribution as

$$-\left[\frac{\partial d_N(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right]^{-1}\cdot\sqrt{N}\mathbf{d}_N(\boldsymbol{\theta}) = (F'\Psi F)^{-1}F'\Psi\cdot\sqrt{N}[\hat{\boldsymbol{\pi}}_N - \mathbf{f}(\boldsymbol{\theta})]. \tag{3A.5}$$

Assumption 3A.3 says that $\sqrt{N}[\hat{\boldsymbol{\pi}}_N - \mathbf{f}(\boldsymbol{\theta})]$ is asymptotically normally distributed, with mean 0 and variance–covariance $\Delta$. Therefore, $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normally distributed, with mean 0 and variance–covariance matrix given by (3A.2). Q.E.D.

**Proposition 3A.3:** *If $\Delta$ is positive definite, then*

$$(F'\Psi F)^{-1}F'\Psi\Delta\Psi F(F'\Psi F)^{-1} - (F'\Delta^{-1}F)^{-1} \tag{3A.6}$$

*is positive semidefinite; hence, an optimal choice for $\Psi$ is $\Delta^{-1}$.*

*Proof:* Because $\Delta$ is positive definite, there is a nonsingular matrix $\tilde{C}$ such that $\Delta = \tilde{C}\tilde{C}'$. Let $\tilde{F} = \tilde{C}^{-1}F$ and $\tilde{B} = (F'\Psi F)^{-1}F'\Psi\tilde{C}$. Then (3A.6) becomes $\tilde{B}[I - \tilde{F}(\tilde{F}'\tilde{F})^{-1}\tilde{F}']\tilde{B}'$, which is positive semidefinite. Q.E.D.

**Proposition 3A.4:** *Assumptions 3A.1–3A.3 are satisfied, if $\Delta$ is positive definite, and if $A_N$ converges to $\Delta^{-1}$ in probability, then*

$$N[\hat{\boldsymbol{\pi}}_N - \mathbf{f}(\hat{\boldsymbol{\theta}})]'A_N[\hat{\boldsymbol{\pi}}_N - \mathbf{f}(\hat{\boldsymbol{\theta}})] \tag{3A.7}$$

*converges to $\chi^2$ distribution, with $KT^2 - p$ degrees of freedom.*

*Proof:* Taking Taylor-series expansion of $\mathbf{f}(\hat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}$, we have

$$\mathbf{f}(\hat{\boldsymbol{\theta}}) \simeq \mathbf{f}(\boldsymbol{\theta}) + \frac{\partial\,\mathbf{f}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}). \tag{3A.8}$$

Therefore, for sufficiently large $N$, $\sqrt{N}[\mathbf{f}(\hat{\boldsymbol{\theta}}) - \mathbf{f}(\boldsymbol{\theta})]$ has the same limiting distribution as $F \cdot \sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$. Thus,

$$\sqrt{N}[\hat{\boldsymbol{\pi}}_N - \mathbf{f}(\hat{\boldsymbol{\theta}})] = \sqrt{N}[\hat{\boldsymbol{\pi}}_N - \mathbf{f}(\boldsymbol{\theta})] - \sqrt{N}[\mathbf{f}(\hat{\boldsymbol{\theta}}) - \mathbf{f}(\boldsymbol{\theta})] \tag{3A.9}$$

converges in distribution to $Q^*\tilde{C}u^*$, where $Q^* = I_{KT^2} - F(F'\Delta^{-1}F)^{-1}F'\Delta^{-1}$, $\tilde{C}$ is a nonsingular matrix such that $\tilde{C}\tilde{C}' = \Delta$, and $\mathbf{u}^*$ is normally distributed, with mean 0 and variance–covariance matrix $I_{KT^2}$. Then the quadratic form, (3A.7), converges in distribution of $\mathbf{u}^{*'}\tilde{C}'Q^{*'}\Delta^{-1}Q^*\tilde{C}\mathbf{u}^*$. Let $\tilde{F} = \tilde{C}^{-1}F$ and $M = I_{KT^2} - \tilde{F}(\tilde{F}'\tilde{F})^{-1}\tilde{F}'$; then $M$ is a symmetric idempotent matrix with rank $KT^2 - p$, and $\tilde{C}'Q^{*'}\Delta^{-1}Q^*\tilde{C} = M^2 = M$; hence, (3A.7) converges in distribution to $\mathbf{u}^{*'}M\mathbf{u}^*$, which is $\chi^2$, with $KT^2 - p$ degrees of freedom. Q.E.D.

## APPENDIX 3B: CHARACTERISTIC VECTORS AND THE INVERSE OF THE VARIANCE–COVARIANCE MATRIX OF A THREE-COMPONENT MODEL

In this appendix we derive the inverse of the variance–covariance matrix (3.6.23) for a three-component model (3.6.8) by means of its characteristic roots and vectors. The material is drawn from the work of Nerlove (1971b).

The matrix $\tilde{V}$ (3.6.23) has three terms, one in $I_{NT}$, one in $I_N \otimes \mathbf{e}\mathbf{e}'$, and one in $\mathbf{e}_N\mathbf{e}_N' \otimes I_T$. Thus, the vector $(\mathbf{e}_N/\sqrt{N}) \otimes (\mathbf{e}/\sqrt{T})$ is a characteristic vector, with the associated root $\sigma_u^2 + T\sigma_\alpha^2 + N\sigma_\lambda^2$. To find $NT - 1$ other characteristic vectors, we note that we can always find $N - 1$ vectors, $\boldsymbol{\psi}_j$, $j = 1, \ldots, N - 1$,

each $N \times 1$ that are orthonormal and orthogonal to $\mathbf{e}_N$:

$$\begin{aligned} &\boldsymbol{\psi}_N' \boldsymbol{\psi}_j = 0, \\ &\boldsymbol{\psi}_j' \boldsymbol{\psi}_{j'} = \begin{cases} 1, & \text{if } j = j', \\ 0, & \text{if } j \neq j', \end{cases} \quad j = 1. \ldots, N-1, \end{aligned} \tag{3B.1}$$

and $T-1$ vectors $\Phi_k$, $k = 1, \ldots, T-1$, each $T \times 1$, that are orthonormal and orthogonal to $\mathbf{e}$:

$$\begin{aligned} &\mathbf{e}' \Phi_k = 0 \\ &\Phi_k' \Phi_{k'} = \begin{cases} 1 & \text{if } k = k', \\ 0, & \text{if } k \neq k', \end{cases} \quad k = 1, \ldots, T-1, \end{aligned} \tag{3B.2}$$

Then the $(N-1)(T-1)$ vectors $\boldsymbol{\psi}_j \otimes \Phi_k$, $j = 1, \ldots, N-1$, $k = 1, \ldots, T-1$, the $N-1$ vectors $\boldsymbol{\psi}_j \otimes (\mathbf{e}/\sqrt{T})$, $j = 1, \ldots, N-1$, and the $T-1$ vectors $\mathbf{e}_N/\sqrt{N} \otimes \Phi_k$, $k = 1, \ldots, T-1$, are also characteristic vectors of $\tilde{V}$, with the associated roots $\sigma_u^2$, $\sigma_u^2 + T\sigma_\alpha^2$, and $\sigma_u^2 + N\sigma_\lambda^2$, which are of multiplicity $(N-1)(T-1)$, $(N-1)$, and $(T-1)$, respectively.

Let

$$\begin{aligned} C_1 &= \frac{1}{\sqrt{T}}[\boldsymbol{\psi}_1 \otimes \mathbf{e}, \ldots, \boldsymbol{\psi}_{N-1} \otimes \mathbf{e}], \\ C_2 &= \frac{1}{\sqrt{N}}[\mathbf{e}_N \otimes \Phi_1, \ldots, \mathbf{e}_N \otimes \Phi_{T-1}], \\ C_3 &= [\boldsymbol{\psi}_1 \otimes \Phi_1, \boldsymbol{\psi}_1 \otimes \Phi_2, \ldots, \boldsymbol{\psi}_{N-1} \otimes \Phi_{T-1}], \\ C_4 &= \left(\mathbf{e}_N/\sqrt{N}\right) \otimes \left(\mathbf{e}/\sqrt{T}\right) = \frac{1}{\sqrt{NT}} \mathbf{e}_{NT}, \end{aligned} \tag{3B.3}$$

and

$$C = [C_1 \quad C_2 \quad C_3 \quad C_4]. \tag{3B.4}$$

Then

$$CC' = C_1 C_1' + C_2 C_2' + C_3 C_3' + C_4 C_4' = I_{NT}, \tag{3B.5}$$

$$C\tilde{V}C' = \begin{bmatrix} (\sigma_u^2 + T\sigma_\alpha^2) I_{N-1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_u^2 + N\sigma_\lambda^2 I_{T-1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_u^2 I_{(N-1)(T-1)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \sigma_u^2 + T\sigma_\alpha^2 + N\sigma_\lambda^2 \end{bmatrix} = \Lambda, \tag{3B.6}$$

and

$$\tilde{V} = C \Lambda C'.$$

Let $A = I_N \otimes \mathbf{e}\mathbf{e}'$, $D = \mathbf{e}_N\mathbf{e}'_N \otimes I_T$, and $J = \mathbf{e}_{NT}\mathbf{e}'_{NT}$. From

$$C_4C'_4 = \frac{1}{NT}J, \tag{3B.7}$$

Nerlove (1971b) showed that by premultiplying (3B.5) by $A$, we have

$$C_1C'_1 = \frac{1}{T}A - \frac{1}{NT}J, \tag{3B.8}$$

and premultiplying (3B.5) by $D$,

$$C_2C'_2 = \frac{1}{N}D - \frac{1}{NT}J. \tag{3B.9}$$

Premultiplying (3B.5) by $A$ and $D$ and using the relations (3B.5), (3B.7), (3B.8), and (3B.9), we have

$$C_3C'_3 = I_{NT} - \frac{1}{T}A - \frac{1}{N}D + \frac{1}{NT}J = \tilde{Q}. \tag{3B.10}$$

Because $\tilde{V}^{-1} = C\Lambda^{-1}C'$, it follows that

$$\tilde{V}^{-1} = \frac{1}{\sigma_u^2 + T\sigma_\alpha^2}\left(\frac{1}{T}A - \frac{1}{NT}J\right) + \frac{1}{\sigma_u^2 + N\sigma_\lambda^2}\left(\frac{1}{N}D - \frac{1}{NT}J\right) + \frac{1}{\sigma_u^2}\tilde{Q} + \frac{1}{\sigma_u^2 + T\sigma_\alpha^2 + N\sigma_\lambda^2}\left(\frac{1}{NT}J\right). \tag{3B.11}$$

CHAPTER 4

# Dynamic Models with Variable Intercepts

## 4.1 INTRODUCTION

In Chapter 3 we discussed the implications of treating the specific effects as fixed or random and the associated estimation methods for the linear static model

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i^* + \lambda_t + u_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{matrix} \tag{4.1.1}$$

where $\mathbf{x}_{it}$ is a $K \times 1$ vector of explanatory variables, including the constant term; $\boldsymbol{\beta}$ is a $K \times 1$ vector of constants; $\alpha_i^*$ and $\lambda_t$ are the (unobserved) individual- and time-specific effects, which are assumed to stay constant for given $i$ over $t$ and for given $t$ over $i$, respectively; and let $u_{it}$ represent the effects of those unobserved variables that vary over $i$ and $t$. Very often we also wish to use panel data to estimate behavioral relationships that are dynamic in character, namely, models containing lagged dependent variables such as[1]

$$y_{it} = \gamma y_{i,t-1} + \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i^* + \lambda_t + u_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{matrix} \tag{4.1.2}$$

where $Eu_{it} = 0$, and $Eu_{it}u_{js} = \sigma_u^2$ if $i = j$ and $t = s$ and $Eu_{it}u_{js} = 0$ otherwise. It turns out that in this circumstance the choice between a fixed-effects formulation and a random-effects formulation has implications for estimation that are of a different nature than those associated with the static model.

Roughly speaking, two issues have been raised in the literature regarding whether the effects, $\alpha_i$ and $\lambda_t$, should be treated as random or as fixed for a linear static model, namely, the efficiency of the estimates and the independence between the effects and the regressors (i.e., the validity of the strict exogeneity assumption of the regressors (3.4.1); e.g., Maddala 1971a; Mundlak 1978a (see Chapter 3)). When all the explanatory variables are fixed constants or strictly exogenous relative to $u$, the covariance estimator is the best linear

[1] We defer the discussion of estimating distributed-lag models to Chapter 11.

unbiased estimator under the fixed-effects assumption and a consistent and unbiased estimator under the random-effects assumption, even though it is not efficient. However, when there exist omitted individual attributes that are correlated with the included exogenous variables, the covariance (CV) estimator does not suffer from bias due to omission of these relevant individual attributes because their impacts have been differenced out, but a generalized least-squares estimator for the random-effects model under the assumption of independence between the effects and explanatory variables is biased. Furthermore, in a linear static model if the effects are correlated with the mean of the explanatory variables, a correctly formulated random-effects model leads to the same CV estimator as the fixed-effects model (Mundlak (1978a); see also Section 3.4 in Chapter 3). Thus, the fixed-effects model has assumed paramount importance in empirical studies (e.g., Ashenfelter 1978; Hausman 1978; Kiefer 1979).

However, if lagged dependent variables also appear as explanatory variables, strict exogeneity of the regressors no longer holds. The initial values of a dynamic process raise another problem. It turns out that with a random-effects formulation, the interpretation of a model depends on the assumption of initial observation. In the case of fixed-effects formulation, the maximum-likelihood estimator (MLE) or the CV estimator is no longer consistent in the typical situation in which a panel involves a large number of individuals, but over only a short period of time. The consistency and asymptotic properties of various fixed-effects estimators to be discussed in this chapter depend on the way in which the number of time series observations $T$ and the number of cross-sectional units $N$ tend to infinity.

For ease of exposition, we shall first assume that the time-specific effects, $\lambda_t$, do not appear. In Section 4.2 we discuss the properties of the CV (or the least squares dummy variable) estimator. Section 4.3 discusses the random-effects model. We discuss the implications of various formulation and methods of estimation. We show that the ordinary least-squares estimator is inconsistent but the MLE, the instrumental variable (IV), and the generalized method of moments (GMM) estimator are consistent. Procedures to test initial conditions are also discussed. In Section 4.4 we use Balestra and Nerlove's (1966) model of demand for natural gas to illustrate the consequences of various assumptions for the estimated coefficients. Section 4.5 discusses the estimation of the fixed-effects dynamic model. We show that although the conventional MLE and CV estimator are inconsistent when $T$ is fixed and $N$ tends to infinity, there exists a transformed likelihood approach that does not involve the incidental parameter and is consistent and efficient under proper formulation of initial conditions. We also discuss the IV and GMM estimator that does not need the formulation of initial conditions. Procedures to test fixed versus random effects are also suggested. In Section 4.6 we relax the assumption on the specific serial-correlation structure of the error term and propose a system approach to estimating dynamic models. Models with both individual- and time-specific effects are discussed in Section 7.

## 4.2 THE CV ESTIMATOR

The CV transformation removes the individual-specific effects from the specification; hence the issue of random- versus fixed-effects specification does not arise. The CV estimator is consistent for the static model when either $N$ or $T$ or both are large. In the case of dynamic model, the properties of CV (or LDSV) depend on the way in which $N$ and $T$ goes to infinity.

Consider[2]

$$y_{it} = \gamma y_{i,t-1} + \alpha_i^* + u_{it}, \quad |\gamma| < 1, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{matrix} \tag{4.2.1}$$

where for simplicity we let $\alpha_i^* = \alpha_i + \mu$ to avoid imposing the restriction that $\sum_{i=1}^N \alpha_i = 0$. We also assume that $y_{i0}$ are observable, $Eu_{it} = 0$, and $Eu_{it}u_{js} = \sigma_u^2$ if $i = j$ and $t = s$, and $Eu_{it}u_{js} = 0$ otherwise.

Let $\overline{y}_i = \sum_{t=1}^T y_{it}/T$, $\overline{y}_{i,-1} = \sum_{t=1}^T y_{i,t-1}/T$, and $\overline{u}_i = \sum_{t=1}^T u_{it}/T$. The LSDV (CV) estimators for $\alpha_i^*$ and $\gamma$ are

$$\hat{\alpha}_i^* = \overline{y}_i - \hat{\gamma}_{\text{cv}}\overline{y}_{i,-1}, \qquad i = 1, \ldots, N, \tag{4.2.2}$$

$$\begin{aligned} \hat{\gamma}_{\text{cv}} &= \frac{\sum_{i=1}^N \sum_{t=1}^T (y_{it} - \overline{y}_i)(y_{i,t-1} - \overline{y}_{i,-1})}{\sum_{i=1}^N \sum_{t=1}^T (y_{i,t-1} - \overline{y}_{i,-1})^2} \\ &= \gamma + \frac{\sum_{i=1}^N \sum_{t=1}^T (y_{i,t-1} - \overline{y}_{i,-1})(u_{it} - \overline{u}_i)/NT}{\sum_{i=1}^N \sum_{t=1}^T (y_{i,t-1} - \overline{y}_{i,-1})^2/NT}. \end{aligned} \tag{4.2.3}$$

The CV estimator exists if the denominator of the second term of (4.2.3) is nonzero. It is consistent if the numerator of the second term of (4.2.3) converges to 0 as sample size increases.

By continuous substitution, we have

$$y_{it} = u_{it} + \gamma u_{i,t-1} + \cdots + \gamma^{t-1} u_{i1} + \frac{1 - \gamma^t}{1 - \gamma}\alpha_i^* + \gamma^t y_{i0}. \tag{4.2.4}$$

[2] The assumption that $|\gamma| < 1$ is made to establish the (weak) stationarity of an autoregressive process (Anderson 1971, Chapters 5 and 7). A stochastic process $\{\xi_t\}$ is stationary if its probability structure does not change with time. A stochastic process is weakly stationary if its mean $E\xi_t = m$ is a constant, independent of its time, and if the covariance of any two variables $E(\xi_t - E\xi_t)(\xi_s - E\xi_s) = \sigma_\xi(t - s)$ depends only on their distance apart in time. The statistical properties of a least-squares estimator for the dynamic model vary with whether or not $|\gamma| < 1$ when $T \to \infty$ (Anderson 1959). When $T$ is fixed and $N \to \infty$, it is not necessary to assume that $|\gamma| < 1$ to establish the asymptotic normality of the least-squares estimator (Anderson 1978; Goodrich and Caines 1979). We keep this conventional assumption for simplicity of exposition and also because it allows us to provide a unified approach toward various assumptions about the initial conditions discussed in Chapter 4, Section 4.3.

Summing $y_{i,t-1}$ over $t$, we have

$$\sum_{t=1}^{T} y_{i,t-1} = \frac{1-\gamma^T}{1-\gamma} y_{i0} + \frac{(T-1) - T\gamma + \gamma^T}{(1-\gamma)^2}\alpha_i^* + \frac{1-\gamma^{T-1}}{1-\gamma}u_{i1} + \frac{1-\gamma^{T-2}}{1-\gamma}u_{i2} + \cdots + u_{i,T-1}. \tag{4.2.5}$$

Under the assumption that $u_{it}$ are uncorrelated with $\alpha_i^*$ and are independently identically distributed, by a law of large numbers (Rao 1973), and using (4.2.5), we can show that when $N$ tends to infinity,

$$\begin{aligned} &\operatorname*{plim}_{N\to\infty} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{i,t-1} - \overline{y}_{i,-1})(u_{it} - \overline{u}_i) \\ &= -\operatorname*{plim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N}\overline{y}_{i,-1}\overline{u}_i \\ &= -\frac{\sigma_u^2}{T^2}\cdot\frac{(T-1) - T\gamma + \gamma^T}{(1-\gamma)^2}. \end{aligned} \tag{4.2.6}$$

By similar manipulations we can show that the denominator of (4.2.3) converges to

$$\frac{\sigma_u^2}{1-\gamma^2}\left\{1 - \frac{1}{T} - \frac{2\gamma}{(1-\gamma)^2}\cdot\frac{(T-1) - T\gamma + \gamma^T}{T^2}\right\}. \tag{4.2.7}$$

as $N \to \infty$. If $T$ is fixed, (4.2.6) is a nonzero constant, and (4.2.2) and (4.2.3) are inconsistent estimators no matter how large $N$ is. The asymptotic bias of the CV of $\gamma$ is

$$\operatorname*{plim}_{N\to\infty}(\hat{\gamma}_{\text{cv}} - \gamma) = -\frac{1+\gamma}{T-1}\left(1 - \frac{1}{T}\frac{1-\gamma^T}{1-\gamma}\right) \cdot\left\{1 - \frac{2\gamma}{(1-\gamma)(T-1)}\left[1 - \frac{1-\gamma^T}{T(1-\gamma)}\right]\right\}^{-1}. \tag{4.2.8}$$

*The bias of* $\hat{\gamma}_{cv}$ is caused by having to eliminate the unknown individual effects $\alpha_i^*$ from each observation, which creates the correlation of the order $(1/T)$ between the regressors and the residuals in the transformed model $(y_{it} - \overline{y}_i) = \gamma(y_{i,t-1} - \overline{y}_{i,-1}) + (u_{it} - \overline{u}_i)$. For small $T$, this bias is always negative if $\gamma > 0$. Nor does the bias go to 0 as $\gamma$ goes to 0. Because a typical panel usually contains a small number of time series observations, this bias can hardly be ignored. For instance, when $T = 2$, the asymptotic bias is equal to $-(1+\gamma)/2$, and when $T = 3$, it is equal to $-(2+\gamma)(1+\gamma)/2$. Even with $T = 10$ and $\gamma = 0.5$, the asymptotic bias is $-0.167$. The CV estimator for the dynamic fixed-effects model remains biased with the introduction of exogenous variables if $T$ is

small; for details of the derivation, see Anderson and Hsiao (1982) and Nickell (1981); for Monte Carlo studies, see Nerlove (1971a).

The process of eliminating the individual-specific effects $\alpha_i$ introduces an estimation error of order $T^{-1}$. When $T$ is large, $(y_{i,t-1} - \bar{y}_{i,-1})$ and $(u_{it} - \bar{u}_i)$ become asymptotically uncorrelated, (4.2.6) converges to zero, and (4.2.7) converges to a nonzero constant $\sigma_u^2/(1-\gamma^2)$. Hence when $T \longrightarrow \infty$, the CV estimator becomes consistent. It can be shown that when $N$ is fixed and $T$ is large, $\sqrt{T}(\hat{\gamma}_{cv} - \gamma)$ is asymptotically normally distributed with mean 0 and variance $1-\gamma^2$. When both $N$ and $T$ are large, the CV estimator remains consistent. However, the standard error of the CV is now of order $(\frac{1}{\sqrt{NT}})$.

The $t$-statistic

$$\frac{(\hat{\gamma}_{CV} - \gamma)}{\text{standard error of } \hat{\gamma}_{CV}} \tag{4.2.9}$$

is no longer centered at 0 because the order $\left(\frac{1}{T}\right)$ correlation between $(y_{i,t-1} - \bar{y}_i)$ and $(u_{it} - \bar{u}_i)$ gets magnified by large $N$. The scale factor $\sqrt{NT} = \sqrt{c}T$ if $\frac{N}{T} = c \neq 0 < \infty$ as $T \longrightarrow \infty$, $(\hat{\gamma}_{cv} - \gamma)$ divided by its standard error is equivalent to multiplying $(\hat{\gamma}_{cv} - \gamma)$ by a scale factor $T$. Equation (4.2.6) multiplied by $T$ will not go to 0 no matter how large $T$ is. Hahn and Kuersteiner (2002) have shown that $\sqrt{NT}(\hat{\gamma}_{cv} - \gamma)$ is asymptotically normally distributed with mean $-\sqrt{c}(1+\gamma)$ and variance $1-\gamma^2$. In other words, the usual $t$-statistic based on $\hat{\gamma}_{cv}$ is not centered at 0, and hence could be subject to severe size distortion when $N$ also increases as $T$ increases such that $\frac{N}{T} \to c \neq 0$ as $T \to \infty$ (e.g., Hsiao and Zhang 2013).

## 4.3 RANDOM-EFFECTS MODELS

When the specific effects are treated as random, they can be considered to be either correlated or not correlated with the explanatory variables. In the case in which the effects are correlated with the explanatory variables, ignoring this correlation and simply using the CV estimator no longer yields the desirable properties as in the case of static regression models. Thus, a more appealing approach here would be to take explicit account of the linear dependence between the effects and the exogenous variables by letting $\alpha_i = \mathbf{a}'\bar{\mathbf{x}}_i + \omega_i$ (Mundlak 1978a) (see Section 3.4) and use a random-effects framework of the model

$$\mathbf{y}_i = \mathbf{y}_{i,-1}\gamma + X_i\boldsymbol{\beta} + \mathbf{e}\bar{\mathbf{x}}_i'\mathbf{a} + \mathbf{e}\omega_i + u_i, \tag{4.3.1}$$

where now $E(\mathbf{x}_{it}\omega_i) = \mathbf{0}$ and $E(\mathbf{x}_{it}u_{it}) = \mathbf{0}$. However, because $\bar{\mathbf{x}}_i$ is time-invariant and the (residual) individual effect $\omega_i$ possesses the same property as $\alpha_i$ when the assumption $E\alpha_i\mathbf{x}_{it}' = \mathbf{0}'$ holds, the estimation of (4.3.1) is formally equivalent to the estimation of the model

$$\mathbf{y}_i = \mathbf{y}_{i,-1}\gamma + X_i\boldsymbol{\beta} + \mathbf{e}\mathbf{z}_i'\boldsymbol{\rho} + \mathbf{e}\alpha_i + u_i, \tag{4.3.2}$$

with $X_i$ now denoting the $T \times K_1$ time-varying explanatory variables, $\mathbf{z}_i'$ being the $1 \times K_2$ time-invariant explanatory variables including the intercept term, and $E\alpha_i = 0$, $E\alpha_i \mathbf{z}_i' = \mathbf{0}'$, and $E\alpha_i \mathbf{x}_{it}' = \mathbf{0}'$. So, for ease of exposition, we assume in this section that the effects are uncorrelated with the exogenous variables.[3]

We first show that the ordinary least-squares (OLS) estimator for dynamic error-component models is biased. We then discuss how the assumption about the initial observations affects interpretation of a model. Finally we discuss estimation methods and their asymptotic properties under various assumptions about initial conditions and sampling schemes.

### 4.3.1 Bias in the OLS Estimator

In the static case in which all the explanatory variables are exogenous and are uncorrelated with the effects, we can ignore the error-component structure and apply the OLS method. The OLS estimator, although less efficient, is still unbiased and consistent. But this is no longer true for dynamic error-component models. The correlation between the lagged dependent variable and individual-specific effects would seriously bias the OLS estimator. We use the following simple model to illustrate the extent of bias. Let

$$y_{it} = \gamma y_{i,t-1} + \alpha_i + u_{it}, \quad |\gamma| < 1, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots T, \end{matrix} \tag{4.3.3}$$

where $u_{it}$ is independently, identically distributed over $i$ and $t$. The OLS estimator of $\gamma$ is

$$\hat{\gamma}_{\mathrm{LS}} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it} \cdot y_{i,t-1}}{\sum_{i=1}^{N}\sum_{t=1}^{T} y_{i,t-1}^2} = \gamma + \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(\alpha_i + u_{it})y_{i,t-1}}{\sum_{i=1}^{N}\sum_{t=1}^{T} y_{i,t-1}^2}. \tag{4.3.4}$$

The asymptotic bias of the OLS estimator is given by the probability limit of the second term on the right-hand side of (4.3.4). Using a manipulation similar to that in Section 4.2, we can show that

$$\begin{aligned} &\operatorname*{plim}_{N\to\infty} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\alpha_i + u_{it})y_{i,t-1} \\ &\quad = \frac{1}{T}\frac{1-\gamma^T}{1-\gamma}\mathrm{Cov}(y_{i0}, \alpha_i) + \frac{1}{T}\frac{\sigma_\alpha^2}{(1-\gamma)^2}\left[(T-1) - T\gamma + \gamma^T\right], \end{aligned} \tag{4.3.5}$$

[3] This does not mean that we have resolved the issue of whether or not the effects are correlated with the exogenous variables. It only means that for estimation purposes we can let $\alpha_i$ stand for $\omega_i$ and treat (4.3.1) as a special case of (4.3.2).

$$\begin{aligned}\operatorname*{plim}_{N\to\infty}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}y_{i,t-1}^2 &= \frac{1-\gamma^{2T}}{T(1-\gamma^2)}\cdot\operatorname*{plim}\frac{\sum_{i=1}^{N}y_{i0}^2}{N}\\ &+\frac{\sigma_\alpha^2}{(1-\gamma)^2}\cdot\frac{1}{T}\left(T-2\frac{1-\gamma^T}{1-\gamma}+\frac{1-\gamma^{2T}}{1-\gamma^2}\right)\\ &+\frac{2}{T(1-\gamma)}\left(\frac{1-\gamma^T}{1-\gamma}-\frac{1-\gamma^{2T}}{1-\gamma^2}\right)\operatorname{Cov}(\alpha_i, y_{i0})\\ &+\frac{\sigma_u^2}{T(1-\gamma^2)^2}\left[(T-1)-T\gamma^2+\gamma^{2T}\right].\end{aligned} \tag{4.3.6}$$

Usually, $y_{i0}$ are assumed either to be arbitrary constants or to be generated by the same process as any other $y_{it}$, so that $\operatorname{Cov}(y_{i0}, \alpha_i)$ is either 0 or positive.[4] Under the assumption that the initial values are bounded, namely, that $\operatorname{plim}_{N\to\infty}\sum_{i=1}^{N}y_{i0}^2/N$ is finite, the OLS method overestimates the true autocorrelation coefficient $\gamma$ when $N$ or $T$ or both tend to infinity. The overestimation is more pronounced the greater the variance of the individual effects, $\sigma_\alpha^2$. This asymptotic result also tends to hold in finite samples according to the Monte Carlo studies conducted by Nerlove (1967) ($N = 25$, $T = 10$).

The addition of exogenous variables to a first-order autoregressive process does not alter the direction of bias of the estimator of the coefficient of the lagged dependent variable, although its magnitude is somewhat reduced. The estimator of the coefficient of the lagged dependent variable remains biased upward, and the estimated coefficients of the exogenous variables are biased downward.

Formulas for the asymptotic bias of the OLS estimator for a $p$th-order autoregressive process and for a model also containing exogenous variables were given by Trognon (1978). The direction of the asymptotic bias for a higher-order autoregressive process is difficult to identify a priori.

### 4.3.2 Model Formulation

Consider a model of the form[5]

$$\begin{aligned}y_{it} = \gamma y_{i,t-1} + \boldsymbol{\rho}'\mathbf{z}_i + \boldsymbol{\beta}'\mathbf{x}_{it} + v_{it}, \quad & i = 1, \ldots, N,\\ & t = 1, \ldots, T,\end{aligned} \tag{4.3.7}$$

[4] For details, see Chapter 4, Section 4.3.2 or Sevestre and Trognon (1982).

[5] The presence of the term $\mathbf{x}_{it}'\boldsymbol{\beta}$ shows that the process $\{y_{it}\}$ is not generally stationary. But the statistical properties of the process $\{y_{it}\}$ vary fundamentally when $T \to \infty$ according to whether or not $\{y_{it}\}$ converges to a stationary process when the sequence of $\mathbf{x}_{it}$ is identically 0. As stated in footnote 2, we shall adopt the first position by letting $|\gamma| < 1$.

where $|\gamma| < 1$, $v_{it} = \alpha_i + u_{it}$,

$$\begin{aligned} E\alpha_i &= Eu_{it} = 0, \\ E\alpha_i \mathbf{z}_i' &= \mathbf{0}', \quad E\alpha_i \mathbf{x}_{it}' = \mathbf{0}', \\ E\alpha_i u_{it} &= 0, \\ E\alpha_i \alpha_j &= \begin{cases} \sigma_\alpha^2 & \text{if } i = j, \\ 0 & \text{otherwise}, \end{cases} \\ Eu_{it}u_{js} &= \begin{cases} \sigma_u^2 & \text{if } i = j, \quad t = s, \\ 0 & \text{otherwise}, \end{cases} \end{aligned} \tag{4.3.8}$$

and where $\mathbf{z}_i$ is a $K_2 \times 1$ vector of time-invariant exogenous variables such as the constant term or an individual's sex or race, $\mathbf{x}_{it}$ is a $K_1 \times 1$ vector of time-varying exogenous variables, $\gamma$ is $1 \times 1$, and $\boldsymbol{\rho}$ and $\boldsymbol{\beta}$ are $K_2 \times 1$ and $K_1 \times 1$ vectors of parameters, respectively. Equation (4.3.7) can also be written in the form

$$w_{it} = \gamma w_{i,t-1} + \boldsymbol{\rho}'\mathbf{z}_i + \boldsymbol{\beta}'\mathbf{x}_{it} + u_{it}, \tag{4.3.9}$$

$$y_{it} = w_{it} + \eta_i, \tag{4.3.10}$$

where

$$\alpha_i = (1-\gamma)\eta_i, \quad E\eta_i = 0, \quad \operatorname{Var}(\eta_i) = \sigma_\eta^2 = \sigma_\alpha^2/(1-\gamma)^2. \tag{4.3.11}$$

Algebraically, (4.3.7) is identical to (4.3.9) and (4.3.10). However, the interpretation of how $y_{it}$ is generated is not the same. Equation (4.3.7) implies that apart from a common response to its own lagged value and the exogenous variables, each individual process is also driven by the unobserved characteristics, $\alpha_i$, which are different for different individuals. Equations (4.3.9) and (4.3.10) imply that the dynamic process $\{w_{it}\}$ is independent of the individual effect $\eta_i$. Conditional on the exogenous variables, individuals are driven by an identical stochastic process with independent (and different) shocks that are random draws from a common population [equation (4.3.9)]. It is the observed value of the latent variable $w_{it}$, $y_{it}$, that is shifted by the individual time-invariant random variable $\eta_i$ [equation (4.3.10)]. This difference in means can be interpreted as a difference in individual endowments or a common measurement error for the $i$th process.

If we observe $w_{it}$, we can distinguish (4.3.7) from (4.3.9) and (4.3.10). Unfortunately, $w_{it}$ are unobservable. However, knowledge of initial observations can provide information to distinguish these two processes. Standard assumptions about initial observations are either that they are fixed or that they are random. If (4.3.7) is viewed as the model, we have two fundamental cases: (I) $y_{i0}$ fixed and (II) $y_{i0}$ random. If (4.3.9) and (4.3.10) are viewed as the basic model, we have (III) $w_{i0}$ fixed and (IV) $w_{i0}$ random.

**Case I:** $y_{i0}$ fixed. A cross-sectional unit may start at some arbitrary position $y_{i0}$ and gradually move toward a level $(\alpha_i + \boldsymbol{\rho}'\mathbf{z}_i)/(1-\gamma) + \boldsymbol{\beta}' \sum_{j=0} \mathbf{x}_{i,t-j}\gamma^j$.

This level is determined jointly by the unobservable effect (characteristic) $\alpha_i$, observable time-invariant characteristics $\mathbf{z}_i$, and time-varying variables $\mathbf{x}_{it}$. The individual effect, $\alpha_i$, is a random draw from a population with mean 0 and variance $\sigma_\alpha^2$. This appears to be a reasonable model. But if the decision about when to start sampling is arbitrary and independent of the values of $y_{i0}$, treating $y_{i0}$ as fixed might be questionable because the assumption $E\alpha_i y_{i0} = 0$ implies that the individual effects, $\alpha_i$, are not brought into the model at time 0, but affect the process at time 1 and later. If the process has been going on for some time, there is no particular reason to believe that $y_{i0}$ should be viewed differently than $y_{it}$.

**Case II:** $y_{i0}$ random. We can assume that the initial observations are random, with a common mean $\mu_{y0}$ and variance $\sigma_{y0}^2$. Namely, let

$$y_{i0} = \mu_{y0} + \epsilon_i. \tag{4.3.12}$$

A rationalization of this assumption is that we can treat $y_{it}$ as a state. We do not care how the initial state, $y_{i0}$, is reached as long as we know that it has a distribution with finite mean and variance. Or, alternatively, we can view $\epsilon_i$ as representing the effect of initial individual endowments (after correction for the mean). Depending on the assumption with regard to the correlation between $y_{i0}$ and $\alpha_i$, we can divide this case into two subcases:

**Case IIa:** $y_{i0}$ independent of $\alpha_i$; that is, $\text{Cov}(\epsilon_i, \alpha_i) = 0$. In this case the impact of initial endowments gradually diminishes over time and eventually vanishes. The model is somewhat like case I, in which the starting value and the effect $\alpha_i$ are independent, except that now the starting observable value is not a fixed constant but a random draw from a population with mean $\mu_{y0}$ and variance $\sigma_{y0}^2$.

**Case IIb:** $y_{i0}$ correlated with $\alpha_i$. We denote the covariance between $y_{i0}$ and $\alpha_i$ by $\phi\sigma_{y0}^2$. Then, as time goes on, the impact of initial endowments ($\epsilon_i$) affects all future values of $y_{it}$ through its correlation with $\alpha_i$ and eventually reaches a level $\phi\epsilon_i/(1-\gamma) = \lim_{t\to\infty} E[y_{it} - \boldsymbol{\rho}'\mathbf{z}_i/(1-\gamma) - \boldsymbol{\beta}' \sum_{j=0}^{t-1} \mathbf{x}_{i,t-j}\gamma^j \mid \epsilon_i]$. In the special case that $\phi\sigma_{y0}^2 = \sigma_\alpha^2$, namely, $\epsilon_i = \alpha_i$, the individual effect can be viewed as completely characterized by the differences in initial endowments. The eventual impact of this initial endowment equals $\alpha_i/(1-\gamma) = \eta_i$.

**Case III:** $w_{i0}$ fixed. Here the unobserved individual process $\{w_{it}\}$ has an arbitrary starting value. In this sense, this case is similar to case I. However, the observed cross-sectional units, $y_{it}$, are correlated with the individual effects, $\eta_i$. That is, each of the observed cross-sectional units may start at some arbitrary position $y_{i0}$ and gradually move toward a level $\eta_i + \boldsymbol{\rho}'\mathbf{z}_i/(1-\gamma)$ $+\boldsymbol{\beta}' \sum_{j=0}^{t-1} \mathbf{x}_{i,t-j}\gamma^j$. Nevertheless, we allow for the possibility that the starting period of the sample observations need not coincide with the beginning of a stochastic process by letting the individual effect $\eta_i$ affect all sample observations, including $y_{i0}$.

**Case IV:** $w_{i0}$ random. Depending on whether or not the $w_{i0}$ are viewed as having common mean, we have four subcases:

**Case IVa:** $w_{i0}$ random, with common mean $\mu_w$ and variance $\sigma_u^2/(1-\gamma^2)$

**Case IVb:** $w_{i0}$ random, with common mean $\mu_w$ and arbitrary variance $\sigma_{w0}^2$

**Case IVc:** $w_{i0}$ random, with mean $\theta_{i0}$ and variance $\sigma_u^2/(1-\gamma^2)$

**Case IVd:** $w_{i0}$ random, with mean $\theta_{i0}$ and arbitrary variance $\sigma_{w0}^2$

In each of these four subcases we allow correlation between $y_{i0}$ and $\eta_i$. In other words, $\eta_i$ affects $y_{it}$ in all periods, including $y_{i0}$. Cases IVa and IVb are similar to the state-space representation discussed in case IIa, in which the initial states are random draws from a distribution with finite mean. Case IVa assumes that the initial state has the same variance as the latter states. Case IVb allows the initial state to be nonstationary (with arbitrary variance). Cases IVc and IVd take a different view in that they assume that the individual states are random draws from different populations with different means. A rationalization for this can be seen through successive substitution of (4.3.9), yielding

$$w_{i0} = \frac{1}{1-\gamma}\boldsymbol{\rho}'\mathbf{z}_i + \boldsymbol{\beta}'\sum_{j=0}^{\infty}\mathbf{x}_{i,-j}\gamma^j + u_{i0} + \gamma u_{i,-1} + \gamma^2 u_{i,-2} + \dots \quad (4.3.13)$$

Because $\mathbf{x}_{i0}, \mathbf{x}_{i,-1}, \dots$ are not observable, we can treat the combined cumulative effects of nonrandom variables for the $i$th individual as an unknown parameter and let

$$\theta_{i0} = \frac{1}{1-\gamma}\boldsymbol{\rho}'\mathbf{z}_i + \boldsymbol{\beta}'\sum_{j=0}^{\infty}\mathbf{x}_{i,-j}\gamma^j \quad (4.3.14)$$

Case IVc assumes that the process $\{w_{it}\}$ was generated from the infinite past and has achieved stationarity of its second moments after conditioning on the exogenous variables (i.e., $w_{i0}$ has the same variance as any other $w_{it}$). Case IVd relaxes this assumption by allowing the variance of $w_{i0}$ to be arbitrary.

### 4.3.3 Estimation of Random-Effects Models

There are various ways to estimate the unknown parameters. Here we discuss four methods: the MLE, the GLS, the instrumental-variable (IV), and the GMM methods.

#### *4.3.3.1 Maximum-Likelihood Estimator*

Different assumptions about the initial conditions imply different forms of the likelihood functions. Under the assumption that $\alpha_i$ and $u_{it}$ are normally

distributed, the likelihood function for case I is[6]

$$L_1 = (2\pi)^{-\frac{NT}{2}} \mid V \mid^{-\frac{N}{2}} \exp\left\{ -\frac{1}{2}\sum_{i=1}^{N}(\mathbf{y}_i - \mathbf{y}_{i,-1}\gamma - Z_i\boldsymbol{\rho} - X_i\boldsymbol{\beta})' \cdot V^{-1}(\mathbf{y}_i - \mathbf{y}_{i,-1}\gamma - Z_i\boldsymbol{\rho} - X_i\boldsymbol{\beta})\right\}, \tag{4.3.15}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$, $\mathbf{y}_{i,-1} = (y_{i0}, \ldots, y_{i,T-1})'$, $Z_i = \mathbf{e}\mathbf{z}_i'$, $\mathbf{e} = (1, \ldots, 1)'$, $X_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})'$, and $V = \sigma_u^2 I_T + \sigma_\alpha^2 \mathbf{e}\mathbf{e}'$. The likelihood function for case IIa is

$$L_{2a} = L_1 \cdot (2\pi)^{-\frac{N}{2}}\left(\sigma_{y0}^2\right)^{-\frac{N}{2}} \exp\left\{ -\frac{1}{2\sigma_{y0}^2}\sum_{i=1}^{N}(y_{i0} - \mu_{y0})^2\right\}. \tag{4.3.16}$$

For case IIb, it is of the form

$$\begin{aligned} L_{2b} = (2\pi)^{-\frac{NT}{2}}\left(\sigma_u^2\right)^{-\frac{N(T-1)}{2}}\left(\sigma_u^2 + Ta\right)^{-\frac{N}{2}} \exp\Bigg\{ & -\frac{1}{2\sigma_u^2}\sum_{i=1}^{N}\sum_{t=1}^{T} \\ & \cdot [y_{it} - \gamma y_{i,t-1} - \boldsymbol{\rho}'\mathbf{z}_i - \boldsymbol{\beta}'\mathbf{x}_{it} - \phi(y_{i0} - \mu_{y0})]^2 \\ & + \frac{\mathrm{a}}{2\sigma_u^2\left(\sigma_u^2 + Ta\right)}\sum_{i=1}^{N}\Bigg\{\sum_{t=1}^{T}[y_{it} - \gamma y_{i,t-1} - \boldsymbol{\rho}'\mathbf{z}_i - \boldsymbol{\beta}'\mathbf{x}_{it} - \phi \\ & \cdot (y_{i0} - \mu_{y0})]\Bigg\}^2\Bigg\} \cdot (2\pi)^{-\frac{N}{2}}\left(\sigma_{y0}^2\right)^{-\frac{N}{2}} \\ & \cdot \exp\left\{ -\frac{1}{2\sigma_{y0}^2}\sum_{i=1}^{N}(y_{i0} - \mu_{y0})^2\right\}, \end{aligned} \tag{4.3.17}$$

where $a = \sigma_\alpha^2 - \phi^2\sigma_{y0}^2$. The likelihood function for case III is

$$\begin{aligned} L_3 = (2\pi)^{-\frac{NT}{2}}\left(\sigma_u^2\right)^{-\frac{NT}{2}} \exp\Bigg\{ & -\frac{1}{2\sigma_u^2}\sum_{i=1}^{N}\sum_{t=1}^{T}[(y_{it} - y_{i0} + w_{i0}) \\ & - \gamma(y_{i,t-1} - y_{i0} + w_{i0}) - \boldsymbol{\rho}'\mathbf{z}_i - \boldsymbol{\beta}'\mathbf{x}_{it}]^2\Bigg\} \cdot (2\pi)^{-\frac{N}{2}}\left(\sigma_\eta^2\right)^{-\frac{N}{2}} \\ & \cdot \exp\left\{ -\frac{1}{2\sigma_\eta^2}\sum_{i=1}^{N}(y_{i0} - w_{i0})^2\right\}, \end{aligned} \tag{4.3.18}$$

[6] $V$ is the same as (3.3.4).

and for Case IVa it is

$$
\begin{aligned}
L_{4a} &= (2\pi)^{-\frac{N(T+1)}{2}} \mid \Omega \mid^{-\frac{N}{2}} \\
&\quad \cdot \exp\left\{ -\frac{1}{2}\sum_{i=1}^{N}(y_{i0}-\mu_w, y_{i1}-\gamma y_{i0}-\boldsymbol{\rho}'\mathbf{z}_i-\boldsymbol{\beta}'\mathbf{x}_{i1}, \ldots, \right. \\
&\quad y_{iT}-\gamma y_{i,T-1}-\boldsymbol{\rho}'\mathbf{z}_i-\boldsymbol{\beta}'\mathbf{x}_{iT}) \\
&\quad \left. \Omega^{-1}(y_{i0}-\mu_w, \ldots, y_{iT}-\gamma y_{i,T-1}-\boldsymbol{\rho}'\mathbf{z}_i-\boldsymbol{\beta}'\mathbf{x}_{iT})' \right\},
\end{aligned} \tag{4.3.19}
$$

where

$$
\begin{aligned}
\underset{(T+1)\times(T+1)}{\Omega} &= \sigma_u^2 \begin{bmatrix} \frac{1}{1-\gamma^2} & \mathbf{0}' \\ \mathbf{0} & I_T \end{bmatrix} + \sigma_\alpha^2 \begin{bmatrix} \frac{1}{1-\gamma} \\ \mathbf{e} \end{bmatrix} \left( \tfrac{1}{1-\gamma}, \mathbf{e}' \right) \\
\mid \Omega \mid &= \frac{\sigma_u^{2T}}{1-\gamma^2}\left( \sigma_u^2 + T\sigma_\alpha^2 + \frac{1+\gamma}{1-\gamma}\sigma_\alpha^2 \right), \\
\Omega^{-1} &= \frac{1}{\sigma_u^2}\left[ \begin{bmatrix} 1-\gamma^2 & \mathbf{0}' \\ \mathbf{0} & I_T \end{bmatrix} \right. \\
&\quad \left. - \left( \frac{\sigma_u^2}{\sigma_\alpha^2} + T + \frac{1+\gamma}{1-\gamma} \right)^{-1} \begin{bmatrix} 1+\gamma \\ \mathbf{e} \end{bmatrix} (1+\gamma, \mathbf{e}') \right].
\end{aligned} \tag{4.3.20}
$$

The likelihood function for case IVb, $L_{4b}$, is of the form (4.3.19), except that $\Omega$ is replaced by $\Lambda$, where $\Lambda$ differs from $\Omega$ only in that the upper left element of the first term, $1/(1-\gamma^2)$, is replaced by $\sigma_{w0}^2/\sigma_u^2$. The likelihood function for case IVc, $L_{4c}$, is similar to that for case IVa, except that the mean of $y_{i0}$ in the exponential term is replaced by $\theta_{i0}$. The likelihood function for case IVd, $L_{4d}$, is of the form (4.3.17), with $\theta_{i0}$, $(1-\gamma)\sigma_\eta^2/(\sigma_\eta^2+\sigma_{w0}^2)$, and $\sigma_\eta^2+\sigma_{w0}^2$ replacing $\mu_{y0}$, $\phi$, and $\sigma_{y0}^2$, respectively.

Maximizing the likelihood function with respect to unknown parameters yields the MLE. The consistency of the MLE depends on the initial conditions and on the way in which the number of time series observations $T$ and the cross-sectional units $N$ tends to infinity. For cases III and IVd, the MLEs do not exist. By letting $y_{i0}$ equal to $w_{i0}$ or $\theta_{i0}$, the exponential term of the second function of their respective likelihood function becomes 1. If we let the variances $\sigma_\eta^2$ or $\sigma_\eta^2+\sigma_{w0}^2$ approach 0, the likelihood functions become unbounded. However, we can still take partial derivatives of these likelihood functions and solve for the first-order conditions. For simplicity of exposition, we shall refer to these interior solutions as the MLEs and examine their consistency properties in the same way as in other cases in which the MLEs exist.

When $N$ is fixed, a necessary condition for $\boldsymbol{\rho}$ being identifiable is that $N \geq K_2$. Otherwise, the model is subject to strict multicollinearity. However, when $T$ tends to infinity, even with $N$ greater than $K_2$, the MLEs for $\boldsymbol{\rho}$ and $\sigma_\alpha^2$ remain inconsistent because of insufficient variation across individuals. On

Table 4.1. *Consistency properties of the MLEs for dynamic random-effects models*[a]

| Case | | $N$ fixed, $T \to \infty$ | $T$ fixed, $N \to \infty$ |
|---|---|---|---|
| Case I: $y_{i0}$ fixed | $\gamma, \boldsymbol{\beta}, \sigma_u^2$ | Consistent | Consistent |
| | $\boldsymbol{\rho}, \sigma_\alpha^2$ | Inconsistent | Consistent |
| Case II: $y_{i0}$ random | | | |
| IIa: $y_{i0}$ independent of $\alpha_i$ | $\gamma, \boldsymbol{\beta}, \sigma_u^2$ | Consistent | Consistent |
| | $\mu_{y0}, \boldsymbol{\rho}, \sigma_\alpha^2, \sigma_{y0}^2$ | Inconsistent | Consistent |
| IIb: $y_{i0}$ correlated with $\alpha_i$ | $\gamma, \boldsymbol{\beta}, \sigma_u^2$ | Consistent | Consistent |
| | $\mu_{y0}, \boldsymbol{\rho}, \sigma_\alpha^2, \sigma_{y0}^2, \phi$ | Inconsistent | Consistent |
| Case III: $w_{i0}$ fixed | $\gamma, \boldsymbol{\beta}, \sigma_u^2$ | Consistent | Inconsistent |
| | $w_{i0}, \boldsymbol{\rho}, \sigma_\eta^2$ | Inconsistent | Inconsistent |
| Case IV: $w_{i0}$ random | | | |
| IVa: mean $\mu_w$ and variance $\sigma_u^2/(1-\gamma^2)$ | $\gamma, \boldsymbol{\beta}, \sigma_u^2$ | Consistent | Consistent |
| | $\mu_w, \boldsymbol{\rho}, \sigma_\eta^2$ | Inconsistent | Consistent |
| IVb: mean $\mu_w$ and variance $\sigma_{w0}^2$ | $\gamma, \boldsymbol{\beta}, \sigma_u^2$ | Consistent | Consistent |
| | $\sigma_{w0}^2, \boldsymbol{\rho}, \sigma_\eta^2, \mu_w$ | Inconsistent | Consistent |
| IVc: mean $\theta_{i0}$ and variance $\sigma_u^2/(1-\gamma^2)$ | $\gamma, \boldsymbol{\beta}, \sigma_u^2$ | Consistent | Inconsistent |
| | $\theta_{i0}, \boldsymbol{\rho}, \sigma_\eta^2$ | Inconsistent | Inconsistent |
| IVd: mean $\theta_{i0}$ and variance $\sigma_{w0}^2$ | $\gamma, \boldsymbol{\beta}, \sigma_u^2$ | Consistent | Inconsistent |
| | $\theta_{i0}, \boldsymbol{\sigma}_\eta^2, \sigma_{w0}^2$ | Inconsistent | Inconsistent |

[a] If an MLE does not exist, we replace it by the interior solution.
*Source:* Anderson and Hsiao (1982, Table 1).

the other hand, the MLEs of $\gamma$, $\boldsymbol{\beta}$, and $\sigma_u^2$ are consistent for all these different cases. When $T$ becomes large, the weight of the initial observations becomes increasingly negligible, and the MLEs for different cases all converge to the same CV estimator.

For cases IVc and IVd, $w_{i0}$ have means $\theta_{i0}$, which introduces incidental parameter problems. The MLE in the presence of incidental parameters is inconsistent. Bhargava and Sargan (1983) suggest predicting $\theta_{i0}$ by all the observed $\mathbf{x}_{it}$ and $\mathbf{z}_i$ as a means to get around the incidental-parameters problem.[7] If $\mathbf{x}_{it}$ is generated by a homogeneous stochastic process

$$\mathbf{x}_{it} = \mathbf{c} + \sum_{j=0}^{\infty} \mathbf{b}_j \boldsymbol{\xi}_{i,t-j}, \tag{4.3.21}$$

[7] Bhargava and Sargan (1983) get around the issue of incidental parameter associated with initial value, $y_{i0}$, by projecting $y_{i0}$ on $\mathbf{x}_i$ under the assumption that $\alpha_i$ and $\mathbf{x}_i$ are uncorrelated. Chamberlain (1984) and Mundlak (1978a) assume that the effects, $\alpha_i$, are correlated with $\mathbf{x}_i$ and get around the issue of incidental parameters by projecting $\alpha_i$ on $\mathbf{x}_i$ under the assumption that $(\alpha_i, \mathbf{x}_i')$ are independently, identically distributed over $i$.

where $\boldsymbol{\xi}_{it}$ is independently, identically distributed, then the minimum mean square error predictor of $\mathbf{x}_{i,-j}$ by $\mathbf{x}_{it}$ is the same for all $i$. Substituting these predictive formulae into (4.3.14) yields

$$y_{i0} = \sum_{t=1}^{T} \boldsymbol{\pi}_{0t}' \mathbf{x}_{it} + \boldsymbol{\rho}^{*\prime} \mathbf{z}_i + v_{i0}, \tag{4.3.22}$$

and

$$v_{i0} = \epsilon_{i0} + u_{i0}^{*} + \eta_i. \qquad i = 1, \ldots, N. \tag{4.3.23}$$

The coefficients $\boldsymbol{\pi}_{0t}$ are identical across $i$ (Hsiao, Pesaran, and Tahmiscioglu 2002). The error term $v_{i0}$ is the sum of three components: the prediction error of $\theta_{i0}$, $\epsilon_{i0}$; the cumulative shocks before time 0, $u_{i0}^{*} = u_{i0} + \gamma u_{i,-1} + \gamma^2 u_{i,-2} + \ldots$; and the individual effects, $\eta_i$. The prediction error $\epsilon_{i0}$ is independent of $u_{it}$ and $\eta_i$, with mean 0 and variance $\sigma_{\epsilon 0}^2$. Depending on whether or not the error process of $w_{i0}$ conditional on the exogenous variables has achieved stationarity (i.e., whether or not the variance of $w_{i0}$ is the same as any other $w_{it}$), we have[8] case IVc′,

$$\begin{aligned} \mathrm{Var}(v_{i0}) &= \sigma_{\epsilon 0}^2 + \frac{\sigma_u^2}{1-\gamma^2} + \frac{\sigma_\alpha^2}{(1-\gamma)^2} \text{ and} \\ \mathrm{Cov}(v_{i0}, v_{it}) &= \frac{\sigma_\alpha^2}{(1-\gamma)}, \quad t = 1, \ldots, T, \end{aligned} \tag{4.3.24}$$

or case IVd′,

$$\mathrm{Var}(v_{i0}) = \sigma_{w0}^2 \quad \text{and} \quad \mathrm{Cov}(v_{i0}, v_{it}) = \sigma_\tau^2, \quad t = 1, \ldots, T. \tag{4.3.25}$$

Cases IVc′ and IVd′ transform cases IVc and IVd, in which the number of parameters increases with the number of observations, into a situation in which $N$ independently distributed $(T+1)$-component vectors depend only on a fixed number of parameters. Therefore, the MLE is consistent when $N \to \infty$ or $T \to \infty$ or both $N, T \to \infty$. Moreover, the MLE multiplied by the scale factor $\sqrt{NT}$ is centered at the true values independent of the way $N$ or $T$ goes to infinity (for details, see Hsiao and Zhang 2013).

The MLE is obtained by solving the first-order conditions of the likelihood function with respect to unknown parameters. If there is a unique solution to these partial derivative equations with $\sigma_\alpha^2 > 0$, the solution is the MLE. However, just as in the static case discussed in Section 3.3, a boundary solution

[8] Strictly speaking, from (4.3.21), the nonstationary analogue of case IVd would imply that

$$\begin{aligned} \mathrm{Var}(v_{i0}) &= \sigma_{\omega 0}^2 + \frac{\sigma_\alpha^2}{(1-\gamma)^2} \quad \text{and} \\ \mathrm{Cov}(v_{i0}, v_{it}) &= \frac{\sigma_\alpha^2}{(1-\gamma)}, \quad t = 1. \ldots, T. \end{aligned}$$

However, given the existence of the prediction-error term $\epsilon_{i0}$, it is not possible to distinguish this case from case IVc' based on the information of $y_{i0}$ alone. So we shall follow Bhargava and Sargan (1983) in treating case IVd′ as the nonstationary analogue of case IVd.

with $\sigma_\alpha^2 = 0$ may occur for dynamic error-components models as well. Anderson and Hsiao (1981) have derived the conditions under which the boundary solution will occur for various cases. Trognon (1978) has provided analytic explanations based on asymptotic approximations where the number of time periods tends to infinity. Nerlove (1967, 1971a) has conducted Monte Carlo experiments to explore the properties of the MLE. These results show that the autocorrelation structure of the exogenous variables is a determinant of the existence of boundary solutions. In general, the more autocorrelated the exogenous variables or the more important the weight of the exogenous variables, the less likely it is that a boundary solution will occur.

The solution for the MLE is complicated. We can apply the Newton–Raphson type iterative procedure or the sequential iterative procedure suggested by Anderson and Hsiao (1982) to obtain a solution. Alternatively, because we have a cross section of size $N$ repeated successively in $T$ time periods, we can regard the problems of estimation (and testing) of (4.3.7) as akin to those for a simultaneous-equations system with $T$ or $T+1$ structural equations with $N$ observations available on each of the equations. That is, the dynamic relationship (4.3.7) in a given time period is written as an equation in a system of simultaneous equations,

$$\Gamma Y' + BX' + PZ' = U', \tag{4.3.26}$$

where we now let[9]

$$\underset{N\times(T+1)}{Y} = \begin{bmatrix} y_{10} & y_{11} & \cdots & y_{1T} \\ y_{20} & y_{21} & \cdots & y_{2T} \\ \vdots & & & \\ y_{N0} & y_{N1} & \cdots & y_{NT} \end{bmatrix},$$

$$\underset{N\times TK_1}{X} = \begin{bmatrix} \mathbf{x}_{11}' & \mathbf{x}_{12}' & \cdots & \mathbf{x}_{1T}' \\ \mathbf{x}_{21}' & \mathbf{x}_{22}' & \cdots & \mathbf{x}_{2T}' \\ \vdots & & & \\ \mathbf{x}_{N1}' & \mathbf{x}_{N2}' & \cdots & \mathbf{x}_{NT}' \end{bmatrix},$$

$$\underset{N\times K_2}{Z} = \begin{bmatrix} \mathbf{z}_1' \\ \mathbf{z}_2' \\ \vdots \\ \mathbf{z}_N' \end{bmatrix}, \qquad i = 1\ldots, N,$$

and $U$ is the $N \times T$ matrix of errors if the initial values, $y_{i0}$, are treated as constants, or the $N \times (T+1)$ matrix of errors if the initial values are treated as stochastic. The structural form coefficient matrix $A = [\Gamma \ B \ P]$ is

[9] Previously we combined the intercept term and the time-varying exogenous variables into the $\mathbf{x}_{it}$ vector because the property of the MLE for the constant is the same as that of the MLE for the coefficients of time-varying exogenous variables. Now we separate $\mathbf{x}_{it}'$ as $(1, \tilde{\mathbf{x}}_{it}')$, because we wish to avoid having the constant term appearing more than once in (4.3.22).

$T \times [(T+1) + TK_1 + K_2]$ or $(T+1) \times [(T+1) + TK_1 + K_2]$, depending on whether the initial values are treated as fixed or random. The earlier serial covariance matrix [e.g., (3.3.4), (4.3.20), (4.3.24), or (4.3.25)] now becomes the variance–covariance matrix of the errors on $T$ or $(T+1)$ structural equations. We can then use the algorithm for solving the full-information maximum-likelihood estimator to obtain the MLE.

There are cross-equation linear restrictions on the structural form coefficient matrix and restrictions on the variance–covariance matrix. For instances, in case I, where $y_{i0}$ are treated as fixed constants, we have

$$A = \begin{bmatrix} -\gamma & 1 & 0 & . & . & 0 & \boldsymbol{\beta}' & \mathbf{0}' & . & . & . & . & \mathbf{0}' & \boldsymbol{\rho}' \\ 0 & -\gamma & 1 & . & . & . & \mathbf{0}' & \boldsymbol{\beta}' & . & . & . & . & . & \boldsymbol{\rho}' \\ . & . & . & . & . & . & . & . & . & . & . & . & . & \\ . & . & . & . & . & . & . & . & . & . & . & . & . & \boldsymbol{\rho}' \\ . & . & . & . & . & 0 & . & . & . & . & . & . & . & \\ 0 & 0 & 0 & . & -\gamma & 1 & \mathbf{0}' & \mathbf{0}' & . & . & . & . & \boldsymbol{\beta}' & \boldsymbol{\rho}' \end{bmatrix}, \tag{4.3.27}$$

The variance–covariance matrix of $U$ is block-diagonal, with the diagonal block equal to $V$ [equation (3.3.4)]. In case IVd′, where $y_{i0}$ are treated as stochastic, the structural form coefficient matrix $A$ is a $(T+1) \times [(T+1) + TK_1 + K_2]$ matrix of the form

$$A = \begin{bmatrix} 1 & 0 & . & . & . & 0 & \boldsymbol{\pi}'_{01} & \boldsymbol{\pi}'_{02} & . & . & . & \boldsymbol{\pi}'_{0T} & \boldsymbol{\rho}^{*\prime} \\ -\gamma & 1 & . & . & . & . & \boldsymbol{\beta}' & \mathbf{0}' & . & . & . & \mathbf{0}' & \boldsymbol{\rho}' \\ 0 & -\gamma & . & . & . & . & \mathbf{0}' & \boldsymbol{\beta}' & . & . & . & \mathbf{0}' & \\ . & . & . & . & . & . & . & . & . & . & . & . & \\ 0 & . & . & . & -\gamma & 1 & \mathbf{0}' & \mathbf{0}' & . & . & . & \boldsymbol{\beta}' & \boldsymbol{\rho}' \end{bmatrix}, \tag{4.3.28}$$

and the variance–covariance matrix of $U$ is block-diagonal, with the diagonal block a $(T+1) \times (T+1)$ matrix of the form

$$\tilde{V} = \begin{bmatrix} \sigma_{w0}^2 & \sigma_\tau^2 \mathbf{e}' \\ \sigma_\tau^2 \mathbf{e} & V \end{bmatrix}. \tag{4.3.29}$$

Bhargava and Sargan (1983) suggest maximizing the likelihood function of (4.3.26) by directly substituting the restrictions into the structural form coefficient matrix $A$ and the variance–covariance matrix of $U'$.

Alternatively, we can ignore the restrictions on the variance–covariance matrix of $U'$ and use three-stage least-squares (3SLS) methods. Because the restrictions on $A$ are linear, it is easy to obtain the constrained 3SLS estimator of $\gamma$, $\boldsymbol{\beta}$, $\boldsymbol{\rho}$, and $\boldsymbol{\rho}^*$ from the unconstrained 3SLS estimator.[10] Or we can use the Chamberlain (1982, 1984) minimum-distance estimator by first obtaining the

[10] For the formula of the constrained estimator, see Theil 1971, p. 285, equation (8.5).

unconstrained reduced form coefficients matrix $\Pi$, then solving for the structural form parameters (see Section 3.9). The Chamberlain minimum-distance estimator has the same limiting distribution as the constrained generalized 3SLS estimator (see Chapter 5). However, because the maintained hypothesis in the model implies that the covariance matrix of $U'$ is constrained and in some cases dependent on the parameter $\gamma$ occurring in the structural form, the constrained 3SLS or the constrained generalized 3SLS is inefficient in comparison with the (full-information) MLE.[11] But if the restrictions on the variance–covariance matrix are not true, the (full information) MLE imposing the wrong restrictions will in general be inconsistent. But the (constrained) 3SLS or the Chamberlain minimum-distance estimator, because it does not impose any restriction on the covariance matrix of U′, remains consistent and is efficient within the class of estimators that do not impose restrictions on the variance–covariance matrix.

#### *4.3.3.2 Generalized Least-Squares Estimator*

We note that except for Cases III, IVc, and IVd, the likelihood function depends only on a fixed number of parameters. Furthermore, conditional on $\Omega$ or $\sigma_u^2$, $\sigma_\alpha^2$, $\sigma_{y0}^2$, and $\phi$, the MLE is equivalent to the generalized least-squares estimator. For instance, under Case I, the covariance matrix of $(y_{i1}, \ldots, y_{iT})$ is the usual error-components form (3.3.4). Under Case IIa, b and Case IVa, b or Case IVc and IVd when the conditional mean of $\theta_{i0}$ can be represented in the form of (4.3.22), the covariance matrix of $\mathbf{v}_i = (v_{i0}, v_{i1}, \ldots, v_{iT})$, $\tilde{V}$, is of similar form to (4.3.29). Therefore, a GLS estimator of $\boldsymbol{\delta}' = (\boldsymbol{\pi}', \boldsymbol{\rho}^{*\prime}, \gamma, \boldsymbol{\beta}', \boldsymbol{\rho}')$, can be applied,

$$\hat{\tilde{\boldsymbol{\delta}}}_{\text{GLS}} = \left( \sum_{i=1}^{N} \tilde{X}_i' \tilde{V}^{-1} \tilde{X}_i \right)^{-1} \left( \sum_{i=1}^{N} \tilde{X}_i' \tilde{V}^{-1} \tilde{\mathbf{y}}_i \right), \tag{4.3.30}$$

where $\tilde{\mathbf{y}}_i' = (y_{io}, \ldots, y_{iT})$,

$$\tilde{X}_i = \begin{pmatrix} \mathbf{x}_{i1}' & \mathbf{x}_{i2}' & \ldots & \mathbf{x}_{iT}' & \mathbf{z}_i' & 0 & \mathbf{0}' & \mathbf{0} \\ \mathbf{0}' & \ldots & \ldots & \ldots & \mathbf{0}' & y_{i0} & \mathbf{x}_{i1}' & \mathbf{z}_i' \\ \vdots & & & & \vdots & y_{i1} & \mathbf{x}_{i2}' & \mathbf{z}_i' \\ \vdots & & & & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0}' & & & & \mathbf{0}' & y_{i,T-1} & \mathbf{x}_{iT}' & \mathbf{z}_i' \end{pmatrix}.$$

The estimator is consistent and asymptotically normally distributed as $N \to \infty$.

[11] See Chapter 5.

Blundell and Smith (1991) suggest a conditional GLS procedure by conditioning $(y_{i1}, \ldots, y_{iT})$ on $v_{i0} = y_{i0} - E(y_{i0} \mid \mathbf{x}_i', \mathbf{z}_i)$,[12]

$$\mathbf{y}_i = \mathbf{y}_{i,-1}\gamma + Z_i\boldsymbol{\rho} + X_i\boldsymbol{\beta} + \boldsymbol{\tau}\, v_{i0} + \mathbf{v}_i^*, \tag{4.3.31}$$

where $\mathbf{v}_i^* = (v_{i1}^*, \ldots, v_{iT}^*)'$, and $\boldsymbol{\tau}$ is a $T \times 1$ vector of constants with the values depending on the correlation pattern between $y_{i0}$ and $\alpha_i$. For Case IIa, $\boldsymbol{\tau} = \mathbf{0}$, Case IIb, $\boldsymbol{\tau} = \mathbf{e}_T \cdot \phi$. When the covariances between $y_{i0}$ and $(y_{i1}, \ldots, y_{iT})$ are arbitrary, $\boldsymbol{\tau}$ is a $T \times 1$ vector of unrestricted constants. Application of the GLS to (4.3.31) is consistent as $N \to \infty$.

When the covariance matrix of $\mathbf{v}_i$ or $\mathbf{v}_i^*$ is unknown, a feasible GLS estimator can be applied. In the first step, we obtain some consistent estimates of the covariance matrix from the estimated $\mathbf{v}_i$ or $\mathbf{v}_i^*$. For instance, we can use the IV estimator to be discussed in Section 4.3.3.3 to obtain consistent estimators of $\gamma$ and $\boldsymbol{\beta}$, then substitute them into $y_{it} - \gamma y_{i,t-1} - \boldsymbol{\beta}'\mathbf{x}_{it}$, and regress the resulting value on $\mathbf{z}_i$ across individuals to obtain a consistent estimate of $\boldsymbol{\rho}$. Substituting estimated $\gamma$, $\boldsymbol{\beta}$ and $\boldsymbol{\rho}$ into (4.3.2), we obtain estimates of $v_{it}$ for $t = 1, \ldots, T$. The estimates of $v_{i0}$ can be obtained as the residuals of the cross-section regression of (4.3.22). The covariance matrix of $\mathbf{v}_i$ can then be estimated using the procedures discussed in Chapter 3. The estimated $\mathbf{v}_i^*$ can also be obtained as the residuals of the cross-sectional regression of $\mathbf{y}_i - \mathbf{y}_{i,-1}\gamma - X_i\boldsymbol{\beta}$ on $Z_i$ and $\mathbf{e}\hat{v}_{i0}$. In the second step, we treat the estimated covariance matrix of $\mathbf{v}_i$ or $\mathbf{v}_i^*$ as if they were known, apply the GLS to the system composed of (4.3.2) and (4.3.22) or the conditional system (4.3.31).

It should be noted that if $\text{Cov}\,(y_{i0}, \alpha_i) \neq 0$, the GLS applied to the system (4.3.2) is inconsistent when $T$ is fixed and $N \to \infty$. This is easily seen by noting that conditional on $y_{i0}$, the system is of the form (4.3.31). Applying GLS to (4.3.2) is therefore subject to omitted variable bias. However, the asymptotic bias of the GLS of (4.3.2) is still smaller than that of the OLS or the within estimator of (4.3.2) (Sevestre and Trognon 1982). When $T$ tends to infinity, GLS of (4.3.2) is again consistent because GLS converges to the within (or LSDV) estimator, which becomes consistent.

It should also be noted that contrary to the static case, the feasible GLS is asymptotically less efficient than the GLS knowing the true covariance matrix because when a lagged dependent variable appears as one of the regressors, the estimation of slope coefficients is no longer asymptotically independent of the estimation of the parameters of the covariance matrix (Amemiya and Fuller 1967; Hsiao, Pesaran, and Tahmiscioglu (2002); or Appendix 4A).

[12] It should be noted that $y_{it}$ conditional on $y_{i,t-1}$ and $y_{i0}$ will not give a consistent estimator because $E(y_{i0}) = \theta_{i0}$. In other words, the residual will have mean different from 0 and the mean varies with $i$ will give rise the incidental parameters problem.

*4.3.3.3 Instrumental–Variable Estimator*

Because the likelihood functions under different initial conditions are different when dealing with panels involving large numbers of individuals over a short period of time, erroneous choices of initial conditions will yield estimators that are not asymptotically equivalent to the correct one, and hence may not be consistent. Sometimes we have little information to rely on in making a correct choice about the initial conditions. A simple consistent estimator that is independent of the initial conditions is appealing in its own right and in addition can be used to obtain initial values for the iterative process that yields the MLE. One estimation method consists of the following procedure.[13]

**Step 1:** Taking the first difference of (4.3.7), we obtain

$$y_{it} - y_{i,t-1} = \gamma(y_{i,t-1} - y_{i,t-2}) + \boldsymbol{\beta}'(\mathbf{x}_{it} - \mathbf{x}_{i,t-1}) + u_{it} - u_{i,t-1}. \quad (4.3.32)$$

Because $y_{i,t-2}$ or $(y_{i,t-2} - y_{i,t-3})$ are correlated with $(y_{i,t-1} - y_{i,t-2})$ but are uncorrelated with $(u_{it} - u_{i,t-1})$ they can be used as an instrument for $(y_{i,t-1} - y_{i,t-2})$ and estimate $\gamma$ and $\boldsymbol{\beta}$ by the instrumental-variable method. Both

$$\begin{aligned}\begin{pmatrix}\hat{\gamma}_{iv}\\ \hat{\boldsymbol{\beta}}_{iv}\end{pmatrix} = &\left[\sum_{i=1}^{N}\sum_{t=3}^{T}\right.\\ &\left.\cdot\begin{pmatrix}(y_{i,t-1}-y_{i,t-2})(y_{i,t-2}-y_{i,t-3}) & (y_{i,t-2}-y_{it-3})(\mathbf{x}_{it}-\mathbf{x}_{i,t-1})'\\ (\mathbf{x}_{it}-\mathbf{x}_{i,t-1})(y_{i,t-1}-y_{i,t-2}) & (\mathbf{x}_{it}-\mathbf{x}_{i,t-1})(\mathbf{x}_{it}-\mathbf{x}_{i,t-1})'\end{pmatrix}\right]^{-1}\\ &\cdot\left[\sum_{i=1}^{N}\sum_{t=3}^{T}\begin{pmatrix}y_{i,t-2}-y_{i,t-3}\\ \mathbf{x}_{it}-\mathbf{x}_{i,t-1}\end{pmatrix}(y_{it}-y_{i,t-1})\right], \quad (4.3.33)\end{aligned}$$

and

$$\begin{aligned}\begin{pmatrix}\tilde{\boldsymbol{\gamma}}_{iv}\\ \boldsymbol{\beta}_{iv}\end{pmatrix} = &\left[\sum_{i=1}^{N}\sum_{t=2}^{T}\right.\\ &\left.\cdot\begin{pmatrix}y_{i,t-2}(y_{i,t-1}-y_{i,t-2}) & y_{i,t-2}(\mathbf{x}_{it}-\mathbf{x}_{i,t-1})'\\ (\mathbf{x}_{it}-\mathbf{x}_{i,t-1})(y_{i,t-1}-y_{i,t-2}) & (\mathbf{x}_{it}-\mathbf{x}_{i,t-1})(\mathbf{x}_{it}-\mathbf{x}_{i,t-1})'\end{pmatrix}\right]^{-1}\\ &\cdot\left[\sum_{i=1}^{N}\sum_{t=2}^{T}\begin{pmatrix}y_{i,t-2}\\ \mathbf{x}_{it}-\mathbf{x}_{i,t-1}\end{pmatrix}(y_{it}-y_{i,t-1})\right], \quad (4.3.34)\end{aligned}$$

are consistent.

Both (4.3.33) and (4.3.34) are derived using the sample moments $\frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=2}^{T}\mathbf{q}_{it}(u_{it} - u_{i,t-1}) = 0$ to approximate the population moments $E[\mathbf{q}_{it}(u_{it} - u_{i,t-1})] = \mathbf{0}$, where $\mathbf{q}_{it} = [(y_{i,t-2} - y_{i,t-3}), (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})']'$ for (4.3.33) and $\mathbf{q}_{it} = [y_{i,t-2}, (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})']$ for (4.3.34). Therefore

[13] See Chapter 3, Section 3.5 for another approach.

(4.3.33) or (4.3.34) is a consistent estimator and $\sqrt{NT}[(\hat{\gamma}_{iv} - \gamma), (\hat{\boldsymbol{\beta}}_{iv} - \boldsymbol{\beta})']'$ is asymptotically normally distributed with mean 0, either $N$ or $T$ or both tend to infinity (in other words, there is no asymptotic bias).

Estimator (4.3.34) has an advantage over (4.3.33) in the sense that the minimum number of time periods required is 2, whereas (4.3.33) requires $T \geq 3$. In practice, if $T \geq 3$, the choice between (4.3.34) and (4.3.33) depends on the correlations between $(y_{i,t-1} - y_{i,t-2})$ and $y_{i,t-2}$ or $(y_{i,t-2} - y_{i,t-3})$. For a comparison of asymptotic efficiencies of the instruments $y_{i,t-2}$ or $(y_{i,t-2} - y_{i,t-3})$, see Anderson and Hsiao (1981).

**Step 2:** Substitute the estimated $\boldsymbol{\beta}$ and $\gamma$ into the equation

$$\overline{y}_i - \gamma \overline{y}_{i,-1} - \boldsymbol{\beta}' \overline{\mathbf{x}}_i = \boldsymbol{\rho}' \mathbf{z}_i + \alpha_i + \overline{u}_i \qquad i = 1, \ldots, N, \tag{4.3.35}$$

where $\overline{y}_i = \sum_{t=1}^{T} y_{it}/T$, $\overline{y}_{i,-1} = \sum_{t=1}^{T} y_{i,t-1}/T$, $\overline{\mathbf{x}}_i = \sum_{t=1}^{T} \mathbf{x}_{it}/T$, and $\overline{u}_i = \sum_{t=1}^{T} u_{it}/T$. Estimate $\boldsymbol{\rho}$ by the OLS method.

**Step 3:** Estimate $\sigma_u^2$ and $\sigma_\alpha^2$ by

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^{N} \sum_{t=2}^{T} \left[ (y_{it} - y_{i,t-1}) - \hat{\gamma}(y_{i,t-1} - y_{i,t-2}) - \hat{\boldsymbol{\beta}}'(\mathbf{x}_{it} - \mathbf{x}_{i,t-1}) \right]^2}{2N(T-1)}, \tag{4.3.36}$$

$$\hat{\sigma}_\alpha^2 = \frac{\sum_{i=1}^{N} \left( \overline{y}_i - \hat{\gamma} \overline{y}_{i,-1} - \hat{\boldsymbol{\rho}}' \mathbf{z}_i - \hat{\boldsymbol{\beta}}' \overline{\mathbf{x}}_i \right)^2}{N} - \frac{1}{T} \hat{\sigma}_u^2. \tag{4.3.37}$$

The consistency of these estimators is independent of initial conditions. The instrumental-variable estimators of $\gamma$, $\boldsymbol{\beta}$, and $\sigma_u^2$ are consistent when $N$ or $T$ or both tend to infinity. The estimators of $\boldsymbol{\rho}$ and $\sigma_\alpha^2$ are consistent only when $N$ goes to infinity. They are inconsistent if $N$ is fixed and $T$ tends to infinity. The instrumental-variable method is simple to implement. But if we also wish to test the maintained hypothesis on initial conditions in the random-effects model, it would seem more appropriate to rely on maximum-likelihood methods.

#### *4.3.3.4 Generalized Method of Moments Estimator*

We note that $y_{i,t-2}$ or $(y_{i,t-2} - y_{i,t-3})$ is not the only instrument for $(y_{i,t-1} - y_{i,t-2})$. In fact, as noted by Amemiya and MaCurdy (1986); Arellano–Bond (1991); Breusch, Mizon, and Schmidt (1989), etc. all $y_{i,t-2-j}$, $j = 0, 1, \ldots$ satisfy the conditions that $E[y_{i,t-2-j}(y_{i,t-1} - y_{i,t-2})] \neq 0$ and $E[y_{i,t-2-j}(u_{it} - u_{i,t-1})] = 0$. Therefore, they all are legitimate instruments for $(y_{i,t-1} - y_{i,t-2})$. Letting $\mathbf{q}_{it} = (y_{i0}, y_{i1}, \ldots, y_{i,t-2}, \mathbf{x}_i')'$, where $\mathbf{x}_i' = (\mathbf{x}_{i1}', \ldots, \mathbf{x}_{iT}')$, we have

$$E\mathbf{q}_{it} \Delta u_{it} = 0, \quad t = 2, \ldots, T. \tag{4.3.38}$$

Stacking the $(T-1)$ first differenced equation of (4.3.32) in matrix form we have

$$\Delta \mathbf{y}_i = \Delta \mathbf{y}_{i,-1}\gamma + \Delta X_i \boldsymbol{\beta} + \Delta \mathbf{u}_i, \quad i = 1, \ldots, N \tag{4.3.39}$$

where $\Delta \mathbf{y}_i$, $\Delta \mathbf{y}_{i,-1}$ and $\Delta \mathbf{u}_i$ are $(T-1)\times 1$ vectors of the form $(y_{i2}-y_{i1}, \ldots, y_{iT}-y_{i,T-1})'$, $(y_{i1}-y_{i0}, \ldots, y_{i,T-1}-y_{i,T-2})'$, $(u_{i2}-u_{i1}, \ldots, u_{iT}-u_{i,T-1})'$, respectively, and $\Delta X_i$ is the $(T-1)\times K_1$ matrix of $(\mathbf{x}_{i2}-\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}-\mathbf{x}_{i,T-1})'$. The $T(T-1)[K_1+\frac{1}{2}]$ orthogonality (or moment) conditions of (4.3.38) can be represented as

$$E W_i \Delta \mathbf{u}_i = \mathbf{0}, \tag{4.3.40}$$

where

$$W_i = \begin{pmatrix} \mathbf{q}_{i2} & \mathbf{0} & & \mathbf{0} \\ \mathbf{0} & \mathbf{q}_{i3} & \cdots & \mathbf{0} \\ \vdots & & \vdots & \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{q}_{iT} \end{pmatrix}, \tag{4.3.41}$$

is of dimension $[T(T-1)(K_1+\frac{1}{2})]\times(T-1)$. The dimension of (4.3.41) in general is much larger than $K_1+1$. Thus, Arellano–Bond (1991) suggest a generalized method of moments estimator (GMM).

The standard method of moments estimator consists of solving the unknown parameter vector $\boldsymbol{\theta}$ by equating the theoretical moments with their empirical counterparts or estimates. For instance, suppose that $\mathbf{m}(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta})$ denote some population moments of $\mathbf{y}$ and/or $\mathbf{x}$, say the first and second moments of $\mathbf{y}$ and/or $\mathbf{x}$, which are functions of the unknown parameter vector $\boldsymbol{\theta}$ and are supposed to equal some known constants, say 0. Let $\hat{\mathbf{m}}(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{m}(\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\theta})$ be their sample estimates based on $N$ independent samples of $(\mathbf{y}_i, \mathbf{x}_i)$. Then the method of moments estimator $\boldsymbol{\theta}$ is the $\hat{\boldsymbol{\theta}}_{mm}$, such that

$$\mathbf{m}(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta}) = \hat{\mathbf{m}}(\mathbf{y}, \mathbf{x}; \hat{\boldsymbol{\theta}}_{mm}) = \mathbf{0}. \tag{4.3.42}$$

For instance, the orthogonality conditions between $QX_i$ and $Q\mathbf{u}_i$ for the fixed-effects linear static model (3.2.2), $E(X_i' Q\mathbf{u}_i) = E[X_i' Q(\mathbf{y}_i - \mathbf{e}\alpha_i^* - X_i\boldsymbol{\beta})] = \mathbf{0}$, lead to the LSDV estimator (3.2.8). In this sense, the IV method is a method of moments estimator.

If the number of equations in (4.3.42) is equal to the dimension of $\boldsymbol{\theta}$, it is in general possible to solve for $\hat{\boldsymbol{\theta}}_{mm}$ uniquely. If the number of equations is greater than the dimension of $\boldsymbol{\theta}$, (4.3.42) in general has no solution. It is then necessary to minimize some norm (or distance measure) of $\hat{\mathbf{m}}(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta}) - \mathbf{m}(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta})$, say

$$[\hat{\mathbf{m}}(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta}) - \mathbf{m}(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta})]' A [\hat{\mathbf{m}}(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta}) - \mathbf{m}(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta})], \tag{4.3.43}$$

where $A$ is some positive definite matrix.

The property of the estimator thus obtained depends on $A$. The optimal choice of $A$ turns out to be

$$A^* = \left\{E[\hat{\mathbf{m}}(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta}) - \mathbf{m}(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta})][\hat{\mathbf{m}}(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta}) - \mathbf{m}(\mathbf{y}, \mathbf{x}; \boldsymbol{\theta})]'\right\}^{-1} \tag{4.3.44}$$

(Hansen 1982). The GMM estimator of $\boldsymbol{\theta}$ is to choose $\hat{\boldsymbol{\theta}}_{\text{GMM}}$ such that it minimizes (4.3.43) when $A = A^*$.

The Arellano–Bond (1991) GMM estimator of $\boldsymbol{\theta} = (\gamma, \boldsymbol{\beta}')'$ is obtained by minimizing

$$\left(\frac{1}{N}\sum_{i=1}^{N} \Delta\mathbf{u}_i' W_i'\right) \Psi^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} W_i \Delta\mathbf{u}_i\right), \tag{4.3.45}$$

with respect to $\boldsymbol{\theta}$, where $\Psi = E[\frac{1}{N^2}\sum_{i=1}^{N} W_i \Delta\mathbf{u}_i \Delta\mathbf{u}_i' W_i']$. Under the assumption that $u_{it}$ is i.i.d. with mean 0 and variance $\sigma_u^2$, $\Psi$ can be approximated by $\frac{\sigma_u^2}{N^2}\sum_{i=1}^{N} W_i \tilde{A} W_i'$, where

$$\underset{(T-1)\times(T-1)}{\tilde{A}} = \begin{bmatrix} 2 & -1 & 0 & . & 0 \\ -1 & 2 & -1 & . & 0 \\ 0 & \ddots & \ddots & & \\ 0 & \ddots & \ddots & . & -1 \\ 0 & & . & -1 & 2 \end{bmatrix}. \tag{4.3.46}$$

Thus, the Arellano and Bond GMM estimator takes the form

$$\begin{aligned} &\hat{\boldsymbol{\theta}}_{\text{GMM,AB}} \\ &= \left\{\left[\sum_{i=1}^{N} \begin{pmatrix} \Delta\mathbf{y}_{i,-1}' \\ \Delta X_i' \end{pmatrix} W_i'\right]\left[\sum_{i=1}^{N} W_i \tilde{A} W_i'\right]^{-1}\left[\sum_{i=1}^{N} W_i(\Delta\mathbf{y}_{i,-1}, \Delta X_i)\right]\right\}^{-1} \\ &\quad \cdot \left\{\left[\sum_{i=1}^{N} \begin{pmatrix} \Delta\mathbf{y}_{i,-1}' \\ \Delta X_i' \end{pmatrix} W_i'\right]\left[\sum_{i=1}^{N} W_i \tilde{A} W_i'\right]^{-1}\left[\sum_{i=1}^{N} W_i \Delta\mathbf{y}_i\right]\right\}, \end{aligned} \tag{4.3.47}$$

with asymptotic covariance matrix

$$\begin{aligned} &\text{Cov}(\hat{\boldsymbol{\theta}}_{\text{GMM,AB}}) \\ &= \sigma_u^2 \left\{\left[\sum_{i=1}^{N} \begin{pmatrix} \Delta\mathbf{y}_{i,-1}' \\ \Delta X_i' \end{pmatrix} W_i'\right]\left[\sum_{i=1}^{N} W_i \tilde{A} W_i'\right]^{-1}\left[\sum_{i=1}^{N} W_i(\Delta\mathbf{y}_{i,-1}, \Delta X_i)\right]\right\}^{-1}. \end{aligned} \tag{4.3.48}$$

In addition to the moment conditions (4.3.38), Arellano and Bover (1995) also note that $E\bar{v}_i = 0$, where $\bar{v}_i = \bar{y}_i - \bar{y}_{i,-1}\gamma - \bar{\mathbf{x}}_i'\boldsymbol{\beta} - \boldsymbol{\rho}'\mathbf{z}_i$.[14] Therefore, if instruments $\tilde{\mathbf{q}}_i$ exist (for instance, the constant 1 is a valid instrument) such that

$$E\tilde{\mathbf{q}}_i\bar{v}_i = \mathbf{0}, \tag{4.3.49}$$

then a more efficient GMM estimator can be derived by incorporating this additional moment condition.

Apart from the linear moment conditions (4.3.40), and (4.3.49), Ahn and Schmidt (1995) note that the homoscedasticity condition of $E(v_{it}^2)$ implies the following $T-2$ linear conditions:

$$E(y_{it}\Delta u_{i,t+1} - y_{i,t+1}\Delta u_{i,t+2}) = 0, \quad t = 1, \ldots, T-2. \tag{4.3.50}$$

Combining (4.3.40), (4.3.49), and (4.3.50), a more efficient GMM estimator can be derived by minimizing[15]

$$\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{u}_i^{+\prime}W_i^{+\prime}\right)\Psi^{+-1}\left(\frac{1}{N}\sum_{i=1}^{N}W_i^{+}\mathbf{u}_i^{+}\right) \tag{4.3.51}$$

with respect to $\boldsymbol{\theta}$, where $\mathbf{u}_i^{+} = (\Delta\mathbf{u}_i', \bar{v}_i)'$, $\Psi^{+} = E\left(\frac{1}{N^2}\sum_{i=1}^{N}W_i^{+}\mathbf{u}_i^{+}\mathbf{u}_i^{+\prime}W_i^{+\prime}\right)$, and

$$W_i^{+'} = \begin{pmatrix} W_i' & W_i^{*\prime} & \mathbf{0} \\ \mathbf{0}' & \mathbf{0}' & \tilde{\mathbf{q}}_i' \end{pmatrix}$$

where

$$\underset{(T-2)\times(T-1)}{W_i^{*}} = \begin{pmatrix} y_{i1} & -y_{i2} & 0 & 0 & \ldots & 0 \\ 0 & y_{i2} & -y_{i3} & 0 & \ldots & . \\ & & & & & \\ & & & & & 0 \\ & & & 0 & y_{i,T-2} & -y_{i,T-1} \end{pmatrix}.$$

However, because the covariance matrix (4.3.50) depends on the unknown $\boldsymbol{\theta}$, it is impractical to implement the GMM. A less efficient but computationally feasible GMM estimator is to ignore the information that $\Psi^{+}$ also depends on $\boldsymbol{\theta}$ and simply substitute $\Psi^{+}$ by its consistent estimator

$$\hat{\Psi}^{+} = \left(\frac{1}{N^2}\sum_{i=1}^{N}W_i^{+}\hat{\mathbf{u}}_i^{+}\hat{\mathbf{u}}_i^{+\prime}W_i^{+'}\right) \tag{4.3.52}$$

[14] Note that we let $\mathbf{z}_i = 0$ for ease of exposition. When $\mathbf{z}_i$ is present, the first differencing step of (4.3.38) eliminates $\mathbf{z}_i$ from the specification; hence the moment conditions (4.3.39) remain valid. However, for $Ev_i = 0$ to hold, it requires the assumption of stationarity in mean (Blundell and Bond 1998).

[15] For ease of notation, we again assume that $\mathbf{z}_i = \mathbf{0}$.

into the objective function (4.3.51) to derive a linear estimator of form (4.3.47) where $\hat{\mathbf{u}}_i^+$ is derived by using some simple consistent estimator of $\gamma$ and $\boldsymbol{\beta}$, say the IV discussed in Section 4.3.3.3, into (4.3.39) and the $\bar{v}_i$ equation.

In principle, one can improve the asymptotic efficiency of the GMM type estimator by adding more moment conditions. For instance, Ahn and Schmidt (1995) note that in addition to the linear moment conditions of (4.3.40), (4.3.49), and (4.3.50), there exist $(T-1)$ nonlinear moment conditions of the form $E((\bar{y}_i - \boldsymbol{\beta}'\bar{\mathbf{x}}_i)\Delta u_{it}) = 0, t = 2, \ldots, T$, implied by the homoscedasticity conditions of $Ev_{it}^2$. Under the additional assumption that $E(\alpha_i y_{it})$ is the same for all $t$, this condition and condition (4.3.50) can be transformed into the $(2T-2)$ linear moment conditions

$$E[(y_{iT} - \boldsymbol{\beta}'\mathbf{x}_{iT})\Delta y_{it}] = 0, \quad t = 1, \ldots, T-1, \tag{4.3.53}$$

and

$$E[(y_{it} - \boldsymbol{\beta}'\mathbf{x}_{it})y_{it} - (y_{i,t-1} - \boldsymbol{\beta}'\mathbf{x}_{i,t-1})y_{i,t-1}] = 0, \quad t = 2, \ldots, T. \tag{4.3.54}$$

Though theoretically it is possible to add additional moment conditions to improve the asymptotic efficiency of GMM, it is doubtful how much efficiency gain one can achieve by using a huge number of moment conditions in a finite sample. Moreover, if higher moment conditions are used, the estimator can be very sensitive to outlying observations. Through a simulation study, Ziliak (1997) has found that the downward bias in GMM is quite severe as the number of moment conditions expands, outweighing the gains in efficiency. The strategy of exploiting all the moment conditions for estimation is actually not recommended for panel-data applications in finite sample, owing mainly to bias. The bias is proportional to the number of instruments for each equation. In addition, when $\gamma$ is close to 1, the lagged instruments $y_{i,t-2-j}$, $j \geq 0$ are weak instruments. There is also a bias-variance tradeoff in the number of moment conditions used for estimation. Koenken and Machado (1999) show that the usual asymptotic theory holds only if the number of moments used is less than the cubic root of the sample size. Okui (2009) proposes a moment selection method based on minimizing (Nagar 1959) the approximate mean square error. In general, when $T$ is small, it is optimal to use all moment conditions. When $T$ is not very small ($\frac{T^2}{N \log T} \longrightarrow \infty$), the optimal number of moment conditions chosen is $O((NT)^{1/3})$ assuming there exists a natural rank ordering of instruments for each $(y_{i,t-1} - y_{i,t-2})$, say $(y_{i0}, y_{i1}, \ldots, y_{i,t-2})$ in increasing order. (Actually, Okui (2009) derives his selection method using the forward orthogonal deviation operator of Arellano and Bover (1995), $\Delta u_{it}^* = \sqrt{\frac{T-t}{T-t+1}}\left[u_{it} - \frac{1}{T-t}(u_{i,t+1} + \ldots + u_{iT})\right]$.) When $\sigma_\alpha^2$ is large relative to $\sigma_u^2$, it is also advisable to use many moment conditions. For further discussions, see Judson and Owen (1999), Kiviet (1995), and Wansbeek and Bekker (1996).

To improve the efficiency of GMM when $\gamma$ is close to 1, Hahn, Hausman, and Kuersteiner (2007) suggest not using the first difference equation as in (4.3.32), but to use the long difference $y_{iT} - y_{i1}$. In the case of first-order autoregressive process (4.3.3),

$$y_{iT} - y_{i1} = \gamma(y_{i,T-1} - y_{io}) + (u_{iT} - u_{i1}). \tag{4.3.55}$$

Then $y_{io}, y_{i,T-1} - \gamma y_{i,T-2}, \ldots, y_{i2} - \gamma y_{i1}$ are valid instruments. Their long difference (LD) estimator is equivalent to applying the GMM based on the "reduced set" of moment conditions

$$E\begin{pmatrix} y_{io} \\ y_{iT-1} - \gamma y_{i,T-2} \\ \cdot \\ \cdot \\ \cdot \\ y_{i2} - \gamma y_{i1} \end{pmatrix} [(y_{iT} - y_{i1}) - \gamma(y_{i,T-1} - y_{io})] = \mathbf{0}. \tag{4.3.56}$$

The instruments $y_{it} - \gamma y_{i,t-1}$ for $t = 2, \ldots, T-1$ require knowledge of $\gamma$. A feasible LD estimator could be to use the Arellano–Bond GMM estimator (4.3.47) to obtain a preliminary consistent estimator $\hat{\gamma}_{\text{GMM,AB}}$, then use $(y_{io}, y_{i,T-1} - y_{i,T-2}\hat{\gamma}_{\text{GMM,AB}}, \ldots, y_{i2} - y_{i1}\hat{\gamma}_{\text{GMM,AB}})$ as instruments.

The reason that the LD estimator can improve the efficiency of the GMM based on the first difference equation of (4.3.3) is because GMM can be viewed as the two-stage least-squares method (Theil 1958). As shown by Donald and Newey (2001), the bias of 2SLS (GMM) depends on four factors: "explained" variance of the first stage reduced form equation, "covariance" between the stochastic disturbance of the structural equation and the reduced form equation, the number of instruments, and the sample size,

$$E\,[\hat{\gamma}_{\text{2SLS}} - \gamma] \simeq \frac{1}{n}a, \tag{4.3.57}$$

where $n$ denotes the sample size and

$$a = \frac{\text{(number of instruments)} \times \text{("covariance")}}{\text{"Explained" variance of the first stage reduced form equation}} \tag{4.3.58}$$

Based on this formula, Hahn, Hausman, and Kuersteiner (2007) show that $a = -\frac{1+\gamma}{1-\gamma}$ for the Arellano–Bond (1991) GMM estimator when $T = 3$. When $\gamma = .9$, $a = -19$. For $N = 100$, this implies a percentage bias of $-105.56$. On the other hand, using the LD estimator, $a = -.37$, which is much smaller than $-19$ in absolute magnitude.

**Remark 4.3.1:** We derive the MLE (or GLS) or the GMM estimator (4.3.47) assuming that $u_{it}$ is independently distributed across $i$ and over $t$. If $u_{it}$ is serially correlated, $E(y_{i,t-2}\Delta u_{it}) \neq 0$ for $j \geq 2$. Then neither (4.3.30) nor (4.3.47) is a consistent estimator. On the other hand, the estimator $\hat{\boldsymbol{\theta}}^*$ that replaces $W_i$ in (4.3.47) by the block diagonal instrument matrix $\tilde{W}_i^*$ whose

$t$th block is given by $\mathbf{x}_i$ if $\mathbf{x}_{it}$ is strictly exogenous (i.e., $E\mathbf{x}_{it}u_{is} = 0$ for all $s$) or $(\mathbf{x}'_{it}, \mathbf{x}'_{i,t-1}, \ldots, \mathbf{x}'_{i1})'$ if $\mathbf{x}_{it}$ is weakly exogenous (i.e., $E(\mathbf{x}_{i,t+j+1}u_{it}) \neq 0$ and $E(u_{it}\mathbf{x}_{i,t-j}) = 0$ for $j \geq 0$)) remains consistent. Therefore, a Hausman (1978) type test statistic can be constructed to test if $u_{it}$ are serially uncorrelated by comparing the difference of $(\hat{\boldsymbol{\theta}}_{\text{GMM,AB}} - \hat{\boldsymbol{\theta}}^*)$.

Arellano–Bond (1991) note that if $u_{it}$ is not serially correlated, $E(\Delta u_{it}\Delta u_{i,t-2}) = 0$. They show that the statistic

$$\frac{\sum_{i=1}^{N}\sum_{t=4}^{T}\Delta\hat{u}_{it}\Delta\hat{u}_{i,t-2}}{\hat{s}} \tag{4.3.59}$$

is asymptotically normally distributed with mean 0 and variance 1 when $T \geq 5$ and $N \longrightarrow \infty$, where

$$\begin{aligned}\hat{s}^2 = &\sum_{i=1}^{N}\left(\sum_{t=4}^{T}\Delta\hat{u}_{it}\Delta\hat{u}_{i,t-2}\right)^2 - 2\left(\sum_{i=1}^{N}\sum_{t=4}^{T}\Delta\hat{u}_{i,t-2}\Delta\mathbf{x}'_{it}\right)\\
&\cdot\left\{\left[\sum_{i=1}^{N}\begin{pmatrix}\Delta\mathbf{y}'_{i,-1}\\ \Delta X'_i\end{pmatrix}W'_i\right]\left(\frac{1}{N}\sum_{i=1}^{N}W_i\tilde{A}W'_i\right)^{-1}\left[\sum_{i=1}^{N}W_i(\Delta\mathbf{y}_{i,-1},\Delta X_i)\right]\right\}^{-1}\\
&\cdot\left[\sum_{i=1}^{N}\begin{pmatrix}\Delta\mathbf{y}'_{i,-1}\\ \Delta X'_i\end{pmatrix}W'_i\right]\left(\frac{1}{N}\sum_{i=1}^{N}W_i\tilde{A}W'_i\right)^{-1}\left[\sum_{i=1}^{N}W_i\Delta\hat{\mathbf{u}}_i\left(\sum_{t=4}^{T}\Delta\hat{u}_{it}\Delta\hat{u}_{i,t-2}\right)\right]\\
&+\left(\sum_{i=1}^{N}\sum_{t=4}^{T}\Delta\hat{u}_{i,t-2}\Delta\mathbf{x}'_{it}\right)(\text{Cov}\,(\hat{\boldsymbol{\theta}}_{\text{GMM,AB}}))\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\Delta\mathbf{x}_{it}\Delta\hat{u}_{i-t-2}\right),\end{aligned} \tag{4.3.60}$$

where $\Delta\hat{u}_{it} = \Delta y_{it} - (\Delta y_{i,t-1}, \Delta x'_{it})\hat{\boldsymbol{\theta}}_{\text{GMM,AB}}$, $\Delta\hat{\mathbf{u}}_i = (\Delta\hat{u}_{i2}, \ldots, \Delta\hat{u}_{iT})'$.

This statistic in lieu of Hausman-type test statistic can be used to test serial correlation in the case there exist no exogenous variables for model (4.3.7). The statistic (4.3.59) is defined only if $T \geq 5$. When $T < 5$, Arellano–Bond (1991) suggest using the Sargan (1958) test of overidentification,

$$\left(\sum_{i=1}^{N}\hat{\Delta\mathbf{u}}'_i W_i^{*'}\right)\left(\sum_{i=1}^{N}W_i^*\hat{\Delta\mathbf{u}}_i\hat{\Delta\mathbf{u}}'_i W_i^{*'}\right)^{-1}\left(\sum_{i=1}^{N}W_i^*\hat{\Delta\mathbf{u}}_i\right), \tag{4.3.61}$$

where $W_i^*$ could be $W_i$ or any number of instruments that satisfy the orthogonality condition $E(W_i^*\Delta\mathbf{u}_i) = 0$. Under the null of no serial correlation, (4.3.61) is asymptotically $\chi^2$ distributed with $p - (K+1)$, degrees of freedom for any $p > (K+1)$, where $p$ denotes the number of rows in $W_i^*$.

**Remark 4.3.2:** Because the individual-specific effects $\alpha_i$ are time-invariant, taking the deviation of individual $y_{it}$ equation from any transformation of $y_{it}$ equation that maintains the time-invariance property of $\alpha_i$ can eliminate $\alpha_i$.

For instance, Alvarez and Arellano (2003) consider the transformation of $y_{it}$ equation into an equation of the form,

$$c_t[y_{it} - \frac{1}{(T-t)}(y_{i,t+1} + \cdots + y_{iT})], \quad t = 1, \ldots, T-1, \quad (4.3.62)$$

where $c_t^2 = \frac{(T-t)}{(T-t+1)}$. The advantage of considering the equation specified by transformation (4.3.62) is that the residuals $u_{it}^* = c_t[u_t - \frac{1}{(T-t)}(u_{i,t+1} + \ldots + u_{iT})], t = 1, \ldots, T-1$ are orthogonal, that is, $Eu_{it}^* u_{is}^* = 0$ if $t \neq s$ and $Eu_{it}^{*2} = \sigma_u^2$. However, if transformation (4.3.62) is used to remove $\alpha_i$, the instruments $\mathbf{q}_{it}$ takes the form $(y_{i0}, \ldots, y_{i,t-1}, \mathbf{x}_i')$ in the application of GMM.

### 4.3.4 Testing Some Maintained Hypotheses on Initial Conditions

As discussed in Sections 4.3.2 and 4.3.3, the interpretation and consistency property for the MLE and GLS of a random-effects model depend on the initial conditions. Unfortunately, in practice we have very little information on the characteristics of the initial observations. Because some of these hypotheses are nested, Bhargava and Sargan (1983) suggest relying on the likelihood principle to test them. For instance, when $y_{i0}$ are exogenous (Case I) we can test the validity of the error-components formulation by maximizing $L_1$ with or without the restrictions on the covariance matrix $V$. Let $L_1^*$ denote the maximum of log $L_1$ subject to the restriction of model (4.3.7), and let $L_1^{**}$ denote the maximum of log $L_1$ with $V$ being an arbitrary positive definite matrix. Under the null hypothesis, the resulting test statistic $2(L_1^{**} - L_1^*)$ is asymptotically $\chi^2$ distributed, with $[T(T+1)/2 - 2]$ degrees of freedom.

Similarly, we can test the validity of the error-components formulation under the assumption that $y_{i0}$ are endogenous. Let the maximum of the log likelihood function under Case IVa and Case IVc′ be denoted by $L_{4a}^*$ and $L_{4c'}^*$, respectively. Let the maximum of the log likelihood function under case IVa or IVc′ without the restriction (4.3.20) or (4.3.24) [namely, the $(T+1) \times (T+1)$ covariance matrix is arbitrary] be denoted by $L_{4a}^{**}$ or $L_{4c'}^{**}$, respectively. Then, under the null, $2(L_{4a}^{**} - L_{4a}^*)$ and $2(L_{4c'}^{**} - L_{4c'}^*)$ are asymptotically $\chi^2$, with $[(T+1)(T+2)/2 - 2]$ and $[(T+1)(T+2)/2 - 3]$ degrees of freedom, respectively.

To test the stationarity assumption, we denote the maximum of the log likelihood function for Case IVb and Case IVd′ as $L_{4b}^*$ and $L_{4d'}^*$, respectively. Then $2(L_{4b}^* - L_{4a}^*)$ and $2(L_{4d'}^* - L_{4c'}^*)$ are asymptotically $\chi^2$, with 1 degree of freedom. The statistics $2(L_{4a}^{**} - L_{4b}^*)$ and $2(L_{4c'}^{**} - L_{4d'}^*)$ can also be used to test the validity of Case IVb and Case IVd′, respectively. They are asymptotically $\chi^2$ distributed, with $[(T+1)(T+2)/2 - 3]$ and $[(T+1)(T+2)/2 - 4]$ degrees of freedom, respectively.

We can also generalize the Bhargava and Sargan principle to test the assumption that the initial observations have a common mean $\mu_w$ or have different means $\theta_{i0}$ under various assumptions about the error process. The statistics $2[L_{4c'}^* - L_{4a}^*]$, $2[L_{4c'}^{**} - L_{4a}^{**}]$, or $2[L_{4d'}^* - L_{4b}^*]$ are asymptotically $\chi^2$

distributed, with $q$, $(q-1)$, and $(q-1)$ degrees of freedom, respectively, where $q$ is the number of unknown coefficients in (4.3.22). We can also test the combined assumption of a common mean and a variance-components formulation by using the statistic $2[L^{**}_{4c'} - L^{*}_{4a}]$ or $2[L^{**}_{4c'} - L^{*}_{4b}]$, both of which are asymptotically $\chi^2$ distributed, with $q + (T+1)(T+2)/2 - 3$ and $q + (T+1)(T+2)/2 - 4$ degrees of freedom, respectively.

With regard to testing that $y_{i0}$ are exogenous, unfortunately it is not possible to directly compare $L_1$ with the likelihood functions of various forms of case IV, because in the former case we are considering the density of $(y_{i1}, \ldots, y_{iT})$ assuming $y_{i0}$ to be exogenous, whereas the latter case is the joint density of $(y_{i0}, \ldots, y_{iT})$. However, we can write the joint likelihood function of (4.3.7) and (4.3.22) under the restriction that $v_{i0}$ are independent of $\eta_i$ (or $\alpha_i$) and have variance $\sigma^2_{\epsilon o}$. Namely, we impose the restriction that $\mathrm{Cov}(v_{i0}, v_{it}) = 0, t = 1, \ldots, T$, in the $(T+1) \times (T+1)$ variance–covariance matrix of $(y_{i0}, \ldots, y_{iT})$. We denote this likelihood function by $L_5$. Let $L^{**}_5$ denote the maximum of log $L_5$ with unrestricted variance–covariance matrix for $(v_{i0}, \ldots, v_{iT})$. Then we can test the exogeneity of $y_{i0}$ using $2(L^{**}_{4c'} - L^{**}_5)$, which is asymptotically $\chi^2$ with $T$ degrees of freedom under the null.

It is also possible to test the exogeneity of $y_{i0}$ by constraining the error terms to have a variance-components structure. Suppose the variance–covariance matrix of $(v_{i1}, \ldots, v_{iT})$ is of the form $V$ [equation (3.3.4)]. Let $L^{*}_5$ denote the maximum of the log likelihood function $L_5$ under this restriction. Let $L^{*}_{4d'}$ denote the maximum of the log likelihood function of $(y_{i0}, \ldots, y_{iT})$ under the restriction that $E\mathbf{v}_i\mathbf{v}_i' = \tilde{V}^*$, but allowing the variance of $v_{i0}$ and the covariance between $v_{i0}$ and $v_{it}, t = 1, \ldots, T$, to be arbitrary constants $\sigma^2_{w0}$ *and* $\sigma^2_{\tau}$. The statistic $2(L^{*}_{4d'} - L^{*}_5)$ is asymptotically $\chi^2$ with 1 degree of freedom if $y_{i0}$ are exogenous. In practice, however, it may not even be necessary to calculate $L^{*}_{4d'}$, because $L^{*}_{4d'} \geq L^{*}_{4c'}$, and if the null is rejected using $2(L^{*}_{4c'} - L^{*}_5)$ against the critical value of $\chi^2$ with 1 degree of freedom, then $2(L^{*}_{4d'} - L^{**}_5)$ must also reject the null.

### 4.3.5 Simulation Evidence

To investigate the performance of maximum-likelihood estimators under various assumptions about the initial conditions, Bhargava and Sargan (1983) conducted Monte Carlo studies. Their true model was generated by

$$y_{it} = 1 + 0.5y_{i,t-1} - 0.16z_i + 0.35x_{it} + \alpha_i + u_{it}, \quad i = 1, \ldots, 100, \quad t = 1, \ldots, 20, \tag{4.3.63}$$

where $\alpha_i$ and $u_{it}$ were independently normally distributed, with means 0 and variances 0.09 and 0.4225, respectively. The time-varying exogenous variables

$x_{it}$ were generated by

$$x_{it} = 0.1t + \phi_i x_{i,t-1} + \omega_{it}, \qquad \begin{matrix} i = 1, \ldots, 100, \\ t = 1, \ldots, 20, \end{matrix} \tag{4.3.64}$$

with $\phi_i$ and $\omega_{it}$ independently normally distributed, with means 0 and variances 0.01 and 1, respectively. The time-invariant exogenous variables $z_i$ were generated by

$$z_i = -0.2x_{i4} + \omega_i^*, \qquad i = 1, \ldots, 100, \tag{4.3.65}$$

and $\omega_i^*$ were independently normally distributed, with mean 0 and variance 1. The $z$ and the $x$ were held fixed over the replications, and the first 10 observations were discarded. Thus, the $y_{i0}$ are in fact stochastic and are correlated with the individual effects $\alpha_i$. Table 4.2 reproduces their results on the biases in the estimates for various models obtained in 50 replications.

In cases where the $y_{i0}$ are treated as endogenous, the MLE performs extremely well, and the biases in the parameters are almost negligible. But this is not so for the MLE where $y_{i0}$ are treated as exogenous. The magnitude of the bias is about 1 standard error. The boundary solution of $\sigma_\alpha^2 = 0$ occurs in a number of replications for the error-components formulation as well. The likelihood-ratio statistics also rejected the exogeneity of $y_{i0}$ 46 and 50 times, respectively, using the tests $2[L_{4c'}^{**} - L_5^{**}]$ and $2[L_{4c'}^{*} - L_5^{*}]$. Under the endogeneity assumption, the likelihood-ratio statistic $2(L_{4c'}^{**} - L_{4c'}^{*})$ rejected the error-components formulation 4 times (out of 50), whereas under the exogeneity assumption, the statistic $2(L_1^{**} - L_1^{*})$ rejected the error-components formulation 7 times.[16]

## 4.4 AN EXAMPLE

We have discussed the properties of various estimators for dynamic models with individual-specific effects. In this section we report results from the study of demand for natural gas conducted by Balestra and Nerlove (1966) to illustrate the specific issues involved in estimating dynamic models using observations drawn from a time series of cross sections.

Balestra and Nerlove (1966) assumed that the new demand for gas (inclusive of demand due to the replacement of gas appliances and the demand due to net increases in the stock of such appliances), $G^*$, was a linear function of the relative price of gas, $P$, and the total new requirements for all types of fuel, $F^*$. Let the depreciation rate for gas appliances be $r$, and assume that the rate of utilization of the stock of appliances is constant; the new demand for gas and the gas consumption at year $t$, $G_t$, follow the relation

$$G_t^* = G_t - (1 - r)G_{t-1}. \tag{4.4.1}$$

[16] Bhargava and Sargan (1983) did not report the significance level of their tests. Presumably they used the conventional 5 percent significance level.

Table 4.2. *Simulation results for the biases of the MLEs for dynamic random-effects models*

| Coefficient of | $y_{i0}$ exogenous, unrestricted covariance matrix | $y_{i0}$ exogenous, error-components formulation | $y_{i0}$ endogenous, unrestricted covariance matrix | $y_{i0}$ endogenous, error-components formulation |
|---|---|---|---|---|
| Intercept | −0.1993<br>(0.142)[a] | −0.1156<br>(0.1155) | −0.0221<br>(0.1582) | 0.0045<br>(0.105) |
| $z_i$ | 0.0203<br>(0.0365) | 0.0108<br>(0.0354) | 0.0007<br>(0.0398) | −0.0036<br>(0.0392) |
| $x_{it}$ | 0.0028<br>(0.0214) | 0.0044<br>(0.0214) | 0.0046<br>(0.0210) | 0.0044<br>(0.0214) |
| $y_{i,t-1}$ | 0.0674<br>(0.0463) | 0.0377<br>(0.0355) | 0.0072<br>(0.0507) | −0.0028<br>(0.0312) |
| $\sigma_\alpha^2 / \sigma_u^2$ | | −0.0499<br>(0.0591) | | 0.0011<br>(0.0588) |

[a] Means of the estimated standard errors in parentheses.

*Source:* Bhargava and Sargan (1983).

They also postulated a similar relation between the total new demand for all types of fuel and the total fuel consumption, $F$, with $F$ approximated by a linear function of total population, $N$, and per capita income, $I$. Substituting these relations into (4.4.1), they obtained

$$G_t = \beta_0 + \beta_1 P_t + \beta_2 \Delta N_t + \beta_3 N_{t-1} + \beta_4 \Delta I_t + \beta_5 I_{t-1} + \beta_6 G_{t-1} + v_t, \tag{4.4.2}$$

where $\Delta N_t = N_t - N_{t-1}$, $\Delta I_t = I_t - I_{t-1}$, and $\beta_6 = 1 - r$.

Balestra and Nerlove used annual U.S. data from 36 states over the period 1957–67 to estimate the model for residential and commercial demand for natural gas (4.4.2). Because the average age of the stock of gas appliances during this period was relatively young, it was expected that the coefficient of the lagged gas consumption variable, $\beta_6$, would be less than 1, but not too much below 1. The OLS estimates of (4.4.2) are reported in the second column of Table 4.3. The estimated coefficient of $G_{t-1}$ is 1.01. It is clearly incompatible with a priori theoretical expectations, as it implies a negative depreciation rate for gas appliances.

One possible explanation for the foregoing result is that when cross-sectional and time series data are combined in the estimation of (4.4.2), certain effects specific to the individual state may be present in the data. To account for such effects, dummy variables corresponding to the 36 different states were introduced into the model. The resulting dummy variable estimates are shown in the

Table 4.3. *Various estimates of the parameters of Balestra and Nerlove's demand-for-gas model (4.4.2) from the pooled sample, 1957–1962*

| Coefficient | OLS | LSDV | GLS |
|---|---|---|---|
| $\beta_0$ | −3.650<br>(3.316)[a] | — | −4.091<br>(11.544) |
| $\beta_1$ | −0.0451<br>(0.0270) | −0.2026<br>(0.0532) | −0.0879<br>(0.0468) |
| $\beta_2$ | 0.0174<br>(0.0093) | −0.0135<br>(0.0215) | −0.00122<br>(0.0190) |
| $\beta_3$ | 0.00111<br>(0.00041) | 0.0327<br>(0.0046) | 0.00360<br>(0.00129) |
| $\beta_4$ | 0.0183<br>(0.0080) | 0.0131<br>(0.0084) | 0.0170<br>(0.0080) |
| $\beta_5$ | 0.00326<br>(0.00197) | 0.0044<br>(0.0101) | 0.00354<br>(0.00622) |
| $\beta_6$ | 1.010<br>(0.014) | 0.6799<br>(0.0633) | 0.9546<br>(0.0372) |

[a] Figures in parentheses are standard errors for the corresponding coefficients.
*Source:* Balestra and Nerlove (1966).

third column of Table 4.3. The estimated coefficient of the lagged endogenous variable is drastically reduced; in fact, it is reduced to such a low level that it implies a depreciation rate of gas appliances of greater than 30 percent – again highly implausible.

Instead of assuming the regional effect to be fixed, they again estimated (4.4.2) by explicitly incorporating individual state-specific effects into the error term, so that $v_{it} = \alpha_i + u_{it}$, where $\alpha_i$ and $u_{it}$ are independent random variables. The two-step GLS estimates under the assumption that the initial observations are fixed are shown in the fourth column of Table 4.3. The estimated coefficient of lagged consumption is 0.9546. The implied depreciation rate is approximately 4.5 percent, which is in agreement with a priori expectation.

The foregoing results illustrate that by properly taking account of the unobserved heterogeneity in the panel data, Balestra and Nerlove were able to obtain results that were reasonable on the basis of a priori theoretical considerations that they were not able to obtain through attempts to incorporate other variables into the equation by conventional procedures. Moreover, the least-squares and the least-squares dummy variables estimates of the coefficient of the lagged gas consumption variable were 1.01 and 0.6799, respectively. In previous sections we showed that for dynamic models with individual-specific effects, the least-squares estimate of the coefficient of the lagged dependent variable is biased upward and the least-squares dummy variable estimate is biased downward if $T$ is small. Their estimates are in agreement with these theoretical results.[17]

## 4.5 FIXED-EFFECTS MODELS

If individual effects are considered fixed and different across individuals, because of strict multicollinearity between the effects and other time-invariant variables, there is no way one can disentangle the individual-specific effects from the impact of other time-invariant variables. We shall therefore assume $\mathbf{z}_i \equiv \mathbf{0}$. When $T$ tends to infinity, even though lagged $y$ does not satisfy the strict exogeneity condition for the regressors, it does satisfy the weak exogeneity condition of $E(u_{it} \mid y_{i,t-1}, y_{i,t-2}, .; \alpha_i) = 0$; hence the least-squares regression of $y_{it}$ on lagged $y_{i,t-j}$ and $\mathbf{x}_{it}$ and the individual-specific constant yields a consistent estimator. In the case that $T$ is fixed and $N$ tends to infinity, the number of parameters in a fixed-effects specification increases with

[17] We do not know the value of the GLS estimates when the initial observations are treated as endogenous. My conjecture is that it is likely to be close to the two-step GLS estimates with fixed initial observations. As mentioned in Chapter 4, Section 4.3, Sevestre and Trognon (1982) have shown that even the initial values are correlated with the effects; the asymptotic bias of the two-step GLS estimator under the assumption of fixed initial observations is still smaller than the OLS or the within estimator. Moreover, if Bhargava and Sargan's simulation result is any indication, the order of bias due to the wrong assumption about initial observations when $T$ is greater than 10 is about 1 standard error or less. Here, the standard error of the lagged dependent variable for the two-step GLS estimates with fixed initial values is only 0.037.

the number of cross-sectional observations. This is the classical incidental parameters problem (Neyman and Scott 1948). In a static model with strict exogeneity assumption, the presence of individual specific constants does not affect the consistency of the CV or MLE estimator of the slope coefficients (see Chapter 3). However, the result no longer holds if lagged dependent variables also appear as explanatory variables. The regularity conditions for the consistency of the MLE are violated. In fact, if $u_{it}$ are normally distributed and $y_{i0}$ are given constants, the MLE of (4.2.1) is the CV of (4.2.2) and (4.2.3). The asymptotic bias is given by (4.2.8).

While the MLE is inconsistent when $T$ is fixed and $N$ is large, the IV estimator of (4.3.32) or the GMM estimator (4.3.43) remains consistent and asymptotically normally distributed with fixed $\alpha_i^*$. The transformed equation (4.3.39) does not involve the incidental parameters $\alpha_i^*$. The orthogonality condition (4.3.40) remains valid.

In addition to the IV type estimator, a likelihood-based approach based on a transformed likelihood function can also yield a consistent and asymptotically normally distributed estimator.

### 4.5.1 Transformed Likelihood Approach

The first difference equation (4.3.32) no longer contains the individual effects $\alpha_i^*$ and is well defined for $t = 2, 3, \ldots, T$, under the assumption that the initial observations $y_{i0}$ and $\mathbf{x}_{i0}$ are available. But (4.3.32) is not defined for $\Delta y_{i1} = (y_{i1} - y_{i0})$ because $\Delta y_{i0}$ and $\Delta \mathbf{x}_{i0}$ are missing. However, by continuous substitution, we can write $\Delta y_{i1}$ as

$$\Delta y_{i1} = a_{i1} + \sum_{j=0}^{\infty} \gamma^j \Delta u_{i,1-j}, \tag{4.5.1}$$

where $a_{i1} = \boldsymbol{\beta}' \sum_{j=0}^{\infty} \Delta \mathbf{x}_{i,1-j} \gamma^j$. Since $\Delta \mathbf{x}_{i,1-j}$, $j = 1, 2, \ldots,$ are unavailable, $a_{i1}$ is unknown. Treating $a_{i1}$ as a free parameter to be estimated will again introduce the incidental parameters problem. To get around this problem, the expected value of $a_{i1}$, conditional on the observables, has to be a function of a finite number of parameters of the form,

$$E(a_{i1} \mid \Delta \mathbf{x}_i) = c^* + \boldsymbol{\pi}' \Delta \mathbf{x}_i, i = 1, \ldots, N, \tag{4.5.2}$$

where $\boldsymbol{\pi}$ is a $TK_1 \times 1$ vector of constants, and $\Delta \mathbf{x}_i$ is a $TK_1 \times 1$ vector of $(\Delta \mathbf{x}_{i1}', \ldots, \Delta \mathbf{x}_{iT}')'$. Hsiao, Pesaran, and Tahmiscioglu (2002) have shown that if $\mathbf{x}_{it}$ are generated by

$$\mathbf{x}_{it} = \boldsymbol{\mu}_i + \mathbf{g}t + \sum_{j=0}^{\infty} \mathbf{b}_j' \boldsymbol{\xi}_{i,t-j} \sum_{j=0}^{\infty} \mid b_j \mid < \infty, \tag{4.5.3}$$

where $\boldsymbol{\xi}_{it}$ are assumed to be i.i.d. with mean 0 and constant covariance matrix, then (4.5.2) holds. The data-generating process of the exogenous variables

$\mathbf{x}_{it}$ (4.5.3) can allow fixed and different intercepts $\boldsymbol{\mu}_i$ across $i$, or to have $\boldsymbol{\mu}_i$ randomly distributed with a common mean. However, if there exists a trend term in the data-generating process of $\mathbf{x}_{it}$, then they must be identical across $i$.

Given (4.5.2), $\Delta y_{i1}$ can be written as

$$\Delta y_{i1} = c^* + \boldsymbol{\pi}' \Delta \mathbf{x}_i + v_{i1}^*. \tag{4.5.4}$$

where $v_{i1}^* = \sum_{j=0}^{\infty} \gamma^j \Delta u_{i,1-j} + [a_{i1} - E(a_{i1} \mid \Delta \mathbf{x}_i)]$. By construction, $E(v_{i1}^* \mid \Delta \mathbf{x}_i) = 0$, $E(v_{i1}^{*2}) = \sigma_{v^*}^2$, $E(v_{i1}^* \Delta u_{i2}) = -\sigma_u^2$, and $E(v_{i1}^* \Delta u_{it}) = 0$, for $t = 3, 4, \ldots, T$. It follows that the covariance matrix of $\Delta \mathbf{u}_i^* = (v_{i1}^*, \Delta \mathbf{u}_i')'$ has the form

$$\Omega^* = \sigma_u^2 \begin{bmatrix} h & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & & \\ 0 & \ddots & \ddots & & \\ \vdots & \ddots & \ddots & & \\ 0 & & & -1 & 2 \end{bmatrix} = \sigma_u^2 \tilde{\Omega}^*, \tag{4.5.5}$$

where $h = \frac{\sigma_{v^*}^2}{\sigma_u^2}$.

Combining (4.3.32) and (4.5.4), we can write the likelihood function of $\Delta \mathbf{y}_i^* = (\Delta y_{i1}, \ldots, \Delta y_{iT})'$, $i = 1, \ldots, N$, in the form of

$$(2\pi)^{-\frac{NT}{2}} \mid \Omega^* \mid^{-\frac{N}{2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{N} \Delta \mathbf{u}_i^{*\prime} \Omega^{*-1} \Delta \mathbf{u}_i^* \right\}, \tag{4.5.6}$$

if $\Delta \mathbf{u}_i^*$ is normally distributed, where

$$\begin{aligned} \Delta \mathbf{u}_i^* = [&\Delta y_{i1} - c^* - \boldsymbol{\pi}' \Delta \mathbf{x}_i, \Delta y_{i2} - \gamma \Delta y_{i1} \\ &- \boldsymbol{\beta}' \Delta \mathbf{x}_{i2}, \ldots, \Delta y_{iT} - \gamma \Delta y_{i,T-1} - \boldsymbol{\beta}' \Delta \mathbf{x}_{iT}]'. \end{aligned} \tag{4.5.7}$$

The likelihood function again depends only on a fixed number of parameters and satisfies the standard regularity conditions, so that the MLE is consistent and asymptotically normally distributed as $N \to \infty$.

Since $\mid \tilde{\Omega}^* \mid = 1 + T(h-1)$ and

$$\tilde{\Omega}^{*-1} = [1 + T(h-1)]^{-1} \cdot \begin{bmatrix} T & T-1 & \cdots & 2 & 1 \\ T-1 & (T-1)h & & 2h & h \\ \vdots & \vdots & & \vdots & \vdots \\ 2 & 2h & & 2[(T-2)h - (T-3)] & (T-2)h - (T-3) \\ 1 & h & & (T-2)h - (T-3) & (T-1)h - (T-2) \end{bmatrix}, \tag{4.5.8}$$

the logarithm of the likelihood function (4.5.6) is

$$\begin{aligned}\ln L = &-\frac{NT}{2}\log 2\pi - \frac{NT}{2}\log\sigma_u^2 - \frac{N}{2}\log\left[1+T(h-1)\right] \\ &-\frac{1}{2}\sum_{i=1}^{N}\left[(\Delta\mathbf{y}_i^* - H_i\boldsymbol{\psi})'\Omega^{*-1}(\Delta\mathbf{y}_i^* - H_i\boldsymbol{\psi})\right],\end{aligned} \tag{4.5.9}$$

where $\boldsymbol{\psi} = (c^*, \boldsymbol{\pi}', \gamma, \boldsymbol{\beta}')'$, and

$$H_i = \begin{bmatrix} 1 & \Delta\mathbf{x}_i' & 0 & \mathbf{0}' \\ 0 & \mathbf{0}' & \Delta y_{i1} & \Delta\mathbf{x}_{i2}' \\ \vdots & & \vdots & \vdots \\ 0 & \mathbf{0}' & \Delta y_{i,T-1} & \Delta\mathbf{x}_{iT}' \end{bmatrix}.$$

The MLE is obtained by solving the following equations simultaneously:

$$\boldsymbol{\psi} = \left(\sum_{i=1}^{N} H_i'\hat{\hat{\Omega}}^{*-1}H_i\right)^{-1}\left(\sum_{i=1}^{N} H_i'\hat{\hat{\Omega}}^{*-1}\Delta\mathbf{y}_i^*\right), \tag{4.5.10}$$

$$\sigma_u^2 = \frac{1}{NT}\sum_{i=1}^{N}\left[(\Delta\mathbf{y}_i^* - H_i\hat{\boldsymbol{\psi}})'(\hat{\hat{\Omega}}^*)^{-1}(\Delta\mathbf{y}_i^* - H_i\hat{\boldsymbol{\psi}})\right], \tag{4.5.11}$$

$$h = \frac{T-1}{T} + \frac{1}{\hat{\sigma}_u^2 NT^2}\sum_{i=1}^{N}\left[(\Delta\mathbf{y}_i^* - H_i\hat{\boldsymbol{\psi}})'(\mathbf{J}\,\mathbf{J}')(\Delta\mathbf{y}_i^* - H_i\hat{\boldsymbol{\psi}})\right], \tag{4.5.12}$$

where $\mathbf{J}' = (T, T-1, \ldots, 2, 1)$. One way to obtain the MLE is to iterate among (4.5.10)–(4.5.12) conditionally on the early round estimates of the other parameters until the solution converges or to use the Newton–Raphson type iterative scheme (Hsiao, Pesaran, and Tahmiscioglu 2002).

For finite $N$, occasionally, the transformed MLE breaks down giving estimated $\gamma$ greater than unity or negative variance estimates. However, the problem quickly disappears as $N$ becomes large. For further discussions on the properties of transformed MLE when $\gamma = 1$ or approaches $-1$ or explosive, see Han and Phillips (2013) and Kruiniger (2009).

### 4.5.2 Minimum Distance Estimator

Conditional on $\Omega^*$, the MLE is the minimum distance estimator (MDE) of the form

$$\text{Min}\sum_{i=1}^{N}\Delta\mathbf{u}_i^{*\prime}\Omega^{*-1}\Delta\mathbf{u}_i^*. \tag{4.5.13}$$

In the case that $\Omega^*$ is unknown, a two-step feasible MDE can be implemented. In the first step we obtain consistent estimators of $\sigma_u^2$ and $\sigma_{v*}^2$. For instance, we can regress (4.5.4) across $i$ to obtain the least-squares residuals $\hat{v}_{i1}^*$, and then estimate

$$\hat{\sigma}_{v^*}^2 = \frac{1}{N - TK_1 - 1} \sum_{i=1}^{N} \hat{v}_{i1}^{*2}. \tag{4.5.14}$$

Similarly, we can apply the IV to (4.3.32) and obtain the estimated residuals $\Delta\hat{u}_{it}$ and

$$\hat{\sigma}_u^2 = \frac{1}{N(T-1)} \sum_{i=1}^{N} \Delta\hat{\mathbf{u}}_i' \tilde{A}^{-1} \Delta\hat{\mathbf{u}}_i \tag{4.5.15}$$

where $\tilde{A}$ is defined in (4.3.46).

In the second step, we substitute estimated $\sigma_u^2$ and $\sigma_{v^*}^2$ into (4.5.5) and treat them as if they were known and use (4.5.10) to obtain the MDE of $\boldsymbol{\psi}$, $\hat{\boldsymbol{\psi}}_{\text{MDE}}$.

The asymptotic covariance matrix of MDE, Var $(\hat{\boldsymbol{\psi}}_{\text{MDE}})$, using the true $\Omega^*$ as the weighting matrix is equal to $(\sum_{i=1}^{N} H_i' \Omega^{*-1} H_i)^{-1}$. The asymptotic covariance of the feasible MDE using a consistently estimated $\Omega^*$, Var $(\hat{\boldsymbol{\psi}}_{\text{FMDE}})$, contrary to the static case, is equal to (Hsiao, Pesaran, and Tahmiscioglu 2002)

$$\begin{aligned} &\left(\frac{1}{N}\sum_{i=1}^{N} H_i' \Omega^{*-1} H_i\right)^{-1} + \left(\frac{1}{N}\sum_{i=1}^{N} H_i' \Omega^{*-1} H_i\right)^{-1} \\ &\begin{bmatrix} 0 & \mathbf{0}' & 0 & \mathbf{0}' \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ 0 & \mathbf{0}' & d & \mathbf{0}' \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \\ &\left(\frac{1}{N}\sum_{i=1}^{N} H_i' \Omega^{*-1} H_i\right)^{-1}, \end{aligned} \tag{4.5.16}$$

where

$$\begin{aligned} d = &\frac{[\gamma^{T-2} + 2\gamma^{T-3} + \cdots + (T-1)]^2}{[1 + T(h-1)]^2 \sigma_u^4} \\ &\cdot \left(\sigma_u^4 \operatorname{Var}\left(\hat{\sigma}_{v^*}^2\right) + \sigma_{v^*}^4 \operatorname{Var}\left(\hat{\sigma}_u^2\right) - 2\sigma_u^2 \sigma_{v^*}^2 \operatorname{Cov}\left(\hat{\sigma}_{v^*}^2, \hat{\sigma}_u^2\right)\right). \end{aligned}$$

The second term of (4.5.16) arises because the estimation of $\boldsymbol{\psi}$ and $\Omega^*$ are not asymptotically independent when the lagged dependent variables also appear as regressors.

### 4.5.3 Relations between the Likelihood-Based Estimator and the GMM

Although normality is assumed to derive the transformed MLE and MDE, it is not required. Both estimators remain consistent and asymptotically normally distributed even the errors are not normally distributed. Under normality, the transformed MLE achieves the Cramér–Rao lower bound, and hence is fully efficient. Even without normality, the transformed MLE (or MDE if $\Omega^*$ is known) is more efficient than the GMM that only uses second moment restrictions.

Using the formula of partitioned inverse (e.g., Amemiya 1985), the covariance matrix of the minimum distance estimator of $(\gamma, \boldsymbol{\beta})$ is of the form

$$\operatorname{Cov}\begin{pmatrix}\boldsymbol{\gamma}_{MDE}\\ \boldsymbol{\beta}_{MDE}\end{pmatrix}=\sigma_u^2\left[\sum_{i=1}^{N}\begin{pmatrix}\Delta\mathbf{y}_{i,-1}'\\ \Delta X_i'\end{pmatrix}\left(\tilde{A}-\frac{1}{h}\mathbf{g}\mathbf{g}'\right)^{-1}(\Delta\mathbf{y}_{i,-1},\ \Delta X_i)\right]^{-1} \tag{4.5.17}$$

where $\mathbf{g}' = (-1, 0, \ldots, 0)$.

We note that (4.5.17) is smaller than

$$\sigma_u^2\left[\sum_{i=1}^{N}\begin{pmatrix}\Delta\mathbf{y}_{i,-1}'\\ \Delta X_i'\end{pmatrix}\tilde{A}^{-1}(\Delta\mathbf{y}_{i,-1},\ \Delta X_i)\right]^{-1}, \tag{4.5.18}$$

in the sense that the difference between the two matrices is a nonpositive semidefinite matrix, because $\tilde{A}-(\tilde{A}-\frac{1}{h}\mathbf{g}\mathbf{g}')$ is a positive semidefinite matrix. Furthermore,

$$\begin{aligned}&\sum_{i=1}^{N}\begin{pmatrix}\Delta\mathbf{y}_{i,-1}'\\ \Delta X_i'\end{pmatrix}\tilde{A}^{-1}(\Delta\mathbf{y}_{i,-1},\ \Delta X_i)-\left[\sum_{i=1}^{N}\begin{pmatrix}\Delta\mathbf{y}_{i,-1}'\\ \Delta X_i'\end{pmatrix}W_i'\right]\left(\sum_{i=1}^{N}W_i\tilde{A}W_i'\right)^{-1}\\ &\qquad\cdot\left[\sum_{i=1}^{N}W_i(\Delta y_{i,-1},\ \Delta X_i)\right]\\ &\quad=D'[I-Q(Q'Q)^{-1}Q]D,\end{aligned} \tag{4.5.19}$$

is a positive semidefinite matrix, where $D=(D_1',\ldots,D_N')'$, $Q=(Q_1', Q_2', \ldots, Q_N')'$, $D_i=\Lambda'(\Delta\mathbf{y}_{i,-1}, \Delta X_i)$, $Q_i=\Lambda^{-1}W_i$, and $\Lambda\Lambda'=\tilde{A}^{-1}$. Therefore, the asymptotic covariance matrix of the GMM estimator (4.3.47), (4.3.48), is greater than (4.5.18), which is greater than (4.5.17) in the sense that the difference of the two covariance matrix is a positive semidefinite matrix.

When $\tilde{\Omega}^*$ is unknown, the asymptotic covariance matrix of (4.3.47) remains as (4.3.48). But the asymptotic covariance matrix of the feasible MDE is (4.5.16). Although the first term of (4.5.16) is smaller than (4.3.47), it is not clear that with the addition of the second term, it will remain smaller than (4.3.48). However, it is very likely so because of several factors. First, additional

information due to the $\Delta y_{i1}$ equation is utilized that can be substantial (e.g., see Hahn 1999). Second, the GMM method uses the $(t-1)$ instruments $(y_{i0}, \ldots, y_{i,t-2})$ for the $\Delta y_{it}$ equation for $t = 2, 3, \ldots, T$. The likelihood-based approach uses the $t$ instruments $(y_{i0}, y_{i1}, \ldots, y_{i,t-1})$. Third, the likelihood approach uses the condition that $E(H_i' \Omega^{*-1} \Delta \mathbf{u}_i^*) = \mathbf{0}$ and the GMM method uses the condition $E(\frac{1}{N} \sum_{i=1}^{N} W_i \Delta \mathbf{u}_i) = \mathbf{0}$. The grouping of observations in general will lead to a loss of information.[18]

Although both the GMM and the likelihood-based estimator are consistent, the process of removing the individual-specific effects in a dynamic model creates the order 1, (O(1)) correlation between $(y_{i,t-1} - y_{i,t-1})$ and $(u_{it} - u_{i,t-1})$. The likelihood approach uses all $NT$ observations to approximate the population moment $E(H_i' \Omega^{*-1} \Delta \mathbf{u}_i^*) = \mathbf{0}$, and hence is asymptotically unbiased independent of the way $N$ or $T \longrightarrow \infty$ (Hsiao and Zhang 2013). The GMM (or instrumental variable) approach transforms the correlation between $(y_{it} - y_{i,t-1})$ and $(u_{it} - u_{i,t-1})$ into the correlation between $\frac{1}{N} \sum_{i=1}^{N} \mathbf{q}_{it}(y_{it} - y_{i,t-1})$ and $\frac{1}{N} \sum_{i=1}^{N} q_{it}(u_{it} - u_{i,t-1})$, which is of order $\frac{1}{N}$, $O(\frac{1}{N})$. Therefore, when $T$ is fixed and $N$ is large, the GMM estimator is consistent and $\sqrt{N}(\hat{\gamma}_{\text{GMM}} - \gamma)$ is centered at 0. However, the number of moment conditions for the GMM (say (4.3.40)) is (or increases) at the order of $T^2$. This could create finite sample bias (e.g., see Ziliak 1997). When both $N$ and $T$ are large, and $\frac{T}{N} \longrightarrow c$, $0 < c < \infty$ as $N \to \infty$, the effects of the correlations due to $\frac{1}{N} \sum_{i=1}^{N} \mathbf{q}_{it}(y_{it} - y_{i,t-1})$ and $\frac{1}{N} \sum_{i=1}^{N} \mathbf{q}_{it}(u_{it} - u_{i,t-1})$ get magnified. Alvarez and Arellano (2003) show that $\sqrt{NT}\hat{\gamma}_{\text{GMM}}$ has asymptotic bias equal to $-\sqrt{c}(1+\gamma)$. On the other hand, the likelihood-based estimator is asymptotically unbiased (Hsiao and Zhang 2013). In other words, the GMM estimator multiplied by the scale factor $\sqrt{NT}$ is not centered at $\sqrt{NT}\gamma$, but the likelihood based estimator is.[19] Whether an estimator is asymptotically biased or not has important implications in statistical inference because in hypothesis testing typically we normalize the estimated $\gamma$ by the inverse of its standard error, which is equivalent to multiplying the estimator by the scale factor $\sqrt{NT}$. The Monte Carlo studies conducted by Hsiao and Zhang (2013) show that there is no size distortion for the MLE or simple IV ((4.3.33), (4.3.34)) but there are significant size distortions for GMM when $N$ and $T$ are of similar magnitude. For a nominal 5% significance level test, the actual size could be 40% when $\gamma = .5$ and 80% when $\gamma = .8$ for cases when $N$ and $T$ are of similar magnitude.

[18] For additional discussions on the contribution of initial observations, see Blundell and Bond (1998) and Hahn (1999).

[19] As a matter of fact, Alvarez and Arellano (2003) show that the least variance ratio estimator (which they call "the limited information maximum likelihood estimator") has asymptotic bias of order $\frac{1}{2N-T}$ when $0 < c < 2$. However, it appears that their forward deviation approach works only under fixed initial conditions. When the initial condition is treated as random, there is no asymptotic bias for the MLE (see Hsiao and Zhang 2013).

Table 4.4. *Monte Carlo design*

| Design number | $\gamma$ | $\beta$ | $\phi$ | $\theta$ | $g$ | $R^2_{\Delta y}$ | $\sigma_\epsilon$ |
|---|---|---|---|---|---|---|---|
| 1 | 0.4 | 0.6 | 0.5 | 0.5 | 0.01 | 0.2 | 0.800 |
| 2 | 0.4 | 0.6 | 0.9 | 0.5 | 0.01 | 0.2 | 0.731 |
| 3 | 0.4 | 0.6 | 1 | 0.5 | 0.01 | 0.2 | 0.711 |
| 4 | 0.4 | 0.6 | 0.5 | 0.5 | 0.01 | 0.4 | 1.307 |
| 5 | 0.4 | 0.6 | 0.9 | 0.5 | 0.01 | 0.4 | 1.194 |
| 6 | 0.4 | 0.6 | 1 | 0.5 | 0.01 | 0.4 | 1.161 |
| 7 | 0.8 | 0.2 | 0.5 | 0.5 | 0.01 | 0.2 | 1.875 |
| 8 | 0.8 | 0.2 | 0.9 | 0.5 | 0.01 | 0.2 | 1.302 |
| 9 | 0.8 | 0.2 | 1 | 0.5 | 0.01 | 0.2 | 1.104 |
| 10 | 0.8 | 0.2 | 0.5 | 0.5 | 0.01 | 0.4 | 3.062 |
| 11 | 0.8 | 0.2 | 0.9 | 0.5 | 0.01 | 0.4 | 2.127 |
| 12 | 0.8 | 0.2 | 1 | 0.5 | 0.01 | 0.4 | 1.803 |

*Source:* Hsiao, Pesaran, and Tahmiscioglu (2002, Table 1).

Hsiao, Pesaran, and Tahmiscioglu (2002) have conducted Monte Carlo studies to compare the performance of the IV of (4.3.34), the GMM of (4.3.47), the MLE, and the MDE when $T$ is small and $N$ is finite. They generate $y_{it}$ by

$$y_{it} = \alpha_i + \gamma y_{i,t-1} + \beta x_{it} + u_{it}, \tag{4.5.20}$$

where the error term $u_{it}$ is generated from two schemes. One is from $N(0, \sigma_u^2)$. The other is from mean adjusted $\chi^2$ with 2 degrees of freedom. The regressor $x_{it}$ is generated according to

$$x_{it} = \mu_i + gt + \xi_{it} \tag{4.5.21}$$

where $\xi_{it}$ follows an autoregressive moving average process

$$\xi_{it} - \phi\xi_{i,t-1} = \epsilon_{it} + \theta\epsilon_{i,t-1} \tag{4.5.22}$$

and $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$. The fixed effects $\mu_i$ and $\alpha_i$ are generated from a variety of schemes such as being correlated with $x_{it}$ or uncorrelated with $x_{it}$ but from a mixture of different distributions. Table 4.4 gives a summary of the different designs of the Monte Carlo study.

In generating $y_{it}$ and $x_{it}$, both are assumed to start from 0. But the first 50 observations are discarded. The bias and root mean square error (RMSE) of various estimators of $\gamma$ and $\beta$ when $T = 5$ and $N = 50$ based on 2500 replications are reported in Tables 4.5 and 4.6, respectively. The results show that the bias of the MLE of $\gamma$ as a percentage of the true value is smaller than 1 percent in most cases. The bias of the IV of $\gamma$ can be significant for certain data generating processes. In particular, if $\gamma$ is close to 1, the GMM method could run into weak IV problem (for an analytical results, see Kruiniger 2009). The MDE and GMM of $\gamma$ also have substantial downward biases in all designs. The bias of the GMM estimator of $\gamma$ can be as large as 15 to 20 percent in many

Table 4.5. *Bias of estimators ($T = 5$ and $N = 50$)*

| | | Bias | | | |
|---|---|---|---|---|---|
| Design | Coeff. | IVE | MDE | MLE | GMM |
| 1 | $\gamma = 0.4$ | 0.0076201 | −0.050757 | −0.000617 | −0.069804 |
| | $\beta = 0.6$ | −0.001426 | 0.0120812 | 0.0023605 | 0.0161645 |
| 2 | $\gamma = 0.4$ | 0.0220038 | −0.052165 | −0.004063 | −0.072216 |
| | $\beta = 0.6$ | −0.007492 | 0.0232612 | 0.0027946 | 0.0321212 |
| 3 | $\gamma = 0.4$ | 1.3986691 | −0.054404 | −0.003206 | −0.075655 |
| | $\beta = 0.6$ | −0.386998 | 0.0257393 | 0.0002997 | 0.0365942 |
| 4 | $\gamma = 0.4$ | 0.0040637 | −0.026051 | −0.001936 | −0.03616 |
| | $\beta = 0.6$ | 0.0004229 | 0.0066165 | 0.0019218 | 0.0087369 |
| 5 | $\gamma = 0.4$ | 0.1253257 | −0.023365 | −0.000211 | −0.033046 |
| | $\beta = 0.6$ | −0.031759 | 0.0113724 | 0.0016388 | 0.0155831 |
| 6 | $\gamma = 0.4$ | −0.310397 | −0.028377 | −0.00351 | −0.040491 |
| | $\beta = 0.6$ | 0.0640605 | 0.0146638 | 0.0022274 | 0.0209054 |
| 7 | $\gamma = 0.8$ | −0.629171 | −0.108539 | 0.009826 | −0.130115 |
| | $\beta = 0.2$ | −0.018477 | 0.0007923 | 0.0026593 | 0.0007962 |
| 8 | $\gamma = 0.8$ | −1.724137 | −0.101727 | 0.0027668 | −0.128013 |
| | $\beta = 0.2$ | 0.0612431 | 0.0109865 | −0.000011 | 0.013986 |
| 9 | $\gamma = 0.8$ | −0.755159 | −0.102658 | 0.00624 | −0.133843 |
| | $\beta = 0.2$ | −0.160613 | 0.0220208 | 0.0002624 | 0.0284606 |
| 10 | $\gamma = 0.8$ | 0.1550445 | −0.045889 | 0.001683 | −0.05537 |
| | $\beta = 0.2$ | 0.0096871 | 0.0000148 | 0.0007889 | −0.000041 |
| 11 | $\gamma = 0.8$ | −0.141257 | −0.038216 | −0.000313 | −0.050427 |
| | $\beta = 0.2$ | 0.0207338 | 0.0048828 | 0.0007621 | 0.0063229 |
| 12 | $\gamma = 0.8$ | 0.5458734 | −0.039023 | 0.0005702 | −0.053747 |
| | $\beta = 0.2$ | −0.069023 | 0.0079627 | 0.0003263 | 0.010902 |

*Source:* Hsiao, Pesaran, and Tahmiscioglu (2002, Table 2).

cases and is larger than the bias of the MDE. The MLE also has the smallest RMSE followed by the MDE, then GMM. The IV has the largest RMSE.

### 4.5.4 Issues of Random versus Fixed-Effects Specification

The GMM or the MLE of the transformed likelihood function (4.5.6) or the MDE (4.5.10) is consistent and asymptotically normally distributed whether $\alpha_i$ are fixed or random. However, when $\alpha_i$ are random and uncorrelated with $\mathbf{x}_{it}$, the likelihood function of the form (4.3.19) uses the level variables whereas (4.5.6) uses the first difference variables. In general, the variation across individuals are greater than the variation within individuals. Moreover, first differencing reduces the number of time series observations by one per cross-sectional unit; hence maximizing (4.5.6) yields estimators that will not be as efficient as

Table 4.6. *Root mean square error ($T = 5$ and $N = 50$)*

| | | Root Mean Square Error | | | |
|---|---|---|---|---|---|
| Design | Coeff. | IVE | MDE | MLE | GMM |
| 1 | $\gamma = 0.4$ | 0.1861035 | 0.086524 | 0.0768626 | 0.1124465 |
| | $\beta = 0.6$ | 0.1032755 | 0.0784007 | 0.0778179 | 0.0800119 |
| 2 | $\gamma = 0.4$ | 0.5386099 | 0.0877669 | 0.0767981 | 0.11512 |
| | $\beta = 0.6$ | 0.1514231 | 0.0855346 | 0.0838699 | 0.091124 |
| 3 | $\gamma = 0.4$ | 51.487282 | 0.0889483 | 0.0787108 | 0.1177141 |
| | $\beta = 0.6$ | 15.089928 | 0.0867431 | 0.0848715 | 0.0946891 |
| 4 | $\gamma = 0.4$ | 0.1611908 | 0.0607957 | 0.0572515 | 0.0726422 |
| | $\beta = 0.6$ | 0.0633505 | 0.0490314 | 0.0489283 | 0.0497323 |
| 5 | $\gamma = 0.4$ | 2.3226456 | 0.0597076 | 0.0574316 | 0.0711803 |
| | $\beta = 0.6$ | 0.6097378 | 0.0529131 | 0.0523433 | 0.0556706 |
| 6 | $\gamma = 0.4$ | 14.473198 | 0.0620045 | 0.0571656 | 0.0767767 |
| | $\beta = 0.6$ | 2.9170627 | 0.0562023 | 0.0550687 | 0.0607588 |
| 7 | $\gamma = 0.8$ | 27.299614 | 0.1327602 | 0.116387 | 0.1654403 |
| | $\beta = 0.2$ | 1.2424372 | 0.0331008 | 0.0340688 | 0.0332449 |
| 8 | $\gamma = 0.8$ | 65.526156 | 0.1254994 | 0.1041461 | 0.1631983 |
| | $\beta = 0.2$ | 3.2974597 | 0.043206 | 0.0435698 | 0.0450143 |
| 9 | $\gamma = 0.8$ | 89.83669 | 0.1271169 | 0.104646 | 0.1706031 |
| | $\beta = 0.2$ | 5.2252014 | 0.0535363 | 0.0523473 | 0.0582538 |
| 10 | $\gamma = 0.8$ | 12.201019 | 0.074464 | 0.0715665 | 0.0884389 |
| | $\beta = 0.2$ | 0.6729934 | 0.0203195 | 0.020523 | 0.0203621 |
| 11 | $\gamma = 0.8$ | 17.408874 | 0.0661821 | 0.0642971 | 0.0822454 |
| | $\beta = 0.2$ | 1.2541247 | 0.0268981 | 0.026975 | 0.02756742 |
| 12 | $\gamma = 0.8$ | 26.439613 | 0.0674678 | 0.0645253 | 0.0852814 |
| | $\beta = 0.2$ | 2.8278901 | 0.0323355 | 0.0323402 | 0.0338716 |

*Source:* Hsiao, Pesaran, and Tahmiscioglu (2002, Table 5).

the MLE of (4.3.19) when $\alpha_i$ are indeed random. However, if $\alpha_i$ are fixed or correlated with $\mathbf{x}_{it}$, the MLE of (4.3.19) yields an inconsistent estimator.

The transformed MLE or MDE is consistent under a more general data-generating process of $\mathbf{x}_{it}$ than the MLE of (4.3.19) or the GLS (4.3.30). For the Bhargava and Sargan (1983) MLE of the random effects model to be consistent, we will have to assume that the $\mathbf{x}_{it}$ are strictly exogenous and are generated from the same stationary process with common means ((4.3.21)). Otherwise, $E(y_{i0} \mid \mathbf{x}_i) = \mathbf{c}_i + \boldsymbol{\pi}_i' \mathbf{x}_i$, where $\mathbf{c}_i$ and $\boldsymbol{\pi}_i$ vary across $i$, and we will have the incidental parameters problem again. On the other hand, the transformed likelihood approach allows $\mathbf{x}_{it}$ to be correlated with individual specific effects, $\alpha_i$, and to have different means (or intercepts) (4.5.3). Therefore it appears that if one is not sure about the assumption of the effects, $\alpha_i$, or the homogeneity assumption about the data-generating process of $\mathbf{x}_{it}$, one should work with the

transformed likelihood function (4.5.6) or the MDE (4.5.10) despite the fact that one may lose efficiency under the ideal condition.

The use of the transformed likelihood approach also offers the possibility of using a Hausman (1978) type test for fixed versus random effects specification or test for the homogeneity and stationarity assumption about the $\mathbf{x}_{it}$ process under the assumption that $\alpha_i$ are random. Under the null of random effects and homogeneity of the $\mathbf{x}_{it}$ process, the MLE of the form (4.3.19) is asymptotically efficient. The transformed MLE of (4.5.6) is consistent, but not efficient. On the other hand, if $\alpha_i$ are fixed or $\mathbf{x}_{it}$ is not generated by a homogeneous process but satisfies (4.5.3), the transformed MLE of (4.5.6) is consistent, but the MLE of (4.3.19) is inconsistent. Therefore, a Hausman type test statistics (3.5.2) can be constructed by comparing the difference between the two estimators.

## 4.6 ESTIMATION OF DYNAMIC MODELS WITH ARBITRARY SERIAL CORRELATIONS IN THE RESIDUALS

In previous sections we discussed estimation of the dynamic model

$$y_{it} = \gamma y_{i,t-1} + \boldsymbol{\beta}'\mathbf{x}_{it} + \alpha_i^* + u_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{matrix} \tag{4.6.1}$$

under the assumption that $u_{it}$ are serially uncorrelated, where we now again let $\mathbf{x}_{it}$ stand for a $K \times 1$ vector of time-varying exogenous variables. When $T$ is fixed and $N$ tends to infinity, we can relax the restrictions on the serial correlation structure of $u_{it}$ and still obtain efficient estimates of $\gamma$ and $\boldsymbol{\beta}$.

Taking the first difference of (4.6.1) to eliminate the individual effect $\alpha_i^*$, and stacking all equations for a single individual, we have a system of $(T-1)$ equations,

$$\begin{aligned}
y_{i2} - y_{i1} &= \gamma(y_{i1} - y_{i0}) + \boldsymbol{\beta}'(\mathbf{x}_{i2} - \mathbf{x}_{i1}) + (u_{i2} - u_{i1}), \\
y_{i3} - y_{i2} &= \gamma(y_{i2} - y_{i1}) + \boldsymbol{\beta}'(\mathbf{x}_{i3} - \mathbf{x}_{i2}) + (u_{i3} - u_{i2}), \\
&\vdots \\
y_{iT} - y_{i,T-1} &= \gamma(y_{i,T-1} - y_{i,T-2}) + \boldsymbol{\beta}'(\mathbf{x}_{iT} - \mathbf{x}_{i,T-1}) \\
&\quad + (u_{iT} - u_{i,T-1}), \quad i = 1, \ldots, N,
\end{aligned} \tag{4.6.2}$$

We complete the system (4.6.2) with the identities

$$\begin{aligned}
y_{i0} &= E^*(y_{i0} \mid \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}) + [y_{i0} - E^*(y_{i0} \mid \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})] \\
&= a_0 + \sum_{t=1}^{T} \boldsymbol{\pi}_{0t}'\mathbf{x}_{it} + \epsilon_{i0}
\end{aligned} \tag{4.6.3}$$

and

$$\begin{aligned} y_{i1} &= E^*(y_{i1} \mid \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}) + [y_{i1} - E^*(y_{i1} \mid \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})] \\ &= a_1 + \sum_{t=1}^{T} \boldsymbol{\pi}_{1t}' \mathbf{x}_{it} + \epsilon_{i1}, \qquad i = 1, \ldots, N. \end{aligned} \tag{4.6.4}$$

where $E^*$ denotes the projection operator. Because (4.6.3) and (4.6.4) are exactly identified equations, we can ignore them and apply the three-stage least squares (3SLS) or generalized 3SLS (see Chapter 5) to the system (4.6.2) only. With regard to the cross equation constraints in (4.6.2), one can either directly substitute them out or first obtain unknown nonzero coefficients of each equation ignoring the cross equation linear constraints, then impose the constraints and use the constrained estimation formula [Theil 1971, p. 281, equation (8.5)].

Because the system (4.6.2) does not involve the individual effects, $\alpha_i^*$, nor does the estimation method rely on specific restrictions on the serial-correlation structure of $u_{it}$, the method is applicable whether $\alpha_i^*$ are treated as fixed or random or as being correlated with $\mathbf{x}_{it}$. However, to implement simultaneous-equations estimation methods to (4.6.2) without imposing restrictions on the serial-correlation structure of $u_{it}$, there must exist strictly exogenous variables $\mathbf{x}_{it}$ such that

$$E(u_{it} \mid \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}) = 0. \tag{4.6.5}$$

Otherwise, the coefficient $\gamma$ and the serial correlations of $u_{it}$ cannot be disentangled (e.g., Binder, Hsiao, and Pesaran 2005).

## 4.7 MODELS WITH BOTH INDIVIDUAL- AND TIME-SPECIFIC ADDITIVE EFFECTS

For notational ease and without loss of generality, we illustrate the fundamental issues of dynamic model with both individual- and time-specific additive effects model by restricting $\boldsymbol{\beta} = \mathbf{0}$ in (4.1.2); thus the model becomes

$$y_{it} = \gamma y_{i,t-1} + v_{it}, \tag{4.7.1}$$

$$\begin{aligned} v_{it} = \alpha_i + \lambda_t + u_{it}, \qquad & i = 1, \ldots, N, \\ & t = 1, \ldots, T, \\ & y_{i0} \text{ observable.} \end{aligned} \tag{4.7.2}$$

The panel data estimators discussed in Sections 4.3–4.6 assume no presence of $\lambda_t$ (i.e., $\lambda_t = 0 \ \forall \ t$). When $\lambda_t$ are indeed present, those estimators are not consistent if $T$ is finite when $N \to \infty$. For instance, the consistency of GMM (4.3.47) is based on the assumption that $\frac{1}{N} \sum_{i=1}^{N} y_{i,t-j} \Delta v_{it}$ converges to the population moments (4.3.40) of 0. However, if $\lambda_t$ are also present as in (4.7.2),

this condition is likely to be violated. To see this, taking the first difference of (4.7.1) yields

$$\begin{aligned} \Delta y_{it} &= \gamma \Delta y_{i,t-1} + \Delta v_{it} \\ &= \gamma \Delta y_{i,t-1} + \Delta \lambda_t + \Delta u_{it}, \\ i &= 1, \ldots, N, \\ t &= 2, \ldots, T. \end{aligned} \tag{4.7.3}$$

Although under the assumption $\lambda_t$ are independently distributed over $t$ with mean 0,

$$E(y_{i,t-j} \Delta v_{it}) = 0 \quad \text{for } j = 2, \ldots, t, \tag{4.7.4}$$

the sample moment, as $N \longrightarrow \infty$,

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^{N} y_{i,t-j} \Delta v_{it} = \frac{1}{N} \sum_{i=1}^{N} y_{i,t-j} \Delta \lambda_t \\ + \frac{1}{N} \sum_{i=1}^{N} y_{i,t-j} \Delta u_{it} \end{aligned} \tag{4.7.5}$$

converges to $\bar{y}_{t-j} \Delta \lambda_t$, which in general is not equal to 0, in particular, if $y_{it}$ has mean different from 0,[20] where $\bar{y}_t = \frac{1}{N} \sum_{i=1}^{N} y_{it}$.

To obtain consistent estimators of $\gamma$, we need to take explicit account of the presence of $\lambda_t$ in addition to $\alpha_i$. When $\alpha_i$ and $\lambda_t$ are fixed constants, under the assumption that $u_{it}$ is independent normal and fixed $y_{i0}$, the MLE of the FE model (4.7.1) is equal to:

$$\tilde{\gamma}_{cv} = \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} y^*_{i,t-1} y^*_{it}}{\sum_{i=1}^{N} \sum_{t=1}^{T} y^{*2}_{i,t-1}}, \tag{4.7.6}$$

where $y^*_{it} = (y_{it} - \bar{y}_i - \bar{y}_t + \bar{y})$; $\bar{y}_i = \frac{1}{T} \sum_{t=1}^{T} y_{it}$, $\bar{y}_t = \frac{1}{N} \sum_{i=1}^{N} y_{it}$; $\bar{y} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} y_{it}$; and similarly for $\bar{y}_t$, $\bar{y}_{i,-1}$, $\bar{\mathbf{x}}_i$, $\bar{\mathbf{x}}_t$, $\mathbf{x}^*_{it}$, $v^*_{it}$, $\bar{v}_i$, $\bar{v}_t$, and $\bar{v}$. The FE MLE of $\gamma$ is also called the covariance estimator because it is equivalent to first applying covariance transformation to sweep out $\alpha_i$ and $\lambda_t$,

$$y^*_{it} = \gamma y^*_{i,t-1} + v^*_{it}, \tag{4.7.7}$$

and then applying the least-squares estimator of (4.7.7).

The probability limit of $\tilde{\gamma}_{cv}$ is identical to the case where $\lambda_t \equiv 0$ for all $t$ (4.2.8) (Hahn and Moon 2006; Hsiao and Tahmiscioglu 2008). The bias

[20] For instance, if $y_{it}$ is also a function of exogenous variables as (4.1.2), where $\bar{y}_t = \frac{1}{N} \sum_{i=1}^{N} y_{it}$.

is to the order of $(1/T)$ and it is identical independent of whether $\alpha_i$ and $\lambda_t$ are fixed or random and are identical whether $\lambda_t$ are present or not (e.g., Hahn and Moon 2006; Hsiao and Tahmiscioglu 2008). When $T \longrightarrow \infty$, the MLE of the FE model is consistent. However, if $N$ also goes to infinity and $\lim \left(\frac{N}{T}\right) = c > 0$, Hahn and Moon (2006) have shown that $\sqrt{NT}(\tilde{\gamma}_{cv} - \gamma)$ is asymptotically normally distributed with mean $-\sqrt{c}(1+\gamma)$ and variance $1-\gamma^2$. In other words, the usual $t$-statistic based on $\gamma_{cv}$ could be subject to severe size distortion unless $T$ increases faster than $N$.

If $\alpha_i$ and $\lambda_t$ are random and satisfy (4.3.8), because $Ey_{i0}v_{it} \neq 0$, we either have to write (4.7.1) conditional on $y_{i0}$ or to complete the system (4.7.1) by deriving the marginal distribution of $y_{i0}$. By continuous substitutions, we have

$$\begin{aligned} y_{i0} &= \frac{1-\gamma^m}{1-\gamma}\alpha_i + \sum_{j=0}^{m-1}\lambda_{i,-j}\gamma^j + \sum_{j=0}^{m-1}\epsilon_{i,-j}\gamma^j \\ &= v_{i0}, \end{aligned} \tag{4.7.8}$$

assuming the process started at period $-m$.

Under (4.3.8), $Ey_{i0} = Ev_{i0} = 0$, $\text{var}(y_{i0}) = \sigma_0^2$, $E(v_{i0}v_{it}) = \frac{1-\gamma^m}{1-\gamma}\sigma_\alpha^2 = c^*$, $Ev_{it}v_{jt} = d^*$. Stacking the $T+1$ time series observations for the $i$th individual into a vector, $\mathbf{y}_i = (y_{i0}, \ldots, y_{i,T})'$ and $\mathbf{y}_{i,-1} = (0, y_{i1}, \ldots, y_{i,T-1})'$, $\mathbf{v}_i = (v_{i0}, \ldots, v_{iT})'$. Let $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_n')'$, $\mathbf{y}_{-1} = (\mathbf{y}_{1,-1}', \ldots, \mathbf{y}_{N,-1}')$, $\mathbf{v} = (\mathbf{v}_1', \ldots, \mathbf{v}_N')'$, then

$$\mathbf{y} = \mathbf{y}_{-1}\gamma + \mathbf{v}, \tag{4.7.9}$$

$$\begin{aligned} E\mathbf{v} &= \mathbf{0}, \\ E\mathbf{v}\mathbf{v}' &= \sigma_u^2 I_N \otimes \begin{pmatrix} \omega & \mathbf{0}' \\ \mathbf{0} & I_T \end{pmatrix} + \sigma_\alpha^2 I_N \otimes \begin{pmatrix} 0 & c^*\mathbf{e}_T' \\ c^*\mathbf{e}_T & \mathbf{e}_T\mathbf{e}_T' \end{pmatrix} \\ &\quad + \sigma_\lambda^2 \mathbf{e}_N\mathbf{e}_N \otimes \begin{pmatrix} d^* & \mathbf{0}' \\ \mathbf{0} & I_T \end{pmatrix}, \end{aligned} \tag{4.7.10}$$

where $\otimes$ denotes the Kronecker product, and $\omega$ denotes the variance of $v_{i0}$ divided by $\sigma_u^2$. The system (4.7.9) has a fixed number of unknowns $(\gamma, \sigma_u^2, \sigma_\alpha^2, \sigma_\lambda^2, \sigma_0^2, c^*, d^*)$ as $N$ and $T$ increase. Therefore, the MLE (or quasi-MLE or GLS of (4.7.9)) is consistent and asymptotically normally distributed.

When $\alpha_i$ and $\lambda_t$ are fixed constants, we note that first differencing only eliminates $\alpha_i$ from the specification. The time-specific effects, $\Delta\lambda_t$, remain at (4.7.3). To further eliminate $\Delta\lambda_t$, we note that the cross-sectional mean $\Delta y_t = \frac{1}{N}\sum_{i=1}^N \Delta y_{it}$ is equal to

$$\Delta y_t = \gamma \Delta y_{t-1} + \Delta\lambda_t + \Delta u_t, \tag{4.7.11}$$

where $\Delta u_t = \frac{1}{N}\sum_{i=1}^{N} \Delta u_{it}$. Taking the deviation of (4.7.3) from (4.7.11) yields

$$\Delta y_{it}^* = \gamma \Delta y_{i,t-1}^* + \Delta u_{it}^*, \qquad \begin{matrix} i = 1, \ldots, N, \\ t = 2, \ldots, T, \end{matrix} \tag{4.7.12}$$

where $\Delta y_{it}^* = (\Delta y_{it} - \Delta y_t)$ and $\Delta u_{it}^* = (\Delta u_{it} - \Delta u_t)$. The system (4.7.12) no longer involves $\alpha_i$ and $\lambda_t$.

Since

$$E[y_{i,t-j}\Delta u_{it}^*] = 0 \quad \text{for} \quad \begin{matrix} j = 2, \ldots, t, \\ t = 2, \ldots, T. \end{matrix} \tag{4.7.13}$$

the $\frac{1}{2}T(T-1)$ orthogonality conditions can be represented as

$$E(W_i \Delta \tilde{\mathbf{u}}_i^*) = \mathbf{0}, \tag{4.7.14}$$

where $\Delta \tilde{\mathbf{u}}_i^* = (\Delta u_{i2}^*, \ldots, \Delta u_{iT}^*)'$,

$$W_i = \begin{pmatrix} \mathbf{q}_{i2} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{q}_{i3} & & \\ . & . & \ddots & \\ \vdots & \vdots & & \\ \mathbf{0} & \mathbf{0} & & \mathbf{q}_{iT} \end{pmatrix}, \quad i = 1, \ldots, N,$$

and $\mathbf{q}_{it} = (y_{i0}, y_{i1}, \ldots, y_{i,t-2})'$, $t = 2, 3, \ldots, T$. Following Arellano and Bond (1991), we can propose a generalized method of moments (GMM) estimator,[21]

$$\begin{aligned} \tilde{\gamma}_{\text{GMM}} = &\left\{ \left[ \frac{1}{N}\sum_{i=1}^{N} \Delta \tilde{\mathbf{y}}_{i,-1}^{*'} W_i' \right] \hat{\Psi}^{-1} \left[ \frac{1}{N}\sum_{i=1}^{N} W_i \Delta \tilde{\mathbf{y}}_{i,-1}^* \right] \right\}^{-1} \\ &\cdot \left\{ \left[ \frac{1}{N}\sum_{i=1}^{N} \Delta \tilde{\mathbf{y}}_{i,-1}^{*'} \mathbf{W}_i' \right] \hat{\Psi}^{-1} \left[ \frac{1}{N}\sum_{i=1}^{N} W_i \Delta \tilde{\mathbf{y}}_i^* \right] \right\}, \end{aligned} \tag{4.7.15}$$

where $\Delta \tilde{\mathbf{y}}_i^* = (\Delta y_{i2}^*, \ldots, \Delta y_{iT}^*)'$, $\Delta \tilde{\mathbf{y}}_{i,-1}^* = (\Delta y_{i1}^*, \ldots, \Delta y_{i,T-1}^*)$, and

$$\hat{\Psi} = \frac{1}{N^2}\left[ \sum_{i=1}^{N} W_i \Delta \hat{\tilde{\mathbf{u}}}_i^* \right] \left[ \sum_{i=1}^{N} W_i \Delta \hat{\tilde{\mathbf{u}}}_i^* \right]' \tag{4.7.16}$$

and $\Delta \hat{\tilde{\mathbf{u}}}_i^* = \Delta \tilde{\mathbf{y}}_i^* - \Delta \tilde{\mathbf{y}}_{i,-1}^* \tilde{\gamma}$, and $\tilde{\gamma}$ denotes some initial consistent estimator of $\gamma$, say a simple instrumental variable estimator.

[21] For ease of exposition, we have considered only the GMM that makes use of orthogonality conditions. For additional moments conditions such as homoscedasticity or initial observations see, for example, Ahn and Schmidt (1995), Blundell and Bond (1998).

The asymptotic covariance matrix of $\tilde{\gamma}_{\text{GMM}}$ can be approximated by

$$\text{asym. Cov}\,(\tilde{\gamma}_{\text{GMM}}) = \left\{ \left[ \sum_{i=1}^{N} \Delta \tilde{\mathbf{y}}_{i,-1}^{*'} W_i' \right] \hat{\Psi}^{-1} \left[ \sum_{i=1}^{N} W_i \Delta \tilde{\mathbf{y}}_{i,-1}^{*} \right] \right\}^{-1}. \quad (4.7.17)$$

To implement the likelihood approach, we need to complete the system (4.7.12) by deriving the marginal distribution of $\Delta y_{i1}^*$ through continuous substitution,

$$\begin{aligned} \Delta y_{i1}^* &= \sum_{j=0}^{m-1} \Delta u_{i,1-j}^* \gamma^j \\ &= \Delta u_{i1}^*, \qquad i = 1, \ldots, N. \end{aligned} \quad (4.7.18)$$

Let $\Delta \mathbf{y}_i^* = (\Delta y_{i1}^*, \ldots, \Delta y_{iT}^*)'$, $\Delta \mathbf{y}_{i,-1}^* = (0, \Delta y_{i1}^*, \ldots, \Delta y_{i,T-1}^*)'$, $\Delta \mathbf{u}_i^* = (\Delta u_{i1}^*, \ldots, \Delta u_{iT}^*)'$; then the system

$$\Delta \mathbf{y}_i^* = \Delta \mathbf{y}_{i,-1}^* \gamma + \Delta u_i^*, \quad (4.7.19)$$

does not involve $\alpha_i$ and $\lambda_t$. The MLE conditional on $\omega = \frac{\text{Var}\,(\Delta y_{i1}^*)}{\sigma_u^2}$ is identical to the GLS:

$$\hat{\gamma}_{\text{GLS}} = \left[ \sum_{i=1}^{N} \Delta \mathbf{y}_{i,-1}^{*'} \tilde{A}^{*-1} \Delta \mathbf{y}_{i,-1}^{*} \right]^{-1} \left[ \sum_{i=1}^{N} \Delta \mathbf{y}_{i,-1}^{*'} \tilde{A}^{*-1} \Delta \mathbf{y}_{i}^{*} \right], \quad (4.7.20)$$

where $\tilde{A}^*$ is a $T \times T$ matrix of the form,

$$\tilde{A}^* = \begin{bmatrix} \omega & -1 & 0 & 0 & \ldots & 0 & 0 \\ -1 & 2 & -1 & 0 & \ldots & . & . \\ 0 & -1 & 2 & -1 & \ldots & . & . \\ . & . & . & . & . & 2 & -1 \\ 0 & . & . & . & & -1 & 2 \end{bmatrix}. \quad (4.7.21)$$

The GLS is consistent and asymptotically normally distributed with covariance matrix equal to

$$\text{Var}\,(\hat{\gamma}_{\text{GLS}}) = \sigma_u^2 \left[ \sum_{i=1}^{N} \Delta \mathbf{y}_{i,-1}^{*'} \tilde{A}^{*-1} \Delta \mathbf{y}_{i,-1}^{*} \right]^{-1}. \quad (4.7.22)$$

**Remark 4.7.1:** The GLS with $\Delta\boldsymbol{\lambda}$ present is basically of the same form as the GLS without the time-specific effects (i.e., $\Delta\boldsymbol{\lambda} = \mathbf{0}$) (Hsiao, Pesaran, and Tahmiscioglu 2002), (4.5.10). However,there is an important difference between the two. The estimator (4.7.20) uses $\Delta y_{i,t-1}^*$ as the regressor for the equation $\Delta y_{it}^*$ (4.7.19), does not use $\Delta y_{i,t-1}$ as the regressor for the equation $\Delta y_{it}$ ((4.7.3)). If there are indeed common shocks that affect all the cross-sectional units, then the estimator (4.5.10) is inconsistent while (4.7.20) is consistent (for details, see Hsiao and Tahmiscioglu 2008). Note also that even though when

there are no time-specific effects, (4.7.20) remains consistent, although it will not be as efficient as (4.5.10).

**Remark 4.7.2:** The estimator (4.7.20) and the estimator (4.7.15) remain consistent and asymptotically normally distributed when the effects are random because the transformation (4.7.11) effectively removes the individual- and time-specific effects from the specification. However, if the effects are indeed random, and uncorrelated with $\mathbf{x}_{it}$ then the MLE or GLS of (4.7.7) is more efficient.

**Remark 4.7.3:** The GLS (4.7.20) assumes known $\omega$. If $\omega$ is unknown, one may substitute it by a consistent estimator $\hat{\omega}$, and then apply the feasible GLS. However, there is an important difference between the GLS and the feasible GLS in a dynamic setting. The feasible GLS is not asymptotically equivalent to the GLS when $T$ is finite. However, if both $N$ and $T \to \infty$ and $\lim(\frac{N}{T}) = c > 0$, then the FGLS will be asymptotically equivalent to the GLS (Hsiao and Tahmiscioglu 2008).

**Remark 4.7.4:** The MLE or GLS of (4.7.20) can also be derived by treating $\Delta\lambda_t$ as fixed parameters in the system (4.7.3). Through continuous substitution, we have

$$\Delta y_{i1} = \lambda_1^* + \Delta\tilde{u}_{i1}, \tag{4.7.23}$$

where $\lambda_1^* = \sum_{j=0}^{m} \gamma^j \Delta\lambda_{1-j}$ and $\Delta\tilde{u}_{i1} = \sum_{j=0}^{m} \gamma^j \Delta u_{i,1-j}$. Let $\Delta\mathbf{y}_i' = (\Delta y_{i1}, \ldots, \Delta y_{iT})$, $\Delta\mathbf{y}_{i,-1}' = (0, \Delta y_{i1}, \ldots, \Delta y_{i,T-1})$, $\Delta\mathbf{u}_i' = (\Delta\tilde{u}_{i1}, \ldots, \Delta u_{iT})$, and $\Delta\boldsymbol{\lambda}' = (\lambda_1^*, \Delta\lambda_2, \ldots, \Delta\lambda_T)$, we may write

$$\underset{NT \times 1}{\Delta\mathbf{y}} = \begin{pmatrix} \Delta\mathbf{y}_1 \\ \vdots \\ \Delta\mathbf{y}_N \end{pmatrix} = \begin{pmatrix} \Delta\mathbf{y}_{1,-1} \\ \vdots \\ \Delta\mathbf{y}_{N,-1} \end{pmatrix}\gamma + (\mathbf{e}_N \otimes I_T)\Delta\boldsymbol{\lambda} + \begin{pmatrix} \Delta\mathbf{u}_1 \\ \vdots \\ \Delta\mathbf{u}_N \end{pmatrix}$$

$$= \Delta\mathbf{y}_{-1}\gamma + (\mathbf{e}_N \otimes I_T)\Delta\boldsymbol{\lambda} + \Delta\mathbf{u}, \tag{4.7.24}$$

If $u_{it}$ is i.i.d. normal with mean 0 and variance $\sigma_u^2$, then $\Delta\mathbf{u}_i'$ is independently normally distributed across $i$ with mean $\mathbf{0}$ and covariance matrix $\sigma_u^2\tilde{A}^*$, and $\omega = \frac{\text{Var}(\Delta\tilde{u}_{i1})}{\sigma_u^2}$.

The log-likelihood function of $\Delta\mathbf{y}$ takes the form

$$\log L = -\frac{NT}{2}\log\sigma_u^2 - \frac{N}{2}\log|\tilde{A}^*| - \frac{1}{2\sigma_u^2}[\Delta\mathbf{y} - \Delta\mathbf{y}_{-1}\gamma - (\mathbf{e}_N \otimes I_T)\Delta\boldsymbol{\lambda}]'$$

$$\cdot (I_N \otimes \tilde{A}^{*-1})[\Delta\mathbf{y} - \Delta\mathbf{y}_{-1}\gamma - (\mathbf{e}_N \otimes I_T)\Delta\boldsymbol{\lambda}]. \tag{4.7.25}$$

Taking the partial derivative of (4.7.25) with respect to $\Delta\boldsymbol{\lambda}$ and solving for $\Delta\boldsymbol{\lambda}$ yields

$$\Delta\hat{\boldsymbol{\lambda}} = (N^{-1}\mathbf{e}_N' \otimes I_T)(\Delta\mathbf{y} - \Delta\mathbf{y}_{-1}\gamma). \tag{4.7.26}$$

Substituting (4.7.26) into (4.7.25) yields the concentrated log-likelihood function.

$$\log L_c = -\frac{NT}{2}\log\sigma_\epsilon^2 - \frac{N}{2}\log\mid \tilde{A}^*\mid - \frac{1}{2\sigma_\epsilon^2}(\Delta\mathbf{y}^* - \Delta\mathbf{y}^*_{-1}\gamma)'(I_N\otimes\tilde{A}^{*-1})(\Delta\mathbf{y}^* - \Delta y^*_{-1}\gamma). \tag{4.7.27}$$

Maximizing (4.7.27) conditional on $\omega$ yields (4.7.20).

**Remark 4.7.5:** When $\gamma$ approaches 1 and $\sigma_\alpha^2$ is large relative to $\sigma_u^2$, the GMM estimator of the form (4.3.47) suffers from the weak instrumental variables issues and performs poorly (e.g., Binder, Hsiao, and Pesaran 2005). On the other hand, the performance of the likelihood or GLS estimator is not affected by these problems.

**Remark 4.7.6:** Hahn and Moon (2006) propose a bias-corrected estimator as

$$\tilde{\gamma}_b = \tilde{\gamma}_{cv} + \frac{1}{T}(1+\tilde{\gamma}_{cv}). \tag{4.7.28}$$

They show that when $N/T \to c$, as both $N$ and $T$ tend to infinity where $0 < c < \infty$,

$$\sqrt{NT}(\tilde{\gamma}_b - \gamma) \Longrightarrow N(0, 1-\gamma^2). \tag{4.7.29}$$

The limited Monte Carlo studies conducted by Hsiao and Tahmiscioglu (2008) to investigate the finite sample properties of the feasible GLS (FGLS), GMM, and bias-corrected (BC) estimator of Hahn and Moon (2006) have shown that in terms of bias and RMSEs, FGLS dominates. However, the BC rapidly improves as $T$ increases. In terms of the closeness of actual size to the nominal size, again FGLS dominates and rapidly approaches the nominal size when $N$ or $T$ increases. The GMM with $T$ fixed and $N$ large also has actual sizes close to nominal sizes except for the cases when $\gamma$ is close to unity (here $\gamma = 0.8$). However, if $N$ and $T$ are of similar magnitude, $\frac{T}{N} = c \neq 0$, there are significant size distortion (e.g., Hsiao and Zhang 2013). The BC has significant size distortion, presumably because of the use of asymptotic covariance matrix, which is significantly downward biased in the finite sample.

**Remark 4.7.7:** Hsiao and Tahmiscioglu (2008) also compared the FGLS and GMM with and without the correction of time-specific effects in the presence of both individual- and time-specific effects or in the presence of individual-specific effects only. It is interesting to note that when both individual- and time-specific effects are present, the biases and RMSEs are large for estimators assuming no time-specific effects; however, their biases decrease as $T$ increases when the time-specific effects are independent of regressors. On the other hand, even in the case of no time-specific effects, there is hardly any efficiency loss for the FGLS or GMM that makes the correction of presumed presence of time-specific effects. Therefore, if an investigator is not sure if the assumption

of cross-sectional independence is valid or not, it might be advisable to use estimators that take account both individual- and time-specific effects when $T$ is finite.

## APPENDIX 4A: DERIVATION OF THE ASYMPTOTIC COVARIANCE MATRIX OF FEASIBLE MDE

The estimation error of $\hat{\boldsymbol{\psi}}_{\text{MDE}}$ is equal to

$$\sqrt{N}(\hat{\boldsymbol{\psi}}_{\text{MDE}} - \boldsymbol{\psi}) = \left(\frac{1}{N}\sum_{i=1}^{N} H_i'\hat{\tilde{\Omega}}^{*-1}H_i\right)^{-1}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} H_i'\hat{\tilde{\Omega}}^{*-1}\Delta\mathbf{u}_i^*\right). \quad (4A.1)$$

When $N \longrightarrow \infty$

$$\frac{1}{N}\sum_{i=1}^{N} H_i'\hat{\tilde{\Omega}}^{*-1}H_i \longrightarrow \frac{1}{N}\sum_{i=1}^{N} H_i'\tilde{\Omega}^{*-1}H_i \quad (4A.2)$$

but

$$\begin{aligned}\frac{1}{\sqrt{N}}\sum_{i=1}^{N} H_i'\hat{\tilde{\Omega}}^{*-1}\Delta\mathbf{u}_i^* \simeq \frac{1}{\sqrt{N}}\sum_{i=1}^{N} H_i'\tilde{\Omega}^{*-1}\Delta\mathbf{u}_i^* \\ + \left[\frac{1}{N}\sum_{i=1}^{N} H_i'\left(\frac{\partial}{\partial h}\tilde{\Omega}^{*-1}\right)\Delta\mathbf{u}_i^*\right]\cdot\sqrt{N}(\hat{h}-h),\end{aligned} \quad (4A.3)$$

where the right-hand side follows from taking a Taylor series expansion of $\hat{\tilde{\Omega}}^{*-1}$ around $\tilde{\Omega}^{*-1}$. By (4.5.8),

$$\begin{aligned}\frac{\partial}{\partial h}\tilde{\Omega}^{*-1} = \frac{-T}{[1+T(h-1)]^2}\tilde{\Omega}^{*-1} \\ + \frac{1}{1+T(h-1)}\begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ 0 & T-1 & \dots & 2 & 1 \\ \ddots & \ddots & & \ddots & \ddots \\ . & 2 & \dots & 2(T-2) & T-2 \\ 0 & 1 & \dots & T-2 & T-1 \end{bmatrix}.\end{aligned} \quad (4A.4)$$

We have

$$\frac{1}{N}\sum_{i=1}^{N} H_i'\tilde{\Omega}^{*-1}\Delta\mathbf{u}_i^* \longrightarrow \mathbf{0},$$

$$\frac{1}{N}\sum_{i=1}^{N}\begin{bmatrix} 1 & \Delta\mathbf{x}_i' & \mathbf{0}' \\ \mathbf{0} & \mathbf{0} & \Delta X_i \end{bmatrix}'\cdot\frac{\partial}{\partial h}\tilde{\Omega}^{*-1}\Delta\mathbf{u}_i^* \longrightarrow \mathbf{0},$$

$$\frac{1}{N}\sum_{i=1}^{N}\Delta\mathbf{y}'_{i,-1}\begin{bmatrix} T-1 & \cdots & 1 \\ & \ddots & \vdots \\ 2 & & T-2 \\ 1 & & T-1 \end{bmatrix}\Delta\mathbf{u}_i^*$$
$$\longrightarrow [\gamma^{T-2}+2\gamma^{T-3}+\cdots+(T-1)]\sigma_u^2.$$

Since plim $\hat{\sigma}_u^2=\sigma_u^2$, and

$$\sqrt{N}(\hat{h}-h)=\sqrt{N}\left[\frac{\hat{\sigma}_{v*}^2}{\hat{\sigma}_u^2}-\frac{\sigma_{v*}^2}{\sigma_u^2}\right]=\sqrt{N}\frac{\sigma_u^2(\hat{\sigma}_{v*}^2-\sigma_{v*}^2)-\sigma_{v*}^2(\hat{\sigma}_u^2-\sigma_u^2)}{\hat{\sigma}_u^2\sigma_u^2},$$

it follows that the limiting distribution of the feasible MDE converges to

$$\sqrt{N}(\hat{\boldsymbol{\psi}}_{\text{MDE}}-\boldsymbol{\psi})\longrightarrow\left(\frac{1}{N}\sum_{i=1}^{N}H_i'\Omega^{*-1}H_i\right)^{-1}\left\{\frac{1}{\sqrt{N}}\sum_{i=1}^{N}H_i'\Omega^{*-1}\Delta\mathbf{u}_i^*\right.$$
$$-\begin{bmatrix}0\\ \mathbf{0}\\ 1\\ \mathbf{0}\end{bmatrix}\frac{[\gamma^{T-2}+2\gamma^{T-3}+\cdots+(T-1)]}{[1+T(h-1)]\sigma_u^2} \tag{4A.5}$$
$$\left.\left[\sigma_u^2\cdot\sqrt{N}\left(\hat{\sigma}_{v*}^2-\sigma_{v*}^2\right)-\sigma_{v*}^2\cdot\sqrt{N}\left(\hat{\sigma}_u^2-\sigma_u^2\right)\right]\right\},$$

with the asymptotic covariance matrix equal to (4.5.16).

## APPENDIX 4B: LARGE $N$ AND $T$ ASYMPTOTICS

In cases when $N$ is fixed and $T$ is large or $T$ is fixed and $N$ is large, standard one-dimensional asymptotic techniques can be applied. However, in some panel data sets, the orders of magnitude of the cross section and time series are similar, for instance, the Penn-World tables. These large $N$, large $T$ panels call for the use of large $N$, $T$ asymptotics rather than just large $N$ asymptotics. Moreover, when $T$ is large, there is a need to consider serial correlations more generally, including both short memory and persistent components. In some panel data sets such as the Penn-World Table, the time series components also have strongly evident nonstationarity. It turns out that panel data in this case can sometimes offer additional insights to the data-generating process than a single time series or cross-sectional data.

In regressions with large $N$, large $T$ panels most of the interesting test statistics and estimators inevitably depend on the treatment of the two indexes, $N$ and $T$, which tend to infinity together. Several approaches are possible:

(a) *Sequential Limits*. A sequential approach is to fix one index, say $N$, and allow the other, say $T$, to pass to infinity, giving an intermediate limit. Then, by letting $N$ pass to infinity subsequently, a sequential limit theory is obtained.

(b) *Diagonal Path Limits*. This approach allows the two indexes, $N$ and $T$, to pass to infinity along a specific diagonal path in the two-dimensional array, say $T = T(N)$ such as $\frac{N}{T} \longrightarrow c \neq 0 < \infty$ as the index $N \to \infty$. This approach simplifies the asymptotic theory of a double-indexed process into a single-indexed process.

(c) *Joint Limits*. A joint limit theory allows both indexes, $N$ and $T$, to pass to infinity simultaneously without placing specific diagonal path restrictions on the divergence, although it may still be necessary to exercise some control over the rate of expansion of the two indexes to get definitive results.

A double-index process in this monograph typically takes the form,

$$X_{N,T} = \frac{1}{k_N} \sum_{i=1}^{N} Y_{i,T}, \tag{4B.1}$$

where $k_N$ is an $N$-indexed standardizing factor, $Y_{i,T}$ are independent $m$-component random vectors across $i$ for all $T$ that is integrable and has the form

$$Y_{i,T} = \frac{1}{d_T} \sum_{t=1}^{T} f(Z_{i,t}), \tag{4B.2}$$

for the $h$-component independently, identically distributed random vectors, $Z_{i,t}$ with finite 4th moments; $f(\cdot)$ is a continuous functional from $R^h$ to $R^m$, and $d_T$ is a $T$-indexed standardizing factor. Sequential limit theory is easy to derive and generally leads to quick results. However, it can also give asymptotic results that are misleading in cases where both indexes pass to infinity simultaneously. A joint limit will give a more robust result than either a sequential limit or diagonal path limit, but will also be substantially more difficult to derive and will usually apply only under stronger conditions, such as the existence of higher moments, which will allow for uniformity in the convergence arguments. Phillips and Moon (1999) give the conditions for sequential convergence to imply joint convergence as:

(i) $X_{N,T}$ converges to $X_N$, for all $N$, in probability as $T \longrightarrow \infty$ uniformly and $X_N$ converges to $X$ in probability as $N \longrightarrow \infty$. Then $X_{N,T}$ converges to $X$ in probability jointly if and only if

$$\lim_{T\to\infty} \sup_{N} P\{\| X_{N,T} - X_N \| > \epsilon\} = 0 \text{ for every } \epsilon > 0. \tag{4B.3}$$

(ii) $X_{N,T}$ converges to $X_N$ in distribution for any fixed $N$ as $T \longrightarrow \infty$ and $X_N$ converges to $X$ in distribution as $N \longrightarrow \infty$. Then, $X_{N,T}$ converges to distribution jointly if and only if

$$\limsup_{N,\,T} \mid E(f(X_{N,T}) - E(f(X)) \mid = 0, \tag{4B.4}$$

for all bounded, continuous, real function on $R^m$.

Suppose $Y_{i,T}$ converges to $Y_i$ in distribution as $T \longrightarrow \infty$, Phillips and Moon (1999) have given the following set of sufficient conditions that ensures the sequential limits are equivalent to joint limits:

(i) $\lim\sup_{N,T} \left(\frac{1}{N}\right) \sum_{i=1}^{N} E \parallel Y_{i,T} \parallel < \infty$;
(ii) $\lim\sup_{N,T} \left(\frac{1}{N}\right) \sum_{i=1}^{N} \parallel EY_{i,T} - EY_i \parallel = 0$;
(iii) $\lim\sup_{N,T} \left(\frac{1}{N}\right) \sum_{i=1}^{N} E \parallel Y_{i,T} \parallel 1\left\{\parallel Y_{i,T} \parallel > N\epsilon\right\} = 0 \; \forall \epsilon > 0$;
(iv) $\lim\sup_{N} \left(\frac{1}{N}\right) \sum_{i=1}^{N} E \parallel Y_i \parallel 1\{\parallel Y_i \parallel > N\epsilon\} = 0 \; \forall \epsilon > 0$,

where $\parallel A \parallel$ is the Euclidean norm $(tr(A'A))^{\frac{1}{2}}$ and $1\{\cdot\}$ is an indicator function.

In general, if an estimator is of the form (4B.1) and $y_{i,T}$ is integrable for all $T$ and if this estimator is consistent in the fixed $T$, large $N$ case, it will remain consistent if both $N$ and $T$ tend to infinity irrespective of how $N$ and $T$ tend to infinity. Moreover, even in the case that an estimator is inconsistent for fixed $T$ and large $N$ case, say, the CV estimator for the fixed effects dynamic model (4.2.1), it can become consistent if $T$ also tends to infinity. The probability limit of an estimator, in general, is identical independent of the sequence of limits one takes. However, the properly scaled limiting distribution may be different depending on how the two indexes, $N$ and $T$, tend to infinity. Consider the double sequence

$$X_{N,T} = \frac{1}{N} \sum_{i=1}^{N} Y_{i,T}. \tag{4B.5}$$

Suppose $Y_{i,T}$ is independently, identically distributed across $i$ for each $T$ with $E(Y_{i,T}) = \frac{1}{\sqrt{T}} b$ and Var $(Y_{i,T}) \leq B < \infty$. For fixed $N$, $X_{N,T}$ converges to $X_N$ in probability as $T \to \infty$ where $E(X_N) = 0$. Because Var $(Y_{i,T})$ is bounded, by a law of large numbers, $X_N$ converges to 0 in probability as $N \to \infty$. Since (4B.5) satisfies (4B.3), the sequential limit is equal to the joint limit as $N, T \to \infty$. This can be clearly seen by writing

$$\begin{aligned} X_{N,T} &= \frac{1}{N} \sum_{i=1}^{N} [Y_{i,T} - E(Y_{i,T})] + \frac{1}{N} \sum_{i=1}^{N} E(Y_{i,T}) \\ &= \frac{1}{N} \sum_{i=1}^{N} [Y_{i,T} - E(Y_{i,T})] + \frac{b}{\sqrt{T}}. \end{aligned} \tag{4B.6}$$

Since the variance of $Y_{i,T}$ is uniformly bounded by $B$,

$$E(X_{N,T}^2) = \frac{1}{N} \text{ Var } (Y_{i,T}) + \frac{b^2}{T} \longrightarrow 0 \tag{4B.7}$$

as $N, T \longrightarrow \infty$. Equation (4B.7) implies that $X_{N,T}$ converges to 0 jointly as $N, T \longrightarrow \infty$.

Alternatively, if we let

$$X_{N,T} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} Y_{i,T}. \tag{4B.8}$$

The sequential limit would imply $X_{N,T}$ is asymptotically normally distributed with $E(X_{N,T}) = 0$. However, under (4B.8) the condition (4B.3) is violated. The joint limit would have

$$EX_{N,T} = \frac{\sqrt{c}}{N} \sum_{i=1}^{N} b \longrightarrow \sqrt{c}b, \tag{4B.9}$$

along some diagonal limit, $\frac{N}{T} \longrightarrow c \neq 0$ as $N \longrightarrow \infty$. In this case, $T$ has to increase faster than $N$ to make the $\sqrt{N}$-standardized sum of the biases small, say $\frac{N}{T} \longrightarrow 0$ to prevent the bias from having a dominating asymptotic effect on the standardized quantity (e.g., Alvarez and Arellano 2003; Hahn and Kuersteiner 2002).

If the time series component is an integrated process (nonstationary), panel regressions in which both $T$ and $N$ are large can behave very differently from time series regressions. For instance, consider the linear regression model

$$y = E(y \mid x) + v = \beta x + v. \tag{4B.10}$$

If $v_t$ is stationary (or $I(0)$ process), the least-squares estimator of $\beta$, $\hat{\beta}$, gives the same interpretation irrespective of whether $y$ and $x$ are stationary or integrated of order 1 $I(1)$ (i.e., the first difference of a variable is stationary or $I(0)$). However, if both $y_{it}$ and $x_{it}$ are $I(1)$ but not cointegrated, then $v_{it}$ is also $I(1)$. It is shown by Phillips (1986) that a time series regression coefficient $\hat{\beta}_i$ has a nondegenerating distribution as $T \longrightarrow \infty$. The estimate $\hat{\beta}_i$ is spurious in the sense that the time series regression of $y_{it}$ on $x_{it}$ does not identify any fixed long-run relation between $y_{it}$ and $x_{it}$. On the other hand, with panel data, such regressions are not spurious in the sense that they do, in fact, identify a long-run average relation between $y_{it}$ and $x_{it}$. To see this, consider the case that the $y$ and $x$ is bivariate normally distributed as $N(\mathbf{0}, \Sigma)$ with

$$\Sigma = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix}, \tag{4B.11}$$

then plim $\hat{\boldsymbol{\beta}} = \Sigma_{yx}\Sigma_{xx}^{-1}$. In a unit root framework of the form

$$\begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} y_{t-1} \\ x_{t-1} \end{pmatrix} + \begin{pmatrix} u_{yt} \\ u_{xt} \end{pmatrix}, \tag{4B.12}$$

where the errors $\mathbf{u}_t = (u_{yt}, u_{xt})'$ are stationary, then the panel regression under the assumption of cross-sectional independence yields

$$\text{plim } \hat{\beta} = \Omega_{yx}\Omega_{xx}^{-1}, \tag{4B.13}$$

which can be viewed as long-run average relation between $y$ and $x$, where $\Omega_{yx}$, $\Omega_{xx}$ denote the long-run covariance between $u_{yt}$ and $u_{xt}$, and the long-run variance of $x_t$ defined by

$$\begin{aligned}\Omega &= \lim_{T\to\infty} E\left[\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbf{u}_t\right)\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbf{u}_t'\right)\right] \\ &= \sum_{\ell=-\infty}^{\infty} E(\mathbf{u}_0\mathbf{u}_\ell') = \begin{pmatrix}\Omega_{yy} & \Omega_{yx}\\ \Omega_{xy} & \Omega_{xx}\end{pmatrix}.\end{aligned} \tag{4B.14}$$

When cross-sectional units have heterogeneous long-run covariance matrices $\Omega_i$ for $(y_{it}, x_{it})$, $i = 1\ldots., N$ with $E\Omega_i = \Omega$, Phillips and Moon (1999) extend this concept of a long-run average relation among cross-sectional units further

$$\beta = E(\Omega_{yx,i})(E\Omega_{xx,i})^{-1} = \Omega_{yx}\Omega_{xx}^{-1}. \tag{4B.15}$$

and show that the least-squares estimator converges to (4B.15) as $N, T \to \infty$.

This generalized concept of average relation between cross-sectional units covers both the cointegrated case (Engle and Granger 1987) in which $\beta$ is a cointegrating coefficient in the sense that the particular linear combination $y_t - \beta x_t$ is stationary, and the correlated but noncointegrated case, which is not available for a single time series. To see this point more clearly, suppose that the two nonstationary time series variables have the following relation:

$$\begin{aligned} y_t &= f_t + w_t, \\ x_t &= f_t, \end{aligned} \tag{4B.16}$$

with

$$\begin{pmatrix} w_t \\ f_t \end{pmatrix} = \begin{pmatrix} w_{t-1} \\ f_{t-1} \end{pmatrix} + \begin{pmatrix} u_{wt} \\ u_{ft} \end{pmatrix}, \tag{4B.17}$$

where $u_{ws}$ is independent of $u_{ft}$ for all $t$ and $s$ and has nonzero long-run variance. Then $f_t$ is a nonstationary common factor variable for $y$ and $x$ and $u_w$ is a nonstationary idiosyncratic factor variable. Since $w_t$ is nonstationary over time, it is apparent that there is no cointegrating relation between $y_t$ and $x_t$. However, since the two nonstationary variables $y_t$ and $x_t$ share a common contributory nonstationary source in $u_{ft}$, we may still expect to find evidence of a long-run correlation between $y_t$ and $x_t$, and this is what is measured by the regression coefficient $\beta$ in (4B.13).

Phillips and Moon (1999, 2000) show that for large $N$ and $T$ panels, the regression coefficient $\beta$ coverages to such a defined long-run average relation. However, if $N$ is fixed, then as $T \to \infty$, the least-squares estimator of $\beta$ is a nondegenerate random variable that is a functional of Brownian motion that does not converge to $\beta$ (Phillips 1986). In other words, with a single time series or a fixed number of time series, the regression coefficient $\beta$ will not converge to the long-run average relation defined by (4B.13) if only $T \to \infty$.

Therefore, if we define spurious regression as yielding nonzero $\beta$ for the two independent variables, then contrary to the case of time series regression of involving two linearly independent I(1) variables (Phillips 1986) the issue of spurious regression will not arise for the panel estimates of $N \to \infty$ (e.g., McCoskey and Kao 1998).

When data on cross-sectional dimension are correlated, the limit theorems become complicated. When there are strong correlations on cross-sectional dimensions, it is unlikely that the law of large numbers or central limit theory will hold if cross-sectional correlations are strong. They can hold only when cross-sectional dependence is weak (in the sense of time series mixing condition in the cross-sectional dimension, e.g., Conley 1999; Pesaran and Tosetti 2010).

CHAPTER 5

# Static Simultaneous-Equations Models

## 5.1 INTRODUCTION

In Chapters 3 and 4, we discussed the approach of decomposing the effect of a large number of factors that affect the dependent variables, but are not explicitly included as explanatory variables, into effects specific to individual units, to time periods, and to both individual units and time periods as a means to take account of the unobserved heterogeneity in panel data in estimating single-equation models. However, the consistency or asymptotic efficiency of various estimators discussed in previous chapters depends on the validity of the single-equation model assumptions. If they are not true, this approach may solve one problem but aggravate other problems.

For instance, consider the income-schooling model,

$$y = \beta_0 + \beta_1 S + \beta_2 A + u, \tag{5.1.1}$$

where $y$ is a measure of income, earnings, or wage rate, $S$ is a measure of schooling, and $A$ is an unmeasured ability variable that is assumed to be positively related to $S$. The coefficients $\beta_1$ and $\beta_2$ are assumed positive. Under the assumption that $S$ and $A$ are uncorrelated with $u$, the least-squares estimate of $\beta_1$ that ignores $A$ is biased upward. The standard left-out-variable formula gives the size of this bias as

$$E(\hat{\beta}_{1,\mathrm{LS}}) = \beta_1 + \beta_2 \frac{\sigma_{AS}}{\sigma_S^2}, \tag{5.1.2}$$

where $\sigma_S^2$ is the variance of $S$, and $\sigma_{AS}$ is the covariance between $A$ and $S$.

If the omitted variable $A$ is a purely "family" one,[1] that is, if siblings have exactly the same level of $A$, then estimating $\beta_1$ from within-family data (i.e., from differences between the brothers' earnings and differences between the brothers' education) will eliminate this bias. But if ability, apart from having a family component, also has an individual component, and this individual

[1] Namely, the family effect $A_i$ has the same meaning as $\alpha_i$ in Chapters 3 and 4.

component is not independent of the schooling variable, the within-family estimates are not necessarily less biased.

Suppose

$$A_{it} = \alpha_i + \omega_{it}, \tag{5.1.3}$$

where $i$ denotes the family, and $t$ denotes members of the family. If $\omega_{it}$ is uncorrelated with $S_{it}$, the combination of (5.1.1) and (5.1.3) is basically of the same form as (3.2.10). The expected value of the within (or LSDV) estimator is unbiased. On the other hand, if the within-family covariance between $A$ and $S$, $\sigma_{s\omega}$, is not equal to 0, the expected value of the within estimator is

$$E(\hat{\beta}_{1,w}) = \beta_1 + \beta_2 \frac{\sigma_{S\omega}}{\sigma^2_{S|w}}, \tag{5.1.4}$$

where $\sigma^2_{S|w}$ is the within-family variance of $S$. The estimator remains biased. Furthermore, if the reasons for the correlation between $A$ and $S$ are largely individual rather than familial, then going to within data will drastically reduce $\sigma^2_{S|w}$, with little change to $\sigma_{AS}$ (or $\sigma_{S\omega}$), which would make this source of bias even more serious.

Moreover, if $S$ is also a function of $A$ and other socioeconomic variables, (5.1.1) is only one behavioral equation in a simultaneous-equations model. Then the probability limit of the least-squares estimate, $\hat{\beta}_{1,\text{LS}}$, is no longer (5.1.2) but is of the form

$$\text{plim}\, \hat{\beta}_{1,\text{LS}} = \beta_1 + \beta_2 \frac{\sigma_{AS}}{\sigma^2_S} + \frac{\sigma_{uS}}{\sigma^2_S}, \tag{5.1.5}$$

where $\sigma_{uS}$ is the covariance between $u$ and $S$. If, as argued by Griliches (1977, 1979), schooling is the result, at least in part, of optimizing behavior by individuals and their family, $\sigma_{uS}$ could be negative. This opens the possibility that the least-squares estimates of the schooling coefficient may be biased downward rather than upward. Furthermore, if the reasons for $\sigma_{uS}$ being negative are again largely individual rather than familial, and the within-family covariance between $A$ and $S$ reduces $\sigma_{AS}$ by roughly the same proportion as $\sigma^2_{S|w}$ is to $\sigma^2_S$, there will be a significant decline in the $\hat{\beta}_{1,w}$ relative to $\hat{\beta}_{1,\text{LS}}$. The size of this decline will be attributed to the importance of ability and "family background," but in fact it reflects nothing more than the simultaneity problems associated with the schooling variable itself. In short, the simultaneity problem could reverse the single-equation conclusions.

In this chapter we focus on the issues of correlations arising from the joint dependence of $G$ endogenous variables $\mathbf{y}_{it} = (y_{1,it}, y_{2,it}, \ldots, y_{G,it})'$ given $K$ exogenous variables $\mathbf{x}_{it} = (x_{1,it}, x_{2,it}, \ldots, x_{K,it})'$. In this chapter, we focus on

issues of static simultaneous equations model of the form,[2]

$$\Gamma \mathbf{y}_{it} + \mathbf{B}\mathbf{x}_{it} + \boldsymbol{\mu} = \mathbf{v}_{it}, \qquad i = 1, \ldots, N, \quad t = 1, \ldots, T, \tag{5.1.6}$$

where $\Gamma$ and $\mathbf{B}$ are $G \times G$ and $G \times K$ matrices of coefficients; $\boldsymbol{\mu}$ is the $G \times 1$ vector of intercepts. We assume that the $G \times 1$ errors $\mathbf{v}_{it}$ has a component structure,

$$\mathbf{v}_{it} = \boldsymbol{\alpha}_i + \boldsymbol{\lambda}_t + \mathbf{u}_{it}, \tag{5.1.7}$$

where $\boldsymbol{\alpha}_i$ and $\boldsymbol{\lambda}_t$ denote the $G \times 1$ individual varying but time-invariant and individual-invariant but time-varying specific effects, respectively, and $\mathbf{u}_{it}$ denote the $G \times 1$ random vector that varies across $i$ and over $t$ and are uncorrelated with $\mathbf{x}_{it}$,

$$E\left(\mathbf{x}_{it}\mathbf{u}'_{js}\right) = \mathbf{0}. \tag{5.1.8}$$

The issue of dynamic dependence is discussed in Chapter 10, Section 10.4.

Model (5.1.6) could give rise to two sources of correlations between the regressors and the errors of the equations: (1) the potential correlations between the individual- and time-specific effects, $\boldsymbol{\alpha}_i$ and $\boldsymbol{\lambda}_t$, with $\mathbf{x}_{it}$, and (2), the correlations between the joint dependent variables and the errors. The first source of correlations could be eliminated through some linear transformation of the original variables. For instance, the covariance transformation of $\mathbf{y}_{it}$ and $\mathbf{x}_{it}$,

$$\dot{\mathbf{y}}_{it} = \mathbf{y}_{it} - \bar{\mathbf{y}}_i - \bar{\mathbf{y}}_t + \bar{\mathbf{y}}, \tag{5.1.9}$$

$$\dot{\mathbf{x}}_{it} = \mathbf{x}_{it} - \bar{\mathbf{x}}_i - \bar{\mathbf{x}}_t + \bar{\mathbf{x}}, \tag{5.1.10}$$

yields

$$\Gamma \dot{\mathbf{y}}_{it} + \mathbf{B}\dot{\mathbf{x}}_{it} = \dot{\mathbf{v}}_{it}, \tag{5.1.11}$$

where $\dot{\mathbf{v}}_{it} = \mathbf{v}_{it} - \bar{\mathbf{v}}_i - \bar{\mathbf{v}}_t + \bar{\mathbf{v}}$ and $(\bar{\mathbf{y}}_i, \bar{\mathbf{x}}_i, \bar{\mathbf{v}}_i)$, $(\bar{\mathbf{y}}_t, \bar{\mathbf{x}}_t, \bar{\mathbf{v}}_t)$, $(\bar{\mathbf{y}}, \bar{\mathbf{x}}, \bar{\mathbf{v}})$ denote the $i$th individual time series mean, cross-sectional mean at $t$, and overall mean of respective variable, for example, $\bar{\mathbf{y}}_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{y}_{it}$, $\bar{\mathbf{y}}_t = \frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_{it}$, and $\bar{\mathbf{y}} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \mathbf{y}_{it}$.

Under the assumption (5.1.8),

$$E\left(\dot{\mathbf{x}}_{it}\dot{\mathbf{v}}'_{js}\right) = \mathbf{0} \tag{5.1.12}$$

[2] The asymptotic property of a fixed-effects linear simultaneous-equations model is the same as that of the single-equation fixed-effects linear static model (see Chapter 3). The MLE of $\boldsymbol{\alpha}_i$ is consistent only when $T$ tends to infinity. The MLE of $\boldsymbol{\lambda}_t$ is consistent only when $N$ tends to infinity. However, just as in the linear static model, the MLE of $\Gamma$ and $B$ do not depend on the MLE of $\boldsymbol{\alpha}_i$ and $\boldsymbol{\lambda}_t$. They are consistent when either $N$ or $T$ or both tend to infinity (Schmidt 1984).

standard identification and estimation methods for a Cowles Commission structural equation model can be applied to model (5.1.11) to obtain consistent and asymptotically normally distributed estimators (e.g., Hood and Koopmans 1953; Hsiao 1983; Intriligator, Bodkin, and Hsiao 1996). However, exploitation of the component structure could lead to more efficient inference of (5.1.6) than those based on the two- or three-stage least-squares methods for model (5.1.11).

We assume $\boldsymbol{\alpha}_i$, $\boldsymbol{\lambda}_t$, and $\mathbf{u}_{it}$ are each $G \times 1$ random vectors that have 0 means and are independent of one another, and

$$
\begin{aligned}
&E\mathbf{x}_{it}\mathbf{v}'_{js} = \mathbf{0}, \\
&E\boldsymbol{\alpha}_i\boldsymbol{\alpha}'_j = \begin{cases} \Omega_\alpha = \left(\sigma^2_{\alpha g\ell}\right) & \text{if } i = j, \\ \mathbf{0} & \text{if } i \neq j, \end{cases} \\
&E\boldsymbol{\lambda}_t\boldsymbol{\lambda}'_s = \begin{cases} \Omega_\lambda = \left(\sigma^2_{\lambda g\ell}\right) & \text{if } t = s, \\ \mathbf{0} & \text{if } t \neq s, \end{cases} \\
&E\mathbf{u}_{it}\mathbf{u}'_{js} = \begin{cases} \Omega_u = \left(\sigma^2_{ug\ell}\right) & \text{if } i = j, \quad \text{and} \quad t = s, \\ \mathbf{0} & \text{otherwise.} \end{cases}
\end{aligned}
\tag{5.1.13}
$$

Multiplying (5.1.6) by $\Gamma^{-1}$, we have the reduced form

$$
\mathbf{y}_{it} = \boldsymbol{\mu}^* + \Pi\mathbf{x}_{it} + \boldsymbol{\epsilon}_{it}, \tag{5.1.14}
$$

where $\boldsymbol{\mu}^* = -\Gamma^{-1}\boldsymbol{\mu}$, $\Pi = -\Gamma^{-1}\mathbf{B}$, and $\boldsymbol{\epsilon}_{it} = \Gamma^{-1}\mathbf{v}_{it}$. The reduced-form error term $\boldsymbol{\epsilon}_{it}$ again has an error-component structure[3]

$$
\boldsymbol{\epsilon}_{it} = \boldsymbol{\alpha}^*_i + \boldsymbol{\lambda}^*_t + \mathbf{u}^*_{it}, \tag{5.1.15}
$$

with

$$
\begin{aligned}
&E\boldsymbol{\alpha}^*_i = E\boldsymbol{\lambda}^*_t = E\mathbf{u}^*_{it} = \mathbf{0}, \qquad E\boldsymbol{\alpha}^*_i\boldsymbol{\lambda}^{*\prime}_t = E\boldsymbol{\alpha}^*_i\mathbf{u}^{*\prime}_{it} = E\boldsymbol{\lambda}^*_t\mathbf{u}^{*\prime}_{it} = \mathbf{0}, \\
&E\boldsymbol{\alpha}^*_i\boldsymbol{\alpha}^{*\prime}_j = \begin{cases} \Omega^*_\alpha = \left(\sigma^{*2}_{\alpha g\ell}\right) & \text{if } i = j, \\ \mathbf{0} & \text{if } i \neq j, \end{cases} \\
&E\boldsymbol{\lambda}^*_t\boldsymbol{\lambda}^{*\prime}_s = \begin{cases} \Omega^*_\lambda = \left(\sigma^{*2}_{\lambda g\ell}\right) & \text{if } t = s, \\ \mathbf{0} & \text{if } t \neq s, \end{cases} \\
&E\mathbf{u}^*_{it}\mathbf{u}^{*\prime}_{js} = \begin{cases} \Omega^*_u = \left(\sigma^{*2}_{ug\ell}\right) & \text{if } i = j \quad \text{and} \quad t = s, \\ \mathbf{0} & \text{otherwise.} \end{cases}
\end{aligned}
\tag{5.1.16}
$$

If the $G \times G$ covariance matrices $\Omega_\alpha$, $\Omega_\lambda$, and $\Omega_u$ are unrestricted, there are no restrictions on the variance–covariance matrix. The usual order and rank

[3] Note that the meaning of these asterisks has been changed from what they were in previous chapters.

conditions are the necessary and sufficient conditions for identifying a particular equation in the system (e.g., Hsiao 1983). If there are restrictions on $\Omega_\alpha$, $\Omega_\lambda$, or $\Omega_u$, we can combine these covariance restrictions with the restrictions on the coefficient matrices to identify a model and obtain efficient estimates of the parameters. We first discuss estimation of the simultaneous-equations model under the assumption that there are no restrictions on the variance–covariance matrix, but the rank condition for identification holds. Estimation of reduced-form or stacked equations is discussed in Section 5.2, and estimation of the structural form is dealt with in Section 5.3. We then discuss the case in which there are restrictions on the variance–covariance matrix in Section 5.4. Because a widely used structure for longitudinal microdata is the triangular structure (e.g., Chamberlain 1976, 1977a,b; Chamberlain and Griliches 1975), we shall use this special case to illustrate how the covariance restrictions can be used to identify an otherwise unidentified model and to improve the efficiency of the estimates.

## 5.2 JOINT GENERALIZED LEAST-SQUARES ESTIMATION TECHNIQUE

We can write an equation of a reduced form (5.1.14) in the more general form in which the explanatory variables in each equation can be different[4]:

$$\mathbf{y}_g = \mathbf{e}_{NT}\boldsymbol{\mu}_g^* + \mathbf{X}_g\boldsymbol{\pi}_g + \boldsymbol{\epsilon}_g, \quad g = 1, \ldots, G, \tag{5.2.1}$$

where $\mathbf{y}_g$ and $\mathbf{e}_{NT}$ are $NT \times 1$, $\mathbf{X}_g$ is $NT \times K_g$, $\boldsymbol{\mu}_g^*$ is the $1 \times 1$ intercept term for the $g$th equation, $\boldsymbol{\pi}_g$ *is* $K_g \times 1$, and $\boldsymbol{\epsilon}_g = (I_N \otimes \mathbf{e}_T)\boldsymbol{\alpha}_g^* + (\mathbf{e}_N \otimes I_T)\boldsymbol{\lambda}_g^* + \mathbf{u}_g^*$, where $\boldsymbol{\alpha}_g^* = (\alpha_{1g}^*, \boldsymbol{\alpha}_{2g}^*, \ldots, \boldsymbol{\alpha}_{Ng}^*)'$, $\boldsymbol{\lambda}_g^* = (\lambda_{1g}^*, \lambda_{2g}^*, \ldots, \lambda_{Tg}^*)'$, and $\mathbf{u}_g^* = (u_{11g}^*, u_{12g}^*, \ldots, u_{1Tg}^*, u_{21g}^*, \ldots, u_{NTg}^*)'$ are $N \times 1$, $T \times 1$, and $NT \times 1$ random vectors, respectively. Stacking the set of $G$ equations, we get

$$\underset{GNT\times 1}{\mathbf{y}} = (I_G \otimes \mathbf{e}_{NT})\boldsymbol{\mu}^* + \mathbf{X}\boldsymbol{\pi} + \boldsymbol{\epsilon}, \tag{5.2.2}$$

where

$$\underset{GNT\times 1}{\mathbf{y}} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_G \end{bmatrix}, \qquad \underset{GNT\times\left(\sum_{g=1}^G K_g\right)}{\mathbf{X}} = \begin{bmatrix} X_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & X_2 & & \vdots \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & X_G \end{bmatrix},$$

$$\underset{G\times 1}{\boldsymbol{\mu}^*} = \begin{bmatrix} \boldsymbol{\mu}_1^* \\ \boldsymbol{\mu}_2^* \\ \vdots \\ \boldsymbol{\mu}_G^* \end{bmatrix}, \qquad \underset{\left(\sum_{g=1}^G K_g\right)\times 1}{\boldsymbol{\pi}} = \begin{bmatrix} \boldsymbol{\pi}_1 \\ \vdots \\ \boldsymbol{\pi}_G \end{bmatrix}, \qquad \boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_G \end{bmatrix},$$

[4] By allowing $X$ to be different, the discussion of estimation of reduced-form equations can proceed along the more general format of seemingly unrelated regression models (Avery 1977; Baltagi 1980).

with

$$V = E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = [V_{g\ell}], \tag{5.2.3}$$

where $V_{g\ell}$ denotes the $g\ell$th block submatrix of $V$, which is given by

$$\underset{NT\times NT}{V_{g\ell}} = E\left(\boldsymbol{\epsilon}_g\boldsymbol{\epsilon}_\ell'\right) = \sigma_{\alpha_{g\ell}}^{*2}A + \sigma_{\lambda_{g\ell}}^{*2}D + \sigma_{u_{g\ell}}^{*2}I_{NT}, \tag{5.2.4}$$

where $A = I_N \otimes \mathbf{e}_T\mathbf{e}_T'$ and $D = \mathbf{e}_N\mathbf{e}_N' \otimes I_T$. Equation (5.2.4) can also be written as

$$V_{g\ell} = \sigma_{1_{g\ell}}^{*2}\left(\frac{1}{T}A - \frac{1}{NT}J\right) + \sigma_{2_{g\ell}}^{*2}\left(\frac{1}{N}D - \frac{1}{NT}J\right) + \sigma_{u_{g\ell}}^{*2}\tilde{Q} + \sigma_{4_{g\ell}}^{*2}\left(\frac{1}{NT}J\right), \tag{5.2.5}$$

where $J = \mathbf{e}_{NT}\mathbf{e}_{NT}'$, $\tilde{Q} = I_{NT} - (1/T)A - (1/N)D + (1/NT)J$, $\sigma_{1_{g\ell}}^{*2} = \sigma_{u_{g\ell}}^{*2} + T\sigma_{\alpha_{g\ell}}^{*2}$, $\sigma_{2_{g\ell}}^{*2} = \sigma_{u_{g\ell}}^{*2} + N\sigma_{\lambda_{g\ell}}^{*2}$, and $\sigma_{4_{g\ell}}^{*2} = \sigma_{u_{g\ell}}^{*2} + T\sigma_{\alpha_{g\ell}}^{*2} + N\sigma_{\lambda_{g\ell}}^{*2}$. It was shown in Appendix 3B that $\sigma_{1_{g\ell}}^{*2}, \sigma_{2_{g\ell}}^{*2}, \sigma_{u_{g\ell}}^{*2}$, and $\sigma_{4_{g\ell}}^{*2}$ are the distinct characteristic roots of $V_{g\ell}$ of multiplicity $N-1, T-1, (N-1)(T-1)$, and 1, with $C_1, C_2, C_3$, and $C_4$ as the matrices of their corresponding characteristic vectors.

We can rewrite $V$ as

$$V = V_1 \otimes \left(\frac{1}{T}A - \frac{1}{NT}J\right) + V_2 \otimes \left(\frac{1}{N}D - \frac{1}{NT}J\right) + \Omega_u^* \otimes \tilde{Q} + V_4 \otimes \left(\frac{1}{NT}J\right), \tag{5.2.6}$$

where $V_1 = (\sigma_{1_{g\ell}}^{*2})$, $V_2 = (\sigma_{2_{g\ell}}^{*2})$, and $V_4 = (\sigma_{4_{g\ell}}^{*2})$ all of dimension $G \times G$. Using the fact that $[(1/T)A - (1/NT)J]$, $[(1/N)D - (1/NT)J]$, $\tilde{Q}$, and $[(1/NT)J]$ are symmetric idempotent matrices, mutually orthogonal, and sum to the identity matrix $I_{NT}$, we can write down the inverse of $V$ explicitly as (Avery 1977; Baltagi 1980)[5]

$$V^{-1} = V_1^{-1} \otimes \left(\frac{1}{T}A - \frac{1}{NT}J\right) + V_2^{-1} \otimes \left(\frac{1}{N}D - \frac{1}{NT}J\right) + \Omega_u^{*-1} \otimes \tilde{Q} + V_4^{-1} \otimes \left(\frac{1}{NT}J\right). \tag{5.2.7}$$

[5] One can check that (5.2.7) is indeed the inverse of (5.2.6) by repeatedly using the formulas of the Kronecker products: $(B + C) \otimes A = B \otimes A + C \otimes A$, $(A \otimes B)(C \otimes D) = AC \otimes BD$, provided the product of these matrices exists (Theil 1971, Chapter 7, Section 7.2).

The generalized least-squares (GLS) estimators of $\boldsymbol{\mu}^*$ and $\boldsymbol{\pi}$ are obtained by minimizing the distance function

$$[\mathbf{y} - (I_G \otimes \mathbf{e}_{NT})\boldsymbol{\mu}^* - \mathbf{X}\boldsymbol{\pi}]' V^{-1} [\mathbf{y} - (I_G \otimes \mathbf{e}_{NT})\boldsymbol{\mu}^* - \mathbf{X}\boldsymbol{\pi}] \tag{5.2.8}$$

Taking partial derivatives of (5.2.8) with respect to $\boldsymbol{\mu}^*$ and $\boldsymbol{\pi}$, we obtain the first-order conditions

$$-(I_G \otimes \mathbf{e}_{NT})' V^{-1} [\mathbf{y} - (I_G \otimes \mathbf{e}_{NT})\boldsymbol{\mu}^* - \mathbf{X}\boldsymbol{\pi}] = \mathbf{0}, \tag{5.2.9}$$

$$-X' V^{-1} [\mathbf{y} - (I_G \otimes \mathbf{e}_{NT})\boldsymbol{\mu}^* - \mathbf{X}\boldsymbol{\pi}] = \mathbf{0}. \tag{5.2.10}$$

Solving (5.2.9) and making use of the relations $[(1/T)A - (1/NT)J]\mathbf{e}_{NT} = \mathbf{0}$, $[(1/N)D - (1/NT)J]\mathbf{e}_{NT} = \mathbf{0}$, $\tilde{Q}\mathbf{e}_{NT} = \mathbf{0}$, and $(1/NT)J\mathbf{e}_{NT} = \mathbf{e}_{NT}$, we have

$$\hat{\boldsymbol{\mu}}^* = \left(I_G \otimes \frac{1}{NT}\mathbf{e}'_{NT}\right)(\mathbf{y} - \mathbf{X}\boldsymbol{\pi}). \tag{5.2.11}$$

Substituting (5.2.11) into (5.2.10), we have the GLS estimator of $\boldsymbol{\pi}$ as[6]

$$\hat{\boldsymbol{\pi}}_{\text{GLS}} = [\mathbf{X}' \tilde{V}^{-1} \mathbf{X}]^{-1} (\mathbf{X}' \tilde{V}^{-1} \mathbf{y}), \tag{5.2.12}$$

where

$$\tilde{V}^{-1} = V_1^{-1} \otimes \left(\frac{1}{T}A - \frac{1}{NT}J\right) + V_2^{-1} \otimes \left(\frac{1}{N}D - \frac{1}{NT}J\right) + \Omega_u^{*-1} \otimes \tilde{Q}. \tag{5.2.13}$$

If $E(\boldsymbol{\epsilon}_g \boldsymbol{\epsilon}'_\ell) = \mathbf{0}$ for $g \neq \ell$ then $V$ is block-diagonal, and equation (5.2.12) is reduced to applying the GLS estimation method to each equation separately. If both $N$ and $T$ tend to infinity and $N/T$ tends to a nonzero constant, then $\lim V_1^{-1} = \mathbf{0}$, $\lim V_2^{-1} = \mathbf{0}$, and $\lim V_4^{-1} = \mathbf{0}$. Equation (5.2.12) becomes the least-squares dummy variable (or fixed-effects) estimator for the seemingly

[6] If only the first $M$ out of $G$ equations have nonzero intercepts, we estimate the first $M$ intercepts by $\{[I_M, (V_4^{MM})^{-1} V_4^{M(G-M)}] \otimes (1/NT)\mathbf{e}'_{NT}\}(\mathbf{y} - X\boldsymbol{\pi})$ and estimate $\boldsymbol{\pi}$ by $[X' V^{*-1} X]^{-1} [X' V^{*-1} \mathbf{y}]$, where $I_M$ is the $M$-rowed identity matrix, $V_4^{MM}$ and $V_4^{M(G-M)}$ are the corresponding $M \times M$ and $M \times (G - M)$ partitioned matrices of

$$V_4^{-1} = \begin{bmatrix} V_4^{MM} & V_4^{M(G-M)} \\ V_4^{(G-M)M} & V_4^{(G-M)(G-M)} \end{bmatrix}$$

and

$$V^{*-1} = \tilde{V}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & V_4^{(G-M)(G-M)} - V_4^{(G-M)M} (V_4^{MM})^{-1} V_4^{M(G-M)} \end{bmatrix} \otimes \frac{1}{NT} J.$$

For details, see Prucha (1983).

unrelated regression case,

$$\operatorname{plim} \hat{\boldsymbol{\pi}}_{\mathrm{GLS}} = \operatorname*{plim}_{\substack{N\to\infty \\ T\to\infty}} \left[\frac{1}{NT}\mathbf{X}'\left(\Omega_u^{*-1}\otimes \tilde{Q}\right)\mathbf{X}\right]^{-1} \cdot \left[\frac{1}{NT}X'\left(\Omega_u^{*-1}\otimes \tilde{Q}\right)\mathbf{y}\right]. \tag{5.2.14}$$

In the case of the standard reduced form, $X_1 = X_2 = \cdots = X_G = \bar{X}$,

$$\begin{aligned}\hat{\boldsymbol{\pi}}_{\mathrm{GLS}} = &\left[V_1^{-1}\otimes \bar{X}'\left(\frac{1}{T}A - \frac{1}{NT}J\right)\bar{X}\right. \\ &\left. + V_2^{-1}\otimes \bar{X}'\left(\frac{1}{N}D - \frac{1}{NT}J\right)\bar{X} + \Omega_u^{*-1}\otimes \bar{X}'\tilde{Q}\bar{X}\right]^{-1} \\ &\cdot \left\{\left[V_1^{-1}\otimes \bar{X}'\left(\frac{1}{T}A - \frac{1}{NT}J\right)\right]\mathbf{y}\right. \\ &\left. + \left[V_2^{-1}\otimes \bar{X}'\left(\frac{1}{N}D - \frac{1}{NT}J\right)\right]\mathbf{y} + \left[\Omega_u^{*-1}\otimes \bar{X}'\tilde{Q}\right]\mathbf{y}\right\}.\end{aligned} \tag{5.2.15}$$

We know that in the conventional case when no restriction is imposed on the reduced-form coefficients vector $\boldsymbol{\pi}$, estimating each equation by the least-squares method yields the best linear unbiased estimate. Equation (5.2.15) shows that in a seemingly unrelated regression model with error components, the fact that each equation has an identical set of explanatory variables is not a sufficient condition for the GLS performed on the whole system to be equivalent to estimating each equation separately.

Intuitively, by stacking different equations together we shall gain efficiency in the estimates, because knowing the residual of the $\ell$th equation helps in predicting the $g$th equation when the covariance terms between different equations are nonzero. For instance, if the residuals are normally distributed, $E(\boldsymbol{\epsilon}_g \mid \boldsymbol{\epsilon}_\ell) = \mathrm{Cov}(\boldsymbol{\epsilon}_g, \boldsymbol{\epsilon}_\ell)\mathrm{Var}(\boldsymbol{\epsilon}_\ell)^{-1}\boldsymbol{\epsilon}_\ell \neq \mathbf{0}$. To adjust for this nonzero mean, it would be appropriate to regress $\mathbf{y}_g - \mathrm{Cov}(\boldsymbol{\epsilon}_g, \boldsymbol{\epsilon}_\ell)\mathrm{Var}(\boldsymbol{\epsilon}_\ell)^{-1}\boldsymbol{\epsilon}_\ell$ on $(\mathbf{e}_{NT}, X_g)$. Although in general $\boldsymbol{\epsilon}_\ell$ is unknown, asymptotically there is no difference if we replace it by the least-squares residual, $\hat{\boldsymbol{\epsilon}}_\ell$. However, if the explanatory variables in different equations are identical, namely, $X_g = X_\ell = \bar{X}$, there is no gain in efficiency by bringing different equations together when the cross equation covariances are unrestricted; because $\mathrm{Cov}(\boldsymbol{\epsilon}_g, \boldsymbol{\epsilon}_\ell) = \sigma_{\epsilon_{g\ell}} I_{NT}$, $\mathrm{Var}(\boldsymbol{\epsilon}_\ell) = \sigma_{\epsilon_{\ell\ell}} I_{NT}$, and $\hat{\boldsymbol{\epsilon}}_\ell$ is orthogonal to $(\mathbf{e}_{NT}, X_g)$ by construction, the variable $\sigma_{\epsilon_{g\ell}}\sigma_{\epsilon_{\ell\ell}}^{-1}\hat{\boldsymbol{\epsilon}}_\ell$ can have no effect on the estimate of $(\mu_g, \boldsymbol{\pi}_g')$ when it is subtracted from $\mathbf{y}_g$. But the same cannot be said for the error-components case, because $\mathrm{Cov}(\boldsymbol{\epsilon}_g, \boldsymbol{\epsilon}_\ell)\mathrm{Var}(\boldsymbol{\epsilon}_\ell)^{-1}$ is not proportional to an identity matrix. The weighted

variable $\text{Cov}(\boldsymbol{\epsilon}_g, \boldsymbol{\epsilon}_\ell)\text{Var}(\boldsymbol{\epsilon}_\ell)^{-1}\hat{\boldsymbol{\epsilon}}_\ell$ is no longer orthogonal to $(\mathbf{e}_{NT}, \bar{X})$. Therefore, in the error-components case it remains fruitful to exploit the covariances between different equations to improve the accuracy of the estimates.

When $V_1$, $V_2$, and $\Omega_u^*$ are unknown, we can replace them by their consistent estimates. In Chapter 3, we discussed methods of estimating variance components. These techniques can be straightforwardly applied to the multiple-equations model as well (Avery 1977; Baltagi 1980).

The model discussed earlier assumes the existence of both individual and time effects. Suppose we believe that the covariances of some of the components are 0. The same procedure can be applied to the simpler model with some slight modifications. For example, if the covariance of the residuals between equations $g$ and $\ell$ is composed of only two components (an individual effect and overall effect), then $\sigma^2_{\lambda_{g\ell}} = 0$. Hence, $\sigma^{*2}_{1_{g\ell}} = \sigma^{*2}_{4_{g\ell}}$, and $\sigma^{*2}_{2_{g\ell}} = \sigma^{*2}_{u_{g\ell}}$. These adjusted roots can be substituted into the appropriate positions in (5.2.6) and (5.2.7), with coefficient estimates following directly from (5.2.12).

## 5.3 ESTIMATION OF STRUCTURAL EQUATIONS

### 5.3.1 Estimation of a Single Equation in the Structural Model

As (5.2.12) shows, the GLS estimator of the slope coefficients is invariant against centering the data around overall sample means; so for ease of exposition we shall assume that there is an intercept term and that all sample observations are measured as deviations from their respective overall means and consider the $g$th structural equation as

$$\begin{aligned}\underset{NT\times 1}{\mathbf{y}_g} &= \mathbf{Y}_g\gamma_g + \mathbf{X}_g\boldsymbol{\beta}_g + \mathbf{v}_g \\ &= \mathbf{W}_g\boldsymbol{\theta}_g + \mathbf{v}_g, \quad g = 1, \ldots, G, \end{aligned} \tag{5.3.1}$$

where $\mathbf{Y}_g$ is an $NT \times (G_g - 1)$ matrix of $NT$ observations of $G_g - 1$ included joint dependent variables, $X_g$ is an $NT \times K_g$ matrix of $NT$ observations of $K_g$ included exogenous variables, $W_g = (Y_g, X_g)$, and $\theta_g = (\gamma_g', \boldsymbol{\beta}_g')'$, The $\mathbf{v}_g$ is an $NT \times 1$ vector of error terms that are independent of $\mathbf{x}$,

$$\mathbf{v}_g = (I_N \otimes \mathbf{e}_T)\boldsymbol{\alpha}_g + (\mathbf{e}_N \otimes I_T)\boldsymbol{\lambda}_g + \mathbf{u}_g, \tag{5.3.2}$$

with $\boldsymbol{\alpha}_g = (\alpha_{1g}, \ldots, \alpha_{Ng})'$, $\boldsymbol{\lambda}_g = (\lambda_{1g}, \ldots, \lambda_{Tg})'$, and $\mathbf{u}_g = (u_{11g}, \ldots, u_{1Tg}, u_{21g}, \ldots, u_{NTg})'$ satisfying assumption (5.1.13). So the covariance matrix between the $g$th and the $\ell$th structural equations is

$$\begin{aligned}\Sigma_{g\ell} &= E(\mathbf{v}_g\mathbf{v}_\ell') = \sigma^2_{\alpha_{g\ell}}A + \sigma^2_{\lambda_{g\ell}}D + \sigma^2_{u_{g\ell}}I_{NT} \\ &= \sigma^2_{1_{g\ell}}\left(\frac{1}{T}A - \frac{1}{NT}J\right) + \sigma^2_{2_{g\ell}}\left(\frac{1}{N}D - \frac{1}{NT}J\right) \\ &\quad + \sigma^2_{3_{g\ell}}\tilde{Q} + \sigma^2_{4_{g\ell}}\left(\frac{1}{NT}J\right), \end{aligned} \tag{5.3.3}$$

where $\sigma^2_{1_{g\ell}} = \sigma^2_{u_{g\ell}} + T\sigma^2_{\alpha_{g\ell}}$, $\sigma^2_{2_{g\ell}} = \sigma^2_{u_{g\ell}} + N\sigma^2_{\lambda_{g\ell}}$, $\sigma^2_{3_{g\ell}} = \sigma^2_{u_{g\ell}}$, and $\sigma^2_{4_{g\ell}} = \sigma^2_{u_{g\ell}} + T\sigma^2_{\alpha_{g\ell}} + N\sigma^2_{\lambda_{g\ell}}$. We also assume that each equation in (5.3.1) satisfies the rank condition for identification with $K \geq G_g + K_g - 1$, $g = 1, \ldots, G$.

We first consider estimation of a single equation in the structural model. To estimate the $g$th structural equation, we take account only of the a priori restrictions affecting that equation and ignore the restrictions affecting all other equations. Therefore, suppose we are interested in estimating the first equation. The *limited-information* principle of estimating this equation is equivalent to the full-information estimation of the system

$$\begin{aligned} y_{1_{it}} &= \mathbf{w}'_{1_{it}}\boldsymbol{\theta}_1 + v_{1_{it}}, \\ y_{2_{it}} &= \mathbf{x}'_{it}\boldsymbol{\pi}_2 + \boldsymbol{\epsilon}_{2_{it}}, \\ &\vdots \\ y_{G_{it}} &= \mathbf{x}'_{it}\boldsymbol{\pi}_G + \boldsymbol{\epsilon}_{G_{it}}, \qquad i = 1, \ldots, N, \\ & \qquad\qquad\qquad\quad t = 1, \ldots, T, \end{aligned} \tag{5.3.4}$$

where there are no restrictions on $\boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_G$.

We can apply the usual two-stage least-squares (2SLS) method to estimate the first equation in (5.3.4). The 2SLS estimator is consistent. However, if the $v_{1_{it}}$ are not independently identically distributed over $i$ and $t$, the 2SLS estimator is not efficient even within the limited-information context. To allow for arbitrary heteroscedasticity and serial correlation in the residuals, we can generalize Chamberlain's (1982, 1984) minimum-distance or generalized 2SLS estimator.

We first consider the minimum-distance estimator. Suppose $T$ is fixed and $N$ tends to infinity. Stacking the $T$ period equations for a single individual's behavioral equation into one system, we create a model of $GT$ equations,

$$\begin{aligned} \underset{T\times 1}{\mathbf{y}_{1_i}} &= \mathbf{W}_{1_i}\boldsymbol{\theta}_1 + \mathbf{v}_{1_i}, \\ \underset{T\times 1}{\mathbf{y}_{2_i}} &= \mathbf{X}_i\boldsymbol{\pi}_2 + \boldsymbol{\epsilon}_{2_i}, \\ &\vdots \\ \mathbf{y}_{G_i} &= \mathbf{X}_i\boldsymbol{\pi}_G + \boldsymbol{\epsilon}_{G_i}, \qquad i = 1, \ldots, N. \end{aligned} \tag{5.3.5}$$

Let $\mathbf{y}'_i = (\mathbf{y}'_{1_i}, \ldots, \mathbf{y}'_{G_i})$. The reduced form of $\mathbf{y}_i$ is

$$\mathbf{y}_i = \begin{bmatrix} \mathbf{y}_{1_i} \\ \mathbf{y}_{2_i} \\ \vdots \\ \mathbf{y}_{G_i} \end{bmatrix} = (I_G \otimes \tilde{\mathbf{X}}_i)\tilde{\boldsymbol{\pi}} + \boldsymbol{\epsilon}_i, \qquad i = 1, \ldots, N, \tag{5.3.6}$$

where

$$\underset{T\times TK}{\tilde{\mathbf{X}}_i} = \begin{bmatrix} \mathbf{x}'_{i1} & & & \mathbf{0} \\ & \mathbf{x}'_{i2} & & \\ & & \ddots & \\ \mathbf{0} & & & \mathbf{x}'_{iT} \end{bmatrix},$$

$$\tilde{\boldsymbol{\pi}} = \operatorname{vec}(\tilde{\Pi}'), \tag{5.3.7}$$

$$\underset{GT\times K}{\tilde{\Pi}} = \Pi \otimes \mathbf{e}_T, \quad \text{and} \quad \Pi = E(\mathbf{y}_{it} \mid \mathbf{x}_{it}). \tag{5.3.8}$$

The unconstrained least-squares regression of $\mathbf{y}_i$ on $(I_G \otimes \tilde{\mathbf{X}}_i)$ yields a consistent estimate of $\tilde{\boldsymbol{\pi}}, \hat{\tilde{\boldsymbol{\pi}}}$. If $\boldsymbol{\epsilon}_i$ are independently distributed over $i$, then $\sqrt{N}(\hat{\tilde{\boldsymbol{\pi}}} - \tilde{\boldsymbol{\pi}})$ is asymptotically normally distributed, with mean 0 and variance–covariance matrix

$$\underset{GTK\times GTK}{\tilde{\Omega}} = \left(I_G \otimes \Phi_{xx}^{-1}\right) \tilde{V} \left(I_G \otimes \Phi_{xx}^{-1}\right), \tag{5.3.9}$$

where $\Phi_{xx} = E\tilde{\mathbf{X}}'_i\tilde{\mathbf{X}}_i = \operatorname{diag}\{E(\mathbf{x}_{i1}\mathbf{x}'_{i1}), \ldots, E(\mathbf{x}_{iT}\mathbf{x}'_{iT})\}$, and $\tilde{V}$ is a $GTK \times GTK$ matrix, with the $g\ell$th block a $TK \times TK$ matrix of the form

$$\tilde{V}_{g\ell} = E \begin{bmatrix} \boldsymbol{\epsilon}_{gi1}\boldsymbol{\epsilon}_{\ell i1}\mathbf{x}_{i1}\mathbf{x}'_{i1} & \boldsymbol{\epsilon}_{gi1}\boldsymbol{\epsilon}_{\ell i2}\mathbf{x}_{i1}\mathbf{x}'_{i2} & \cdots & \boldsymbol{\epsilon}_{gi1}\boldsymbol{\epsilon}_{\ell iT}\mathbf{x}_{i1}\mathbf{x}'_{iT} \\ \boldsymbol{\epsilon}_{gi2}\boldsymbol{\epsilon}_{\ell i1}\mathbf{x}_{i2}\mathbf{x}'_{i1} & \boldsymbol{\epsilon}_{gi2}\boldsymbol{\epsilon}_{\ell i2}\mathbf{x}_{i2}\mathbf{x}'_{i2} & \cdots & \boldsymbol{\epsilon}_{gi2}\boldsymbol{\epsilon}_{\ell iT}\mathbf{x}_{i2}\mathbf{x}'_{iT} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{\epsilon}_{giT}\boldsymbol{\epsilon}_{\ell i1}\mathbf{x}_{iT}\mathbf{x}'_{i1} & \boldsymbol{\epsilon}_{giT}\boldsymbol{\epsilon}_{\ell i2}\mathbf{x}_{iT}\mathbf{x}'_{i2} & \cdots & \boldsymbol{\epsilon}_{giT}\boldsymbol{\epsilon}_{\ell iT}\mathbf{x}_{iT}\mathbf{x}'_{iT} \end{bmatrix}. \tag{5.3.10}$$

One can obtain a consistent estimator of $\tilde{\Omega}$ by replacing the population moments in $\tilde{\Omega}$ by the corresponding sample moments (e.g., $E\mathbf{x}_{i1}\mathbf{x}'_{i1}$ is replaced by $\sum_{i=1}^{N} \mathbf{x}_{i1}\mathbf{x}'_{i1}/N$).

Let $\boldsymbol{\theta}' = (\boldsymbol{\theta}'_1, \boldsymbol{\pi}'_2, \ldots, \boldsymbol{\pi}'_G)$, and specify the restrictions on $\tilde{\boldsymbol{\pi}}$ by the condition that $\tilde{\boldsymbol{\pi}} = \tilde{\mathbf{f}}(\boldsymbol{\theta})$. Choose $\boldsymbol{\theta}$ to minimize the following distance function:

$$[\hat{\tilde{\boldsymbol{\pi}}} - \tilde{\mathbf{f}}(\boldsymbol{\theta})]'\hat{\tilde{\Omega}}^{-1}[\hat{\tilde{\boldsymbol{\pi}}} - \tilde{\mathbf{f}}(\boldsymbol{\theta})]. \tag{5.3.11}$$

Then $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normally distributed with mean 0 and variance–covariance matrix $(\tilde{F}'\tilde{\Omega}^{-1}\tilde{F})^{-1}$, where $\tilde{F} = \partial\tilde{\mathbf{f}}/\partial\boldsymbol{\theta}'$. Noting that $\tilde{\Pi} = \Pi \otimes \mathbf{e}_T$, and evaluating the partitioned inverse, we obtain the asymptotic variance–covariance matrix of $\sqrt{N}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1)$ as

$$\left(\tilde{\Phi}_{w_1x}\Psi_{11}^{-1}\tilde{\Phi}'_{w_1x}\right)^{-1}, \tag{5.3.12}$$

where $\tilde{\Phi}_{w_1x} = [E(\mathbf{w}_{1_{i1}}\mathbf{x}'_{i1}), E(\mathbf{w}_{1_{i2}}\mathbf{x}'_{i2}), \ldots, E(\mathbf{w}_{1_{iT}}\mathbf{x}'_{iT})]$, and

$$\Psi_{11} = E\begin{bmatrix} v_{1_{i1}}^2\mathbf{x}_{i1}\mathbf{x}'_{i1} & v_{1_{i1}}v_{1_{i2}}\mathbf{x}_{i1}\mathbf{x}'_{i2} & \cdots & v_{1_{i1}}v_{1_{iT}}\mathbf{x}_{i1}\mathbf{x}'_{iT} \\ v_{1_{i2}}v_{1_{i1}}\mathbf{x}_{i2}\mathbf{x}'_{i1} & v_{1_{i2}}^2\mathbf{x}_{i2}\mathbf{x}'_{i2} & \cdots & v_{1_{i2}}v_{1_{iT}}\mathbf{x}_{i2}\mathbf{x}'_{iT} \\ \vdots & \vdots & & \vdots \\ v_{1_{iT}}v_{1_{i1}}\mathbf{x}_{iT}\mathbf{x}'_{i1} & v_{1_{iT}}v_{1_{i2}}\mathbf{x}_{iT}\mathbf{x}'_{i2} & \cdots & v_{1_{iT}}v_{1_{iT}}\mathbf{x}_{iT}\mathbf{x}'_{iT} \end{bmatrix}. \tag{5.3.13}$$

The limited-information minimum-distance estimator of (5.3.11) is asymptotically equivalent to the following generalization of the 2SLS estimator:

$$\hat{\theta}_{1,\text{G2SLS}} = \left(\tilde{\mathbf{S}}_{w_1x}\hat{\Psi}_{11}^{-1}\tilde{\mathbf{S}}'_{w_1x}\right)^{-1}\left(\tilde{\mathbf{S}}_{w_1x}\hat{\Psi}_{11}^{-1}\mathbf{s}_{xy_1}\right), \tag{5.3.14}$$

where

$$\tilde{\mathbf{S}}_{w_1x} = \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{w}_{1_{i1}}\mathbf{x}'_{i1}, \frac{1}{N}\sum_{i=1}^{N}\mathbf{w}_{1_{i2}}\mathbf{x}'_{i2}, \ldots, \frac{1}{N}\sum_{i=1}^{N}\mathbf{w}_{1_{iT}}\mathbf{x}'_{iT}\right),$$

$$\mathbf{s}_{xy_1} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{i1}y_{1_{i1}} \\ \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{i2}y_{1_{i2}} \\ \vdots \\ \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{iT}y_{1_{iT}} \end{bmatrix},$$

$$\hat{\Psi}_{11} = \frac{1}{N}\begin{bmatrix} \sum_{i=1}^{N}\hat{v}_{1_{i1}}^2\mathbf{x}_{i1}\mathbf{x}'_{i1} & \sum_{i=1}^{N}\hat{v}_{1_{i1}}\hat{v}_{1_{i2}}\mathbf{x}_{i1}\mathbf{x}'_{i2} & \cdots & \sum_{i=1}^{N}\hat{v}_{1_{i1}}\hat{v}_{1_{iT}}\mathbf{x}_{i1}\mathbf{x}'_{iT} \\ \vdots & \vdots & & \vdots \\ \sum_{i=1}^{N}\hat{v}_{1_{iT}}\hat{v}_{1_{i1}}\mathbf{x}_{iT}\mathbf{x}'_{i1} & \sum_{i=1}^{N}\hat{v}_{1iT}\hat{v}_{1i2}\mathbf{x}_{iT}\mathbf{x}'_{i2} & \cdots & \sum_{i=1}^{N}\hat{v}_{1_{iT}}\hat{v}_{1_{iT}}\mathbf{x}_{iT}\mathbf{x}'_{iT} \end{bmatrix},$$

and $\hat{v}_{1_{it}} = y_{1_{it}} - \mathbf{w}'_{1_{it}}\hat{\boldsymbol{\theta}}_1$, with $\hat{\boldsymbol{\theta}}_1$ any consistent estimator of $\boldsymbol{\theta}_1$. The generalized 2SLS converges to the 2SLS if $v_{1_{it}}$ is independently identically distributed over $i$ and $t$ and $E\mathbf{x}_{it}\mathbf{x}'_{it} = E\mathbf{x}_{is}\mathbf{x}'_{is}$. But the generalized 2SLS, like the minimum-distance estimator of (5.3.11), makes allowance for the heteroscedasticity and arbitrary serial correlation in $v_{1_{it}}$, whereas the 2SLS does not.

When the variance–covariance matrix $\sum_{gg}$ possesses an error-component structure as specified in (5.3.3), although both the 2SLS estimator and the minimum-distance estimator of (5.3.11) (or the generalized 2SLS estimator) remain consistent, they are no longer efficient even within a limited-information framework, because, as shown in the last section, when there are restrictions on the variance–covariance matrix the least-squares estimator of the unconstrained $\Pi$ is not as efficient as the generalized least-squares estimator.[7] An efficient estimation method has to exploit the known restrictions on the error structure. Baltagi (1981a) has suggested using the following error-component two-stage

[7] See Chapter 3, footnote 22.

least-squares (EC2SLS) method to obtain a more efficient estimator of the unknown parameters in the $g$th equation.

Transforming (5.3.1) by the eigenvectors of $\sum_{gg}$, $C_1'$, $C_2'$, and $C_3'$, we have[8]

$$\mathbf{y}_g^{(h)} = Y_g^{(h)}\gamma_g + \mathbf{X}_g^{(h)}\boldsymbol{\beta}_g + \mathbf{v}_g^{(h)} = \mathbf{W}_g^{(h)}\boldsymbol{\theta}_g + \mathbf{v}_g^{(h)}, \tag{5.3.15}$$

where $\mathbf{y}_g^{(h)} = C_h'\mathbf{y}_g$, $W_g^{(h)} = C_h'W_g$, $\mathbf{v}_g^{(h)} = C_h'\mathbf{v}_g$ for $h = 1, 2, 3$, and $C_1'$, $C_2'$, and $C_3'$ are as defined in Appendix 3B. The transformed disturbance term $\mathbf{v}_g^{(h)}$ is mutually orthogonal and has a covariance matrix proportional to an identity matrix. We can therefore use $X^{(h)} = C_h'X$ as the instruments and apply the Aitken estimation procedure to the system of equations

$$\begin{bmatrix} X^{(1)\prime}\mathbf{y}_g^{(1)} \\ X^{(2)\prime}\mathbf{y}_g^{(2)} \\ X^{(3)\prime}\mathbf{y}_g^{(3)} \end{bmatrix} = \begin{bmatrix} X^{(1)\prime}W_g^{(1)} \\ X^{(2)\prime}W_g^{(2)} \\ X^{(3)\prime}W_g^{(3)} \end{bmatrix} \begin{bmatrix} \gamma_g \\ \boldsymbol{\beta}_g \end{bmatrix} + \begin{bmatrix} X^{(1)\prime}\mathbf{v}_g^{(1)} \\ X^{(2)\prime}\mathbf{v}_g^{(2)} \\ X^{(3)\prime}\mathbf{v}_g^{(3)} \end{bmatrix}. \tag{5.3.16}$$

The resulting Aitken estimator of $(\gamma_g', \boldsymbol{\beta}_g')$ is

$$\hat{\boldsymbol{\theta}}_{g,\text{EC2SLS}} = \left\{ \sum_{h=1}^{3} \left[ \frac{1}{\sigma_{hgg}^2} W_g^{(h)\prime} P_X(h) W_g^{(h)} \right] \right\}^{-1} \left\{ \sum_{h=1}^{3} \left[ \frac{1}{\sigma_{hgg}^2} W_g^{(h)\prime} P_X(h) \mathbf{y}_g^{(h)} \right] \right\}, \tag{5.3.17}$$

where $P_X(h) = X^{(h)}(X^{(h)\prime}X^{(h)})^{-1}X^{(h)\prime}$. It is a weighted combination of the between-groups, between-time-periods, and within-groups 2SLS estimators of $(\gamma_g', \boldsymbol{\beta}_g')$. The weights $\sigma_{hgg}^2$ can be estimated by substituting the transformed 2SLS residuals in the usual variance formula,

$$\hat{\sigma}_{hgg}^2 = \left(\mathbf{y}_g^{(h)} - W_g^{(h)}\hat{\boldsymbol{\theta}}_{g,\text{2SLS}}^{(h)}\right)' \left(\mathbf{y}_g^{(h)} - W_g^{(h)}\hat{\boldsymbol{\theta}}_{g,\text{2SLS}}^{(h)}\right) / n(h) , \tag{5.3.18}$$

where $\hat{\boldsymbol{\theta}}_{g,\text{2SLS}}^{(h)} = [W_g^{(h)\prime}P_X(h)W_g^{(h)}]^{-1}[W_g^{(h)\prime}P_X(h)\mathbf{y}_g^{(h)}]$, and $n(1) = N - 1$, $n(2) = T - 1$, $n(3) = (N - 1)(T - 1)$. If $N \to \infty$, $T \to \infty$, and $N/T$ tends to a nonzero constant, then the probability limit of the EC2SLS tends to the 2SLS estimator based on the within-groups variation alone.

In the special case in which the source of correlation between some of the regressors and residuals comes from the unobserved time-invariant individual effects alone, the correlations between them can be removed by removing

[8] As indicated earlier, we have assumed here that all variables are measured as deviations from their respective overall means. There is no loss of generality in this formulation, because the intercept $\mu_g$ is estimated by $\hat{\mu}_g = (1/NT)\mathbf{e}_{NT}'(\mathbf{y}_g - W_g\hat{\boldsymbol{\theta}}_g)$. Because $C_h'\mathbf{e}_{NT} = \mathbf{0}$ for $h = 1, 2, 3$, the only terms pertinent to our discussion are $C_h$ for $h = 1, 2, 3$.

the time-invariant component from the corresponding variables. Thus, instruments for the correlated regressors can be chosen from "inside" the equation, as opposed to the conventional method of being chosen from "outside" the equation. Hausman and Taylor (1981) noted that for variables that are time-varying and are correlated with $\alpha_{ig}$, transforming them into deviations from their corresponding time means provides legitimate instruments, because they will no longer be correlated with $\alpha_{ig}$. For variables that are time-invariant, the time means of those variables that are uncorrelated with $\alpha_{ig}$ can be used as instruments. Hence, a necessary condition for identification of all the parameters within a single-equation framework is that the number of time-varying variables that are uncorrelated with $\alpha_{ig}$ be at least as great as the number of time-invariant variables that are correlated with $\alpha_{ig}$. They further showed that when the variance-component structure of the disturbance term is taken account of, the instrumental-variable estimator with instruments chosen this way is efficient among the single-equation estimators.

### 5.3.2 Estimation of the Complete Structural System

The single-equation estimation method considered earlier ignores restrictions in all equations in the structural system except the one being estimated. In general, we expect to get more efficient estimates if we consider the additional information contained in the other equations. In this subsection we consider the full-information estimation methods.

Let $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_G')'$, $\mathbf{v} = (\mathbf{v}_1', \ldots, \mathbf{v}_G')'$,

$$W = \begin{bmatrix} W_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & W_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & W_G \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}_1 \\ \vdots \\ \boldsymbol{\theta}_G \end{bmatrix}.$$

We write the set of $G$ structural equations as

$$\mathbf{y} = W\boldsymbol{\theta} + \mathbf{v}. \tag{5.3.19}$$

We can estimate the system (5.3.19) by the three-stage least-squares (3SLS) method. But just as in the limited-information case, the 3SLS estimator is efficient only if $(v_{1_{it}}, v_{2_{it}}, \ldots, v_{G_{it}})$ are independently identically distributed over $i$ and $t$. To allow for arbitrary heteroscedasticity or serial correlation, we can use the full-information minimum-distance estimator or the generalized 3SLS estimator.

We first consider the minimum-distance estimator. When $T$ is fixed and $N$ tends to infinity, we can stack the $T$ period equations for an individual's

behavioral equation into a system to create a model of $GT$ equations,

$$\begin{aligned}
\underset{T\times 1}{\mathbf{y}_{1_i}} &= W_{1_i}\boldsymbol{\theta}_1 + \mathbf{v}_{1_i},\\
\mathbf{y}_{2_i} &= W_{2i}\boldsymbol{\theta}_2 + \mathbf{v}_{2_i},\\
&\vdots\\
\mathbf{y}_{G_i} &= W_{G_i}\boldsymbol{\theta}_G + \mathbf{v}_{G_i}, \qquad i = 1, \ldots, N.
\end{aligned} \tag{5.3.20}$$

We obtain a minimum-distance estimator of $\boldsymbol{\theta}$ by choosing $\hat{\boldsymbol{\theta}}$ to minimize $[\hat{\tilde{\boldsymbol{\pi}}} - \tilde{\mathbf{f}}(\boldsymbol{\theta})]'\hat{\tilde{\Omega}}^{-1}[\hat{\tilde{\boldsymbol{\pi}}} - \tilde{\mathbf{f}}(\boldsymbol{\theta})]$, where $\hat{\tilde{\boldsymbol{\pi}}}$ is the unconstrained least-squares estimator of regressing $\mathbf{y}_i$ on $(I_G \otimes \tilde{\mathbf{X}}_i)$, and $\hat{\tilde{\Omega}}$ is a consistent estimate of $\tilde{\Omega}$ [equation (5.3.9)]. Noting that $\tilde{\Pi} = \Pi \otimes \mathbf{e}_T$ and $\operatorname{vec}(\Pi') = \boldsymbol{\pi} = \operatorname{vec}([-\Gamma^{-1}B]')$ for all elements of $\Gamma$ and $B$ not known a priori, and making use of the formula $\partial\boldsymbol{\pi}/\partial\boldsymbol{\theta}'$ [equation (3.8.25)], we can show that if $\mathbf{v}_i$ are independently distributed over $i$, then $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normally distributed, with mean 0 and variance–covariance matrix

$$\{\Phi_{wx}\Psi^{-1}\Phi'_{wx}\}^{-1}, \tag{5.3.21}$$

where

$$\begin{aligned}
\Phi_{wx} &= \begin{bmatrix} \tilde{\Phi}_{w_1x} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \tilde{\Phi}_{w_2x} & & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \tilde{\Phi}_{w_Gx} \end{bmatrix},\\
\tilde{\Phi}_{w_gx} &= [E(\mathbf{w}_{g_{i1}}\mathbf{x}'_{i1}), E(\mathbf{w}_{g_{i2}}\mathbf{x}'_{i2}), \ldots, E(\mathbf{w}_{g_{iT}}\mathbf{x}'_{iT})],\\
\underset{GTK\times GTK}{\Psi} &= \begin{bmatrix} \Psi_{11} & \Psi_{12} & \cdots & \Psi_{1G} \\ \Psi_{21} & \Psi_{22} & \cdots & \Psi_{2G} \\ \vdots & \vdots & & \vdots \\ \Psi_{G1} & \Psi_{G2} & \cdots & \Psi_{GG} \end{bmatrix},\\
\underset{TK\times TK}{\Psi_{g\ell}} &= E\begin{bmatrix} v_{g_{i1}}v_{\ell_{i1}}\mathbf{x}_{i1}\mathbf{x}'_{i1} & v_{g_{i1}}v_{\ell_{i2}}\mathbf{x}_{i1}\mathbf{x}'_{i2} & \cdots & v_{g_{i1}}v_{\ell_{iT}}\mathbf{x}_{i1}\mathbf{x}'_{iT} \\ \vdots & \vdots & & \vdots \\ v_{g_{iT}}v_{\ell i1}\mathbf{x}_{iT}\mathbf{x}'_{i1} & v_{g_{iT}}v_{\ell_{i2}}\mathbf{x}_{iT}\mathbf{x}'_{i2} & \cdots & v_{g_{iT}}v_{\ell_{iT}}\mathbf{x}_{iT}\mathbf{x}'_{iT} \end{bmatrix}.
\end{aligned} \tag{5.3.22}$$

We can also estimate (5.3.20) by using a generalized 3SLS estimator,

$$\hat{\boldsymbol{\theta}}_{\text{G3SLS}} = (S_{wx}\hat{\Psi}^{-1}S'_{wx})^{-1}(S_{wx}\hat{\Psi}^{-1}S_{xy}), \tag{5.3.23}$$

where

$$S_{wx} = \begin{bmatrix} \tilde{S}_{w_1x} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \tilde{S}_{w_2x} & & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \tilde{S}_{w_Gx} \end{bmatrix},$$

$$\tilde{S}_{w_gx} = \left[ \frac{1}{N}\sum_{i=1}^{N} \mathbf{w}_{g_{i1}}\mathbf{x}'_{i1}, \frac{1}{N}\sum_{i=1}^{N} \mathbf{w}_{g_{i2}}\mathbf{x}'_{i2}, \ldots, \frac{1}{N}\sum_{i=1}^{N} \mathbf{w}_{g_{iT}}\mathbf{x}'_{iT} \right],$$

$$S_{xy} = \begin{bmatrix} \mathbf{s}_{xy_1} \\ \mathbf{s}_{xy_2} \\ \vdots \\ \mathbf{s}_{xy_G} \end{bmatrix},$$

$$\underset{TK\times 1}{\mathbf{s}_{xy_g}} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_{i1}y_{g_{i1}} \\ \vdots \\ \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_{iT}y_{g_{iT}}, \end{bmatrix},$$

and $\hat{\Psi}$ is $\Psi$ [equation (5.3.22)] with $\mathbf{v}_{it}$ replaced by $\hat{\mathbf{v}}_{it} = \hat{\Gamma}\mathbf{y}_{it} + \hat{B}\mathbf{x}_{it}$, where $\hat{\Gamma}$ and $\hat{B}$ are any consistent estimates of $\Gamma$ and $B$. The generalized 3SLS is asymptotically equivalent to the minimum-distance estimator.

Both the 3SLS and the generalized 3SLS are consistent. But just as in the limited-information case, if the variance–covariance matrix possesses an error-component structure, they are not fully efficient. To take advantage of the known structure of the covariance matrix, Baltagi (1981a) suggested the following error-component three-stage least-squares estimator (EC3SLS).

The $g\ell$th block of the covariance matrix $\Sigma$ is of the form (5.3.3). A key point that is evident from Appendix 3B is that the set of eigenvectors $C_1$, $C_2$, $C_3$, and $C_4$ of (5.3.3) is invariant with respect to changes in the parameters $\sigma^2_{\lambda_{g\ell}}$, $\sigma^2_{\alpha_{g\ell}}$, and $\sigma^2_{u_{g\ell}}$. Therefore, premultiplying (5.3.19) by $I_G \otimes C'_h$, we have[9]

$$\mathbf{y}^{(h)} = W^{(h)}\boldsymbol{\theta} + \mathbf{v}^{(h)}, \qquad h = 1, 2, 3, \tag{5.3.24}$$

where $\mathbf{y}^{(h)} = (I_G \otimes C'_h)\mathbf{y}$, $W^{(h)} = (I_G \otimes C'_h)W$, $\mathbf{v}^{(h)} = (I_G \otimes C'_h)\mathbf{v}$, with $E(\mathbf{v}^{(h)}\mathbf{v}^{(h)\prime}) = \Sigma^{(h)} \otimes I_{n(h)}$, where $\Sigma^{(h)} = (\sigma^2_{h_{g\ell}})$ for $h = 1$, 2, and 3. Because $W^{(h)}$ contains endogenous variables that are correlated with $\mathbf{v}^{(h)}$, we first premultiply (5.3.24) by $(I_G \otimes X^{(h)})'$ to purge the correlation between $W^{(h)}$ and $\mathbf{v}^{(h)}$. Then apply the GLS estimation procedure to the resulting systems of

[9] Again, we ignore $C_4 = \mathbf{e}_{NT}/\sqrt{NT}$ because we have assumed that there is an intercept for each equation and because $C'_h\mathbf{e}_{NT} = \mathbf{0}$ for $h = 1, 2, 3$.

equations to obtain

$$\hat{\boldsymbol{\theta}}_{\text{GLS}} = \left[\sum_{h=1}^{3}\{W^{(h)\prime}[(\Sigma^{(h)})^{-1} \otimes P_X(h)]W^{(h)}\}\right]^{-1} \cdot \left[\sum_{h=1}^{3}\{W^{(h)\prime}[(\Sigma^{(h)})^{-1} \otimes P_X(h)]\mathbf{y}^{(h)}\right]. \tag{5.3.25}$$

Usually we do not know $\Sigma^{(h)}$. Therefore, the following three-stage procedure is suggested:

1. Estimate the $\hat{\boldsymbol{\theta}}_{\text{g}}^{(h)}$ by 2SLS.
2. Use the residuals from the $h$th 2SLS estimate to estimate $\hat{\sigma}^2_{h_{gl}}$ [equation (5.3.18)].
3. Replace $\Sigma^{(h)}$ by the estimated covariance matrix. Estimate $\boldsymbol{\theta}$ by (5.3.25).

The resulting estimator is called the EC3SLS estimator. It is a weighted combination of three 3SLS (within, between-groups, and between-time-periods) estimators of the structural parameters (Baltagi 1981a).

The EC3SLS estimator is asymptotically equivalent to the full-information maximum-likelihood estimator. In the case in which $\Sigma$ is block-diagonal, the EC3SLS reduces to the EC2SLS. But, contrary to the usual simultaneous-equations models, when the error terms have an error-component structure, the EC3SLS does not necessarily reduce to the EC2SLS, even if all the structural equations are just identified. For details, see Baltagi (1981a).

## 5.4 TRIANGULAR SYSTEM

Under the assumption that $\boldsymbol{\alpha}_i$, $\boldsymbol{\lambda}_t$ and $\mathbf{u}_{it}$ are independent of $\mathbf{x}_{is}$, the model discussed earlier assumes that residuals of different equations in a multiequation model have an unrestricted variance-component structure. Under this assumption, the panel data improve the precision of the estimates only by providing a large number of sample observations. They do not offer additional opportunities that are not standard. However, quite often the residual correlations may simply be due to one or two common omitted or unobservable variables (Chamberlain 1976, 1977a, 1977b; Chamberlain and Griliches 1975; Goldberger 1972; Zellner 1970). For instance, in the estimation of income and schooling relations or individual-firm production and factor-demand relations, it is sometimes postulated that the biases in different equations are caused by a common left-out "ability" or "managerial-differences" variable. When panel data are used, this common omitted variable is again assumed to have a within- and between-group structure. The combination of this factor-analytic structure with error-components formulations puts restrictions on the residual covariance matrix that can be used to identify an otherwise unidentified model and

improve the efficiency of the estimates. Because a widely used structure for longitudinal microdata is the triangular structure, and because its connection with the general simultaneous-equations model in which the residuals have a factor-analytic structure holds in general, in this section we focus on the triangular structure to illustrate how such information can be used to identify and estimate a model.

### 5.4.1 Identification

A convenient way to model correlations across equations, as well as the correlation of a given individual at different times (or different members of a group), is to use latent variables to connect the residuals. Let $y_{g_{it}}$ denote the value of the variable $y_g$ for the $i$th individual (or group) at time $t$ (or tth member). We can assume that

$$v_{g_{it}} = d_g h_{it} + u_{g_{it}}, \tag{5.4.1}$$

where the $u_g$ are uncorrelated across equations and across $i$ and $t$. The correlations across equations are all generated by the common omitted variable $h$, which is assumed to have a variance-component structure:

$$h_{it} = \alpha_i + \omega_{it}, \tag{5.4.2}$$

where $\alpha_i$ is invariant over $t$ but is independently identically distributed across $i$ (groups), with mean 0 and variance $\sigma_\alpha^2$, and $\omega_{it}$ is independently identically distributed across $i$ and $t$, with mean 0 and variance $\sigma_\omega^2$ and is uncorrelated with $\alpha_i$.

An example of the model with $\mathbf{\Gamma}$ lower-triangular and $\mathbf{v}$ of the form (5.4.1) is (Chamberlain 1977a, 1977b; Chamberlain and Griliches 1975; Griliches 1979)

$$\begin{aligned} y_{1_{it}} &= \boldsymbol{\beta}_1' \mathbf{x}_{it} + d_1 h_{it} + u_{1_{it}}, \\ y_{2_{it}} &= -\gamma_{21} y_{1_{it}} + \boldsymbol{\beta}_2' \mathbf{x}_{it} + d_2 h_{it} + u_{2_{it}}, \\ y_{3_{it}} &= -\gamma_{31} y_{1_{it}} - \gamma_{32} y_{2_{it}} + \boldsymbol{\beta}_3' \mathbf{x}_{it} + d_3 h_{it} + u_{3_{it}}, \end{aligned} \tag{5.4.3}$$

where $y_1$, $y_2$, and $y_3$ denote years of schooling, a late (postschool) test score, and earnings, respectively, and $\mathbf{x}_{it}$ are exogenous variables (which may differ from equation to equation via restrictions on $\boldsymbol{\beta}_g$). The unobservable $h$ can be interpreted as early "ability," and $u_2$ as measurement error in the test. The index $i$ indicates groups (or families), and $t$ indicates members in each group (or family).

Without the $h$ variables, or if $d_g = 0$, equation (5.4.3) would be only a simple recursive system that could be estimated by applying least-squares separately to each equation. The simultaneity problem arises when we admit the possibility that $d_g \neq 0$. In general, if there were enough exogenous variables in the first (schooling) equation that did not appear again in the other equations, the system

could be estimated using 2SLS or EC2SLS procedures. Unfortunately, in the income–schooling–ability model using sibling data [e.g., see the survey by Griliches 1979] there usually are not enough distinct $\mathbf{x}$'s to identify all the parameters. Thus, restrictions imposed on the variance–covariance matrix of the residuals will have to be used.

Given that $h$ is unobservable, we have an indeterminate scale

$$d_g^2\left(\sigma_\alpha^2+\sigma_\omega^2\right)=cd_g^2\left(\frac{1}{c}\sigma_\alpha^2+\frac{1}{c}\sigma_\omega^2\right). \tag{5.4.4}$$

So we normalize $h$ by letting $\sigma_\alpha^2=1$. Then

$$E\mathbf{v}_{it}\mathbf{v}_{it}'=\left(1+\sigma_\omega^2\right)\mathbf{d}\mathbf{d}'+\operatorname{diag}\left(\sigma_1^2,\ldots,\sigma_G^2\right)=\Omega, \tag{5.4.5}$$

$$E\mathbf{v}_{it}\mathbf{v}_{is}'=\mathbf{d}\mathbf{d}'=\Omega_w \qquad \text{if } t\neq s, \tag{5.4.6}$$

$$E\mathbf{v}_{it}\mathbf{v}_{js}'=\mathbf{0} \qquad \text{if } i\neq j, \tag{5.4.7}$$

where $\mathbf{d}=(d_1,\ldots,d_G)$, and $\operatorname{diag}(\sigma_1^2,\ldots,\sigma_G^2)$ denotes a $G\times G$ diagonal matrix with $\sigma_1^2,\sigma_2^2,\ldots,\sigma_G^2$ on the diagonal.

Under the assumption that $\alpha_i$, $\omega_{it}$, and $u_{g_{it}}$ are normally distributed, or if we limit our attention to second-order moments, all the information with regard to the distribution of $\mathbf{y}$ is contained in

$$C_{y_{tt}}=\mathbf{\Gamma}^{-1}\mathbf{B}C_{x_{tt}}\mathbf{B}'\mathbf{\Gamma}'^{-1}+\mathbf{\Gamma}^{-1}\Omega\mathbf{\Gamma}'^{-1}, \tag{5.4.8}$$

$$C_{y_{ts}}=\mathbf{\Gamma}^{-1}\mathbf{B}C_{x_{ts}}\mathbf{B}'\Gamma'^{-1}+\mathbf{\Gamma}^{-1}\Omega_w\mathbf{\Gamma}'^{-1}, \qquad t\neq s, \tag{5.4.9}$$

$$C_{yx_{ts}}=-\mathbf{\Gamma}^{-1}\mathbf{B}C_{x_{ts}}, \tag{5.4.10}$$

where $C_{y_{ts}}=E\mathbf{y}_{it}\mathbf{y}_{is}'$, $C_{yx_{ts}}=E\mathbf{y}_{it}\mathbf{x}_{is}'$, and $C_{x_{ts}}=E\mathbf{x}_{it}\mathbf{x}_{is}'$.

Stack the coefficient matrices $\Gamma$ and $\mathbf{B}$ into a $1\times G(G+K)$ vector $\boldsymbol{\theta}'=(\gamma_1',\ldots,\gamma_G',\beta_1',\ldots,\beta_G')$. Suppose $\boldsymbol{\theta}$ is subject to $M$ a priori constraints:

$$\Phi(\boldsymbol{\theta})=\boldsymbol{\phi}, \tag{5.4.11}$$

where $\boldsymbol{\phi}$ is an $M\times 1$ vector of constants. Then a necessary and sufficient condition for local identification of $\mathbf{\Gamma}$, $\mathbf{B}$, $\mathbf{d}$, $\sigma_\omega^2$, and $\sigma_1^2,\ldots,\sigma_G^2$ is that the rank of the Jacobian formed by taking partial derivatives of (5.4.8)–(5.4.11) with respect to the unknowns is equal to $G(G+K)+2G+1$ (e.g., Hsiao 1983).

Suppose there is no restriction on the matrix $\mathbf{B}$. The $GK$ equations (5.4.10) can be used to identify $\mathbf{B}$ provided that $\mathbf{\Gamma}$ is identifiable. Hence, we can concentrate on

$$\mathbf{\Gamma}\left(C_{y_{tt}}-C_{yx_{tt}}C_{x_{tt}}^{-1}C_{yx_{tt}}'\right)\mathbf{\Gamma}'=\Omega, \tag{5.4.12}$$

$$\mathbf{\Gamma}\left(C_{y_{ts}}-C_{yx_{ts}}C_{x_{ts}}^{-1}C_{yx_{ts}}'\right)\mathbf{\Gamma}'=\Omega_w, \qquad t\neq s, \tag{5.4.13}$$

We note that $\Omega$ is symmetric, and we have $G(G+1)/2$ independent equations from (5.4.12). But $\Omega_w$ is of rank 1; therefore, we can derive only $G$ independent equations from (5.4.13). Suppose $\mathbf{\Gamma}$ is lower-triangular and the diagonal elements of $\mathbf{\Gamma}$ are normalized to be unity; there are $G(G-1)/2$ unknowns in $\mathbf{\Gamma}$, and $2G+1$ unknowns of $(d_1, \ldots, d_G)$, $(\sigma_1^2, \ldots, \sigma_G^2)$, and $\sigma_\omega^2$. We have one less equation than the number of unknowns. In order for the Jacobian matrix formed by (5.4.12), (5.4.13), and a priori restrictions to be nonsingular, we need at least one additional a priori restriction. Thus, for the system

$$\mathbf{\Gamma} y_{it} + \mathbf{B}\mathbf{x}_{it} = \mathbf{v}_{it}, \tag{5.4.14}$$

where $\mathbf{\Gamma}$ is lower-triangular, $\mathbf{B}$ is unrestricted, and $\mathbf{v}_{it}$ satisfies (5.4.1) and (5.4.2), a necessary condition for the identification under exclusion restrictions is that at least one $\gamma_{g\ell} = 0$ for $g > \ell$. (For details, see Chamberlain 1976 or Hsiao 1983.)

### 5.4.2 Estimation

We have discussed how the restrictions in the variance–covariance matrix can help identify the model. We now turn to the issues of estimation. Two methods are discussed: the purged-instrumental-variable method (Chamberlain 1977a) and the maximum-likelihood method (Chamberlain and Griliches 1975). The latter method is efficient, but computationally complicated. The former method is inefficient, but it is simple and consistent. It also helps to clarify the previous results on the sources of identification.

For simplicity, we assume that there is no restriction on the coefficients of exogenous variables. Under this assumption we can further ignore the existence of exogenous variables without loss of generality, because there are no excluded exogenous variables that can legitimately be used as instruments for the endogenous variables appearing in the equation. The instruments have to come from the group structure of the model. We illustrate this point by considering the following triangular system:

$$\begin{aligned}
y_{1_{it}} &= && + v_{1_{it}}, \\
y_{2_{it}} &= \gamma_{21} y_{1_{it}} && + v_{2_{it}}, \\
&\vdots \\
y_{G_{it}} &= \gamma_{G1} y_{1_{it}} + \cdots + \gamma_{G,G-1} y_{G-1_{it}} && + v_{G_{it}},
\end{aligned} \tag{5.4.15}$$

where $v_{g_{it}}$ satisfy (5.4.1) and (5.4.2). We assume one additional $\gamma_{\ell k} = 0$ for some $\ell$ and $k$, $\ell > k$, for identification.

The reduced form of (5.4.15) is

$$y_{g_{it}} = a_g h_{it} + \epsilon_{g_{it}}, \qquad g = 1, \ldots, G, \tag{5.4.16}$$

where

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_G \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 + \gamma_{21} d_1 \\ d_3 + \gamma_{31} d_1 + \gamma_{32}(d_2 + \gamma_{21} d_1) \\ \vdots \end{bmatrix}, \tag{5.4.17}$$

$$\boldsymbol{\epsilon}_{it} = \begin{bmatrix} \epsilon_{1_{it}} \\ \epsilon_{2_{it}} \\ \epsilon_{3_{it}} \\ \vdots \\ \epsilon_{g_{it}} \\ \vdots \end{bmatrix} = \begin{bmatrix} u_{1_{it}} \\ u_{2_{it}} + \gamma_{21} u_{1_{it}} \\ u_{3_{it}} + \gamma_{31} u_{1_{it}} + \gamma_{32}(u_{2_{it}} + \gamma_{21} u_{1_{it}}) \\ \vdots \\ u_{g_{it}} + \sum_{k=1}^{g-1} \gamma_{gk}^* u_{k_{it}} \\ \vdots \end{bmatrix}, \tag{5.4.18}$$

where $\gamma_{gk}^* = \gamma_{gk} + \sum_{i=k+1}^{g-1} \gamma_{gi} \gamma_{ik}^*$ if $g > 1$ and $k + 1 < g$, and $\gamma_{gk}^* = \gamma_{gk}$ if $k + 1 = g$.

### *5.4.2.1 Instrumental-Variable Method*

The trick of the purged instrumental-variable (IV) method is to leave $h$ in the residual and construct instruments that are uncorrelated with $h$. Before going to the general formula, we use several simple examples to show where the instruments come from.

Consider the case that $G = 3$. Suppose $\gamma_{21} = \gamma_{31} = 0$. Using $y_1$ as a proxy for $h$ in the $y_3$ equation, we have

$$y_{3_{it}} = \gamma_{32} y_{2_{it}} + \frac{d_3}{d_1} y_{1_{it}} + u_{3_{it}} - \frac{d_3}{d_1} u_{1_{it}}. \tag{5.4.19}$$

If $T \geq 2$ then $y_{1_{is}}$, $s \neq t$, is a legitimate instrument for $y_{1_{it}}$, because it is uncorrelated with $u_{3_{it}} - (d_3/d_1)u_{1_{it}}$ but it is correlated with $y_{1_{it}}$ provided that $d_1 \sigma_\alpha^2 \neq 0$. Therefore, we can use $(y_{2_{it}}, y_{1_{is}})$ as instruments to estimate (5.4.19).

Next, suppose that only $\gamma_{32} = 0$. The reduced form of the model becomes

$$\begin{aligned} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} &= \begin{bmatrix} d_1 \\ d_2 + \gamma_{21} d_1 \\ d_3 + \gamma_{31} d_1 \end{bmatrix} h_{it} + \begin{bmatrix} u_{1_{it}} \\ u_{2_{it}} + \gamma_{21} u_{1_{it}} \\ u_{3_{it}} + \gamma_{31} u_{1_{it}} \end{bmatrix} \\ &= \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} h_{it} + \begin{bmatrix} \epsilon_{1_{it}} \\ \epsilon_{2_{it}} \\ \epsilon_{3_{it}} \end{bmatrix}. \end{aligned} \tag{5.4.20}$$

In this case, the construction of valid instruments is more complicated. It requires two stages. The first stage is to use $y_1$ as a proxy for $h$ in the

reduced-form equation for $y_2$:

$$y_{2_{it}} = \frac{a_2}{a_1} y_{1_{it}} + \epsilon_{2_{it}} - \frac{a_2}{a_1}\epsilon_{1_{it}}. \tag{5.4.21}$$

Equation (5.4.21) can be estimated by using $y_{1_{is}}$, $s \neq t$, as an instrument for $y_{1_{it}}$, provided that $d_1\sigma_\alpha^2 \neq 0$. Then form the residual, thereby purging $y_2$ of its dependence on $h$:

$$z_{2_{it}} = y_{2_{it}} - \frac{a_2}{a_1} y_{1_{it}} = \epsilon_{2_{it}} - \frac{a_2}{a_1}\epsilon_{1_{it}}. \tag{5.4.22}$$

The second stage is to use $z_2$ as an instrument for $y_1$ in the structural equation $y_3$:

$$y_{3_{it}} = \gamma_{31} y_{1_{it}} + d_3 h_{it} + u_{3_{it}}. \tag{5.4.23}$$

The variable $z_2$ is an appropriate IV because it is uncorrelated with $h$ and $u_3$, but it is correlated with $y_1$, provided $d_2\sigma_1^2 \neq 0$. (If $d_2 = 0$, then $z_2 = y_2 - \gamma_{21} y_1 = u_2$. It is no longer correlated with $y_1$.) Therefore, we require that $h$ appear directly in the $y_2$ equation and that $y_1$ not be proportional to $h$ – otherwise we could never separate the effects of $y_1$ and $h$.

In order to identify the $y_2$ equation

$$y_{2_{it}} = \gamma_{21} y_{1_{it}} + d_2 h_{it} + u_{2_{it}}, \tag{5.4.24}$$

we can interchange the reduced-form $y_2$ and $y_3$ equations and repeat the two stages. With $\gamma_{21}$ and $\gamma_{31}$ identified, in the third stage we form the residuals

$$\begin{aligned} v_{2_{it}} &= y_{2_{it}} - \gamma_{21} y_{1_{it}} = d_2 h_{it} + u_{2_{it}}, \\ v_{3_{it}} &= y_{3_{it}} - \gamma_{31} y_{1_{it}} = d_3 h_{it} + u_{3_{it}}. \end{aligned} \tag{5.4.25}$$

Then use $y_1$ as a proxy for $h$:

$$\begin{aligned} v_{2_{it}} &= \frac{d_2}{d_1} y_{1_{it}} + u_{2_{it}} - \frac{d_2}{d_1} u_{1_{it}}, \\ v_{3_{it}} &= \frac{d_3}{d_1} y_{1_{it}} + u_{3_{it}} - \frac{d_3}{d_1} u_{1_{it}}. \end{aligned} \tag{5.4.26}$$

Now $d_2/d_1$ and $d_3/d_1$ can be identified by a third application of instrumental variables, using $y_{1_{is}}$, $s \neq t$, as an instrument for $y_{1_{it}}$. (Note that only the ratio of the $d$'s is identified, because of the indeterminate scale of the latent variable.)

Now come back to the construction of IVs for the general system (5.4.15)–(5.4.18). We assume that $T \geq 2$. The instruments are constructed over several stages. At the first stage, let $y_1$ be a proxy for $h$. Then the reduced-form equation for $y_g$ becomes

$$y_{g_{it}} = \frac{a_g}{a_1} y_{1_{it}} + \epsilon_{g_{it}} - \frac{a_g}{a_1}\epsilon_{1_{it}}, \qquad g = 2, \ldots, \ell - 1. \tag{5.4.27}$$

If $T \geq 2$, $a_g/a_1$ can be consistently estimated by using different members in the same group (e.g., $y_{1_{is}}$ and $y_{1_{it}}$, $t \neq s$) as instruments for the $y_g$ equation

(5.4.27) when $d_1\sigma_\alpha^2 \neq 0$. Once $a_g/a_1$ is consistently estimated, we form the residual

$$z_{g_{it}} = y_{g_{it}} - \frac{a_g}{a_1} y_{1_{it}} = \epsilon_{g_{it}} - \frac{a_g}{a_1}\epsilon_{1_{it}}, \quad g = 2, \ldots, \ell - 1. \tag{5.4.28}$$

The $z_g$ are uncorrelated with $h$. They are valid instruments for $y_g$ provided $d_g\sigma_1^2 \neq 0$. There are $\ell - 2$ IVs for the $\ell - 2$ variables that remain on the right-hand side of the $\ell$th structural equation after $y_k$ has been excluded.

To estimate the equations that follow $y_\ell$, we form the transformed variables

$$\begin{aligned} y^*_{2_{it}} &= y_{2_{it}} - \gamma_{21} y_{1_{it}}, \\ y^*_{3_{it}} &= y_{3_{it}} - \gamma_{31} y_{1_{it}} - \gamma_{32} y_{2_{it}}, \\ &\vdots \\ y^*_{\ell_{it}} &= y_{\ell_{it}} - \gamma_{\ell 1} y_{1_{it}} - \cdots - \gamma_{\ell,\ell-1} y_{\ell-1_{it}}, \end{aligned} \tag{5.4.29}$$

and rewrite the $y_{\ell+1}$ equation as

$$\begin{aligned} y_{\ell+1_{it}} = {} & \gamma^*_{\ell+1,1} y_{1_{it}} + \gamma^*_{\ell+1,2} y^*_{2_{it}} + \cdots + \gamma^*_{\ell+1,\ell-1} y^*_{\ell-1_{it}} + \gamma_{\ell+1,\ell} y^*_{\ell_{it}} \\ & + d_{\ell+1} h_{it} + u_{\ell+1_{it}}, \end{aligned} \tag{5.4.30}$$

where $\gamma^*_{\ell+1,j} = \gamma_{\ell+1,j} + \sum_{m=j+1}^{\ell} \gamma_{\ell+1,m}\gamma^*_{mj}$ for $j < \ell$. Using $y_1$ as a proxy for $h$, we have

$$\begin{aligned} y_{\ell+1_{it}} = {} & \gamma^*_{\ell+1,2} y^*_{2_{it}} + \cdots + \gamma_{\ell+1,\ell} y^*_{\ell_{it}} \\ & + \left(\gamma^*_{\ell+1,1} + \frac{d_{\ell+1}}{d_1}\right) y_{1_{it}} + u_{\ell+1_{it}} - \frac{d_{\ell+1}}{d_1} u_{1_{it}}, \end{aligned} \tag{5.4.31}$$

Because $u_1$ is uncorrelated with $y^*_g$ for $2 \leq g \leq \ell$, we can use $y^*_{g_{it}}$ together with $y_{1_{is}}$, $s \neq t$ as instruments to identify $\gamma_{\ell+1,j}$. Once $\gamma_{\ell+1,j}$ are identified, we can form $y^*_{\ell+1} = y_{\ell+1} - \gamma_{\ell+1,1} y_1 - \cdots - \gamma_{\ell+1,\ell} y_\ell$ and proceed in a similar fashion to identify the $y_{\ell+2}$ equation, and so on.

Once all the $\gamma$ are identified, we can form the estimated residuals, $\hat{\mathbf{v}}_{it}$. From $\hat{\mathbf{v}}_{it}$ we can estimate $d_g/d_1$ by the same procedure as (5.4.26). Or we can form the matrix $\hat{\Omega}$ of variance–covariances of the residuals, and the matrix $\hat{\bar{\Omega}}$ of variance–covariances of averaged residuals $(1/T)\sum_{t=1}^{T} \hat{\mathbf{v}}_{it}$, then solve for $\mathbf{d}$, $(\sigma_1^2, \ldots, \sigma_G^2)$, and $\sigma_\omega^2$ from the relations

$$\hat{\Omega} = \left(1 + \sigma_\omega^2\right)\mathbf{d}\mathbf{d}' + \operatorname{diag}\left(\sigma_1^2, \ldots, \sigma_G^2\right), \tag{5.4.32}$$

$$\hat{\bar{\Omega}} = \left(1 + \sigma_\omega^2\right)\mathbf{d}\mathbf{d}' + \frac{1}{T}\operatorname{diag}\left(\sigma_1^2, \ldots, \sigma_G^2\right). \tag{5.4.33}$$

The purged IV estimator is consistent. It also will often indicate quickly if a new model is identified. For instance, to see the necessity of having at least

one more $\gamma_{g\ell} = 0$ for $g > \ell$ to identify the foregoing system, we can check if the instruments formed by the foregoing procedure satisfy the required rank condition. Consider the example where $G = 3$ and all $\gamma_{g\ell} \neq 0$ for $g > \ell$. In order to follow the strategy of allowing $h$ to remain in the residual, in the third equation we need IVs for $y_1$ and $y_2$ that are uncorrelated with $h$. As indicated earlier, we can purge $y_2$ of its dependence on $h$ by forming $z_2 = y_2 - (a_2/a_1)y_1$. A similar procedure can be applied to $y_1$. We use $y_2$ as a proxy for $h$, with $y_{2_{is}}$ as an IV for $y_{2_{it}}$. Then form the residual $z_1 = y_1 - (a_1/a_2)y_2$. Again $z_1$ is uncorrelated with $h$ and $u_3$. But $z_1 = -(a_1/a_2)z_2$, and so an attempt to use both $z_2$ and $z_1$ as IVs fails to meet the rank condition.

### *5.4.2.2 Maximum-Likelihood Method*

Although the purged IV method is simple to use, it is likely to be inefficient, because the correlations between the endogenous variables and the purged IVs will probably be small. Also, the restriction that (5.4.6) is of rank 1 is not being utilized. To obtain efficient estimates of the unknown parameters, it is necessary to estimate the covariance matrices simultaneously with the equation coefficients. Under the normality assumptions for $\alpha_i$, $\omega_{it}$ and $u_{it}$, we can obtain efficient estimates of (5.4.15) by maximizing the log likelihood function

$$\log L = -\frac{N}{2}\log|V| - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{y}'_{1i}, \mathbf{y}'_{2i}, \ldots, \mathbf{y}'_{Gi})V^{-1}(\mathbf{y}'_{1i}, \ldots, \mathbf{y}'_{Gi})', \tag{5.4.34}$$

where

$$\begin{aligned} \underset{T\times 1}{\mathbf{y}_{g_i}} &= (y_{g_{i1}}, \ldots, y_{g_{iT}})', \qquad g = 1, \ldots, G, \\ \underset{GT\times GT}{V} &= \Lambda \otimes I_T + \mathbf{a}\mathbf{a}' \otimes \mathbf{e}_T\mathbf{e}'_T, \\ \underset{G\times G}{\Lambda} &= E(\boldsymbol{\epsilon}_{it}\boldsymbol{\epsilon}'_{it}) + \sigma^2_\omega \mathbf{a}\mathbf{a}'. \end{aligned} \tag{5.4.35}$$

Using the relations[10]

$$V^{-1} = \Lambda^{-1} \otimes I_T - \mathbf{c}\mathbf{c}' \otimes \mathbf{e}_T\mathbf{e}'_T, \tag{5.4.36}$$

$$|V| = |\Lambda|^T |1 - T\mathbf{c}'\Lambda\mathbf{c}|^{-1}, \tag{5.4.37}$$

[10] For the derivations of (5.4.36) and (5.4.37), see Appendix 5.

we can simplify the log likelihood function as[11]

$$\log L = -\frac{NT}{2}\log|\Lambda| + \frac{N}{2}\log(1 - T\mathbf{c}'\Lambda\mathbf{c}) - \frac{NT}{2}\text{tr}(\Lambda^{-1}R) + \frac{NT^2}{2}\mathbf{c}'\bar{R}\mathbf{c}, \tag{5.4.38}$$

where $\mathbf{c}$ is a $G \times 1$ vector proportional to $\Lambda^{-1}\mathbf{a}$, $R$ is the matrix of the sums of the squares and cross-products of the residuals divided by $NT$, and $\bar{R}$ is the matrix of sums of squares and cross-products of the averaged residuals (over $t$ for $i$) divided by $N$. In other words, we simplify the log likelihood function (5.4.34) by reparameterizing it in terms of $\mathbf{c}$ and $\Lambda$.

Taking partial derivatives of (5.4.38), we obtain the first-order conditions[12]

$$\frac{\partial \log L}{\partial \Lambda^{-1}} = \frac{NT}{2}\Lambda + \frac{NT}{2}\frac{1}{(1 - T\mathbf{c}'\Lambda\mathbf{c})}\Lambda\mathbf{c}\mathbf{c}'\Lambda - \frac{NT}{2}R = \mathbf{0}, \tag{5.4.39}$$

$$\frac{\partial \log L}{\partial \mathbf{c}} = -\frac{NT}{1 - T\mathbf{c}'\Lambda\mathbf{c}}\Lambda\mathbf{c} + NT^2\bar{R}\mathbf{c} = \mathbf{0}. \tag{5.4.40}$$

Postmultiplying (5.4.39) by $\mathbf{c}$ and regrouping the terms, we have

$$\Lambda\mathbf{c} = \frac{1 - T\mathbf{c}'\Lambda\mathbf{c}}{1 - (T-1)\mathbf{c}'\Lambda\mathbf{c}}R\mathbf{c}. \tag{5.4.41}$$

Combining (5.4.40) and (5.4.41), we obtain

$$\left(\bar{R} - \frac{1}{T[1 - (T-1)\mathbf{c}'\Lambda\mathbf{c}]}R\right)\mathbf{c} = \mathbf{0}. \tag{5.4.42}$$

Hence, the MLE of $\mathbf{c}$ is a characteristic vector corresponding to a root of

$$|\bar{R} - \lambda R| = 0. \tag{5.4.43}$$

The determinate equation (5.4.43) has $G$ roots. To find which root to use, substitute (5.4.39) and (5.4.40) into (5.4.38):

$$\begin{aligned}\log L &= -\frac{NT}{2}\log|\Lambda| + \frac{N}{2}\log(1 - T\mathbf{c}'\Lambda\mathbf{c}) \\ &\quad - \frac{NT}{2}(G + T\ \text{tr}\,\mathbf{c}'\bar{R}\mathbf{c}) + \frac{NT^2}{2}\text{tr}(\mathbf{c}'\bar{R}\mathbf{c}) \\ &= -\frac{NT}{2}\log|\Lambda| + \frac{N}{2}\log(1 - T\mathbf{c}'\Lambda\mathbf{c}) - \frac{NTG}{2}.\end{aligned} \tag{5.4.44}$$

Let the $G$ characteristic vectors corresponding to the $G$ roots of (5.4.43) be denoted as $\mathbf{c}_1(=\mathbf{c}), \mathbf{c}_2, \ldots, \mathbf{c}_G$. These characteristic vectors are determined

[11] From $V \cdot V^{-1} = I_{GT}$, we have $-\Lambda\mathbf{c}\mathbf{c}' - T\mathbf{a}\mathbf{a}'\mathbf{c}\mathbf{c}' + \mathbf{a}\mathbf{a}'\Lambda^{-1} = \mathbf{0}$. Premultiplying this equation by $\mathbf{c}'$ we obtain $(b_1 + Tb_2^2)\mathbf{c}' = b_2\mathbf{a}'\Lambda^{-1}$, where $b_1 = \mathbf{c}'\Lambda\mathbf{c}$ and $b_2 = \mathbf{c}'\mathbf{a}$. In Appendix 5 we give the values of $b_1$ and $b_2$ explicitly in terms of the eigenvalue of $[\mathbf{a}\mathbf{a}' - \lambda\Lambda\ | = 0$.

[12] We make use of the formula $\partial \log |\ \Lambda\ |\ /\partial\Lambda^{-1} = -\Lambda'$ and $\partial(\mathbf{c}'\Lambda\mathbf{c})/\partial\Lambda^{-1} = -\Lambda\mathbf{c}\mathbf{c}'\Lambda$ (Theil 1971, pp. 32–33).

only up to a scalar. Choose the normalization $\mathbf{c}_g^{*\prime} R \mathbf{c}_g^* = 1$, $g = 1, \ldots, G$, where $\mathbf{c}_g^* = (\mathbf{c}_g' R \mathbf{c}_g)^{-1/2} \mathbf{c}_g$. Let $C^* = [\mathbf{c}_1^*, \ldots, \mathbf{c}_G^*]$; then $C^{*\prime} R C^* = I_G$. From (5.4.39) and (5.4.41) we have

$$
\begin{aligned}
C^{*\prime} \Lambda C^* &= C^{*\prime} R C^* - \frac{1 - T\mathbf{c}'\Lambda\mathbf{c}}{[1 - (T-1)\mathbf{c}'\Lambda\mathbf{c}]^2} C^{*\prime} R \mathbf{c}\mathbf{c}' R C^* \\
&= I_G - \frac{1 - T\mathbf{c}'\Lambda\mathbf{c}}{[1 - (T-1)\mathbf{c}'\Lambda\mathbf{c}]^2} \\
&\quad \cdot \begin{bmatrix} (\mathbf{c}' R \mathbf{c})^{1/2} \\ 0 \\ \vdots \\ 0 \end{bmatrix} [(\mathbf{c}' R \mathbf{c})^{1/2} \; 0 \; \cdots \; 0]. \qquad (5.4.45)
\end{aligned}
$$

Equation (5.4.41) implies that $(\mathbf{c}' R \mathbf{c}) = \{[1 - (T-1)\mathbf{c}'\Lambda\mathbf{c}]/[1 - T\mathbf{c}'\Lambda\mathbf{c}]\}\mathbf{c}'\Lambda\mathbf{c}$. Therefore, the determinant of (5.4.45) is $\{[1 - T\mathbf{c}'\Lambda\mathbf{c}]/[1 - (T-1)\mathbf{c}'\Lambda\mathbf{c}]\}$. Using $C^{*\prime-1} C^{*-1} = R$, we have $|\Lambda| = \{[1 - T\mathbf{c}'\Lambda\mathbf{c}]/[1 - (T-1)\mathbf{c}'\Lambda\mathbf{c}]\}|R|$. Substituting this into (5.4.44), the log likelihood function becomes

$$
\begin{aligned}
\log L = &-\frac{NT}{2}\{\log|R| + \log(1 - T\mathbf{c}'\Lambda\mathbf{c}) \\
&- \log[1 - (T-1)\mathbf{c}'\Lambda\mathbf{c}]\} \\
&+ \frac{N}{2}\log[1 - T\mathbf{c}'\Lambda\mathbf{c}] - \frac{NTG}{2}, \qquad (5.4.46)
\end{aligned}
$$

which is positively related to $\mathbf{c}'\Lambda\mathbf{c}$ within the admissible range $(0, 1/T)$.[13] So the MLE of $\mathbf{c}$ is the characteristic vector corresponding to the largest root of (5.4.43). Once $\mathbf{c}$ is obtained, from Appendix 5A and (5.4.39) and (5.4.40) we can estimate $\mathbf{a}$ and $\Lambda$ by

$$
\mathbf{a}' = T(1 + T^2 \mathbf{c}' \bar{R} \mathbf{c})^{-1/2} \mathbf{c}' \bar{R}, \qquad (5.4.47)
$$

and

$$
\Lambda = R - \mathbf{a}\mathbf{a}'. \qquad (5.4.48)
$$

Knowing $\mathbf{a}$ and $\Lambda$, we can solve for the coefficients of the joint dependent variables $\Gamma$.

When exogenous variables also appear in the equation, and with no restrictions on the coefficients of exogenous variables, we need only replace the exponential term of the likelihood function (5.4.34),

$$
-\frac{1}{2}\sum_{i=1}^{N} (\mathbf{y}_{1i}', \ldots, \mathbf{y}_{Gi}') V^{-1} (\mathbf{y}_{1i}', \ldots, \mathbf{y}_{Gi}')',
$$

[13] See Appendix 5, equation (5A.7), in which $\psi_1$ is positive.

with

$$-\frac{1}{2}\sum_{i=1}^{N}(\mathbf{y}'_{1i} - \boldsymbol{\pi}'_1\mathbf{X}'_i, \ldots, \mathbf{y}'_{Gi} - \boldsymbol{\pi}'_G X'_i)$$

$$\cdot\, V^{-1}(\mathbf{y}'_{1i} - \boldsymbol{\pi}'_1 X'_i, \ldots, \mathbf{y}'_{Gi} - \boldsymbol{\pi}'_G X'_i)'.$$

The MLEs of $\mathbf{c}$, $\mathbf{a}$, and $\Lambda$ remain the solutions of (5.4.43), (5.4.47), and (5.4.48). From knowledge of $\Lambda$ and $\mathbf{a}$ we can solve for $\Gamma$ and $\sigma_\omega^2$. The MLE of $\Pi$ conditional on $V$ is the GLS of $\Pi$. Knowing $\Pi$ and $\Gamma$, we can solve for $B = -\Gamma\Pi$.

Thus, Chamberlain and Griliches (1975) suggested the following iterative algorithm to solve for the MLE. Starting from the least-squares reduced-form estimates, we can form consistent estimates of $R$ and $\bar{R}$. Then estimate $\mathbf{c}$ by maximizing[14]

$$\frac{\mathbf{c}'\bar{R}\mathbf{c}}{\mathbf{c}'R\mathbf{c}}. \tag{5.4.49}$$

Once $\mathbf{c}$ is obtained, we solve for $\mathbf{a}$ and $\Lambda$ by (5.4.47) and (5.4.48). After obtaining $\Lambda$ and $\mathbf{a}$, the MLE of the reduced-form parameters is just the generalized least-squares estimate. With these estimated reduced-form coefficients, one can form new estimates of $R$ and $\bar{R}$ and continue the iteration until the solution converges. The structural-form parameters are then solved from the convergent reduced-form parameters.

### 5.4.3 An Example

Chamberlain and Griliches (1975) used the Gorseline (1932) data of the highest grade of schooling attained ($y_1$), the logarithm of the occupational (Duncan's SES) standing ($y_2$), and the logarithm of 1927 income ($y_3$) for 156 pairs of brothers from Indiana (U.S.) to fit a model of the type (5.4.1)–(5.4.3). Specifically, they let

$$\begin{aligned} y_{1_{it}} &= \boldsymbol{\beta}'_1\mathbf{x}_{it} + d_1 h_{it} + u_{1_{it}}, \\ y_{2_{it}} &= \gamma_{21} y_{1_{it}} + \boldsymbol{\beta}'_2\mathbf{x}_{it} + d_2 h_{it} + u_{2_{it}}, \\ y_{3_{it}} &= \gamma_{31} y_{1_{it}} + \boldsymbol{\beta}'_3\mathbf{x}_{it} + d_3 h_{it} + u_{3_{it}}. \end{aligned} \tag{5.4.50}$$

The set $X$ contains a constant, age, and age squared, with age squared appearing only in the income equation.

[14] Finding the largest root of (5.4.43) is equivalent to maximizing (5.4.49). If we normalize $\mathbf{c}'R\mathbf{c} = 1$, then to find the maximum of (5.4.49) we can use Lagrangian multipliers and maximize $\mathbf{c}'\bar{R}\mathbf{c} + \lambda(1 - \mathbf{c}'R\mathbf{c})$. Taking partial derivatives with respect to $\mathbf{c}$ gives $(\bar{R} - \lambda R)\mathbf{c} = \mathbf{0}$. Premultiplying by $\mathbf{c}'$, we have $\mathbf{c}'\bar{R}\mathbf{c} = \lambda$. Thus, the maximum of (5.4.49) is the largest root of $|\bar{R} - \lambda R| = 0$, and $\mathbf{c}$ is the characteristic vector corresponding to the largest root.

The reduced form of (5.4.50) is

$$\mathbf{y}_{it} = \Pi \mathbf{x}_{it} + \mathbf{a}h_{it} + \boldsymbol{\epsilon}_{it}, \tag{5.4.51}$$

where

$$\begin{aligned}
\boldsymbol{\Pi} &= \begin{bmatrix} \boldsymbol{\beta}_1' \\ \gamma_{21}\boldsymbol{\beta}_1' + \boldsymbol{\beta}_2' \\ \gamma_{31}\boldsymbol{\beta}_1' + \boldsymbol{\beta}_3' \end{bmatrix}, \\
\mathbf{a} &= \begin{bmatrix} d_1 \\ d_2 + \gamma_{21}d_1 \\ d_3 + \gamma_{31}d_1 \end{bmatrix}, \\
\boldsymbol{\epsilon}_{it} &= \begin{bmatrix} u_{1_{it}} \\ u_{2_{it}} + \gamma_{21}u_{1_{it}} \\ u_{3_{it}} + \gamma_{31}u_{1_{it}} \end{bmatrix}.
\end{aligned} \tag{5.4.52}$$

Therefore,

$$E\boldsymbol{\epsilon}_{it}\boldsymbol{\epsilon}_{it}' = \begin{bmatrix} \sigma_{u1}^2 & \gamma_{21}\sigma_{u1}^2 & \gamma_{31}\sigma_{u1}^2 \\ & \sigma_{u2}^2 + \gamma_{21}^2\sigma_{u1}^2 & \gamma_{21}\gamma_{31}\sigma_{u1}^2 \\ & & \sigma_{u3}^2 + \gamma_{31}^2\sigma_{u1}^2 \end{bmatrix}, \tag{5.4.53}$$

and

$$\Lambda = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ & \sigma_{22} & \sigma_{23} \\ & & \sigma_{33} \end{bmatrix} = E(\boldsymbol{\epsilon}_{it}\boldsymbol{\epsilon}_{it}') + \sigma_\omega^2 \mathbf{a}\mathbf{a}'. \tag{5.4.54}$$

We show that knowing $\mathbf{a}$ and $\Lambda$ identifies the structural coefficients of the joint dependent variables as follows: For a given value of $\sigma_\omega^2$, we can solve for

$$\sigma_{u1}^2 = \sigma_{11} - \sigma_\omega^2 a_1^2, \tag{5.4.55}$$

$$\gamma_{21} = \frac{\sigma_{12} - \sigma_\omega^2 a_1 a_2}{\sigma_{u1}^2}, \tag{5.4.56}$$

$$\gamma_{31} = \frac{\sigma_{13} - \sigma_\omega^2 a_1 a_3}{\sigma_{u1}^2}. \tag{5.4.57}$$

Equating

$$\gamma_{21}\gamma_{31} = \frac{\sigma_{23} - \sigma_\omega^2 a_2 a_3}{\sigma_{u1}^2} \tag{5.4.58}$$

with the product of (5.4.56) and (5.4.57), and making use of (5.4.55), we have

$$\sigma_\omega^2 = \frac{\sigma_{12}\sigma_{13} - \sigma_{11}\sigma_{23}}{\sigma_{12}a_1a_3 + \sigma_{13}a_1a_2 - \sigma_{11}a_2a_3 - \sigma_{23}a_1^2}. \tag{5.4.59}$$

Table 5.1. *Parameter estimates and their standard errors for the income-occupation-schooling model*

| Coefficients of the structural equations | Method | | |
|---|---|---|---|
| | Least-squares estimate | Covariance estimate | MLE |
| Schooling in the: | | | |
| Income equation | 0.082 | 0.080 | 0.088 |
| | (0.010)[a] | (0.011) | (0.009) |
| Occupation equation | 0.104 | 0.135 | 0.107 |
| | (0.010) | (0.015) | (0.010) |
| "Ability" in the: | | | |
| Income equation | | | 0.416 |
| | | | (0.038) |
| Occupation equation | | | 0.214 |
| | | | (0.046) |
| Schooling equation | | | −0.092 |
| | | | (0.178) |

[a] Standard errors in parentheses.
*Source:* Chamberlain and Griliches (1975, p. 429).

The problem then becomes one of estimating $\mathbf{a}$ and $\Lambda$. Table 5.1 presents the MLE of Chamberlain and Griliches (1975) for the coefficients of schooling and (unobservable) ability variables with $\sigma_\alpha^2$ normalized to equal 1. Their least-squares estimates ignore the familial information, and the covariance estimates in which each brother's characteristics (his income, occupation, schooling, and age) are measured around his own family's mean are also presented in Table 5.1.

The CV estimate of the coefficient-of-schooling variable in the income equation is smaller than the least-squares estimate. However, the simultaneous-equations model estimate of the coefficient for the ability variable is negative in the schooling equation. As discussed in Section 5.1, if schooling and ability are negatively correlated, the single-equation within-family estimate of the schooling coefficient could be less than the least-squares estimate (here 0.080 vs. 0.082). To attribute this decline to "ability" or "family background" is erroneous. In fact, when schooling and ability were treated symmetrically, the coefficient-of-schooling variable (0.088) became greater than the least-squares estimate 0.082.

## APPENDIX 5A:

Let

$$V = \Lambda \otimes I_T + \mathbf{a}\mathbf{a}' \otimes \mathbf{e}_T \mathbf{e}_T'. \tag{5A.1}$$

Because $\Lambda$ is positive definite and $\mathbf{a}\mathbf{a}'$ is positive semidefinite, there exists a $G \times G$ nonsingular matrix $F$ such that (Anderson 1985, p. 341)

$$F'\Lambda F = I_G \text{ and } F'\mathbf{a}\mathbf{a}'F = \begin{bmatrix} \psi_1 & & & \mathbf{0} \\ & 0 & & \\ & & \ddots & \\ \mathbf{0} & & & 0 \end{bmatrix},$$

where $\psi_1$ is the root of

$$\left|\mathbf{a}\mathbf{a}' - \lambda\Lambda\right| = 0. \tag{5A.2}$$

Next, choose a $T \times T$ orthogonal matrix $E$, with the first column of $E$ being the vector $(1/\sqrt{T})\mathbf{e}_T$.Then

$$E'E = I_T \text{ and } E'\mathbf{e}_T\mathbf{e}_T'E = \begin{bmatrix} T & \mathbf{0}' \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{5A.3}$$

Now $F \otimes E$ can be used to diagonalize $V$,

$$(F \otimes E)'V(F \otimes E) = I_{GT} + \underset{G\times G}{\begin{bmatrix} \psi_1 & \mathbf{0}' \\ \mathbf{0} & \mathbf{0} \end{bmatrix}} \otimes \underset{T\times T}{\begin{bmatrix} T & \mathbf{0}' \\ \mathbf{0} & \mathbf{0} \end{bmatrix}}, \tag{5A.4}$$

and factor $V^{-1}$,

$$\begin{aligned} V^{-1} &= \Lambda^{-1} \otimes I_T - F' \begin{bmatrix} \frac{\psi_1}{1+T\psi_1} & \mathbf{0}' \\ \mathbf{0} & \mathbf{0} \end{bmatrix} F \otimes \mathbf{e}_T\mathbf{e}_T' \\ &= \Lambda^{-1} \otimes I_T - \mathbf{c}\mathbf{c}' \otimes \mathbf{e}_T\mathbf{e}_T', \end{aligned} \tag{5A.5}$$

where $\mathbf{c}' = [\psi_1/(1 + T\psi_1)]^{\frac{1}{2}}\,\mathbf{f}_1'$,and $\mathbf{f}_1$ is the first column of $F$.

The determinant of $V$ can be obtained from (5A.4):

$$|V| = |\Lambda|^T \cdot (1 + T\psi_1). \tag{5A.6}$$

This can be expressed in terms of $\mathbf{c}$ and $\Lambda$ by nothing that

$$\mathbf{c}'\Lambda\mathbf{c} = \frac{\psi_1}{1 + T\psi_1}. \tag{5A.7}$$

Thus, we have

$$1 - T\mathbf{c}'\Lambda\mathbf{c} = \frac{1}{1 + T\psi_1}, \tag{5A.8}$$

and

$$|V| = |\Lambda|^T \cdot (1 - T\mathbf{c}'\Lambda\mathbf{c})^{-1}. \tag{5A.9}$$

From $V \cdot V^{-1} = I_{GT}$ it is implied that

$$-\Lambda\mathbf{c}\mathbf{c}' + \mathbf{a}\mathbf{a}'\Lambda^{-1} - T\mathbf{a}\mathbf{a}'\mathbf{c}\mathbf{c}' = 0. \tag{5A.10}$$

Premultiplying (5A.10) by $\mathbf{c}'$, we obtain

$$\mathbf{a} = \frac{\mathbf{c}'\mathbf{a}}{\mathbf{c}'\Lambda\mathbf{c} + (\mathbf{c}'\mathbf{a})^2}\Lambda\mathbf{c}. \tag{5A.11}$$

Also, from $\mathbf{f}_1'\mathbf{a} = \psi^{\frac{1}{2}}$ and $\mathbf{a}$ proportional to $\mathbf{c}_1$ [equation (5A.11)] and hence $\mathbf{f}_1$, we have

$$\mathbf{a} = \frac{\psi^{\frac{1}{2}}}{\mathbf{f}_1'\mathbf{f}_1}\mathbf{f}_1 = \frac{1}{(1 + T\psi_1)^{\frac{1}{2}}(\mathbf{c}\mathbf{c}')}\mathbf{c} \tag{5A.12}$$

Premultiplying (5.4.40) by $\mathbf{c}'$, we obtain

$$\mathbf{c}'\bar{R}\mathbf{c} = \frac{\mathbf{c}'\Lambda\mathbf{c}}{T(1 - T\mathbf{c}'\Lambda\mathbf{c})} = \frac{1}{T}\psi_1. \tag{5A.13}$$

Combining (5.4.40) with (5A.8), (5A.12), and (5A.13), and using $\Lambda\mathbf{f}_1 = (1/\mathbf{f}_1'\mathbf{f}_1)\mathbf{f}_1$, we obtain

$$\begin{aligned}\bar{R}\mathbf{c} &= \frac{1}{T}(1 + T\psi_1)\Lambda\mathbf{c} \\ &= \frac{1}{T}(1 + T\psi_1)^{\frac{1}{2}}\mathbf{a} \\ &= \frac{1}{T}(1 + T^2\mathbf{c}'\bar{R}\mathbf{c})^{\frac{1}{2}}\mathbf{a}. \end{aligned} \tag{5A.14}$$

From (5.4.39) and (5A.12), we have

$$\begin{aligned}\Lambda &= R - \frac{1}{1 - T\mathbf{c}'\Lambda\mathbf{c}}\Lambda\mathbf{c}\mathbf{c}'\Lambda \\ &= R - \mathbf{a}\mathbf{a}' \end{aligned} \tag{5A.15}$$

CHAPTER 6

# Variable-Coefficient Models

## 6.1 INTRODUCTION

So far we have confined our discussion to models in which the effects of omitted variables are either individual-specific or time-specific or both. But there are cases in which there are changing economic structures or unobserved different socioeconomic and demographic background factors that imply that the response parameters of the included variables may be varying over time and/or may be different for different cross-sectional units. For instance, in farm production it is likely that variables not included in the specification could also impact the marginal productivity of fertilizer used such as soil characteristics (e.g., slope, soil fertility, water reserve, etc.) or climatic conditions. The same applies to empirical studies of economic growth. The per capita output growth rates are assumed to depend on two sets of variables over a common horizon. One set of variables consists of initial per capita output, savings, and population growth rates, variables that are suggested by the Solow growth model. The second set of variables consists of control variables that correspond to whatever additional determinants of growth a researcher wishes to examine (e.g., Durlauf 2001; Durlauf and Quah 1999). However, there is nothing in growth theory that would lead one to think that the marginal effect of a change in high school enrollment percentages on the per capita growth of the United States should be the same as the effect on a country in sub-Saharan Africa. Had all these factors been taken into account in the specification, a common slope coefficients model may seem reasonable. However, these variables could be unavailable or could be difficult to observe with precision. Moreover, a model is not a mirror; it is a simplification of the real world to capture the relationships among the essential variables. As a matter of fact, any parsimonious regression will necessarily leave out many factors that would, from the perspective of economic theory, be likely to affect the parameters of the included variables (e.g., Canova 1999; Durlauf and Johnson 1995). In these situations, varying parameter models appear to be more capable of capturing the unobserved heterogeneity than a model with only individual-specific and/or time-specific effects (variable-intercept models).

In Chapter 2 we reported a study (Kuh 1963) on investment expenditures of 60 small and middle-sized firms in capital-goods-producing industries

from 1935 to 1955, excluding the war years (1942–45). In a majority of the cases Kuh investigated, the common intercept and common slope coefficients for all firms, as well as the variable-intercept common-slope hypotheses, were rejected (Tables 2.3 and 2.4). Similar results were found by Swamy (1970), who used the annual data of 11 U.S. corporations from 1935 to 1954 to fit the Grunfeld (1958) investment functions. His preliminary test of variable-intercept–common slope coefficients against the variable-intercept and variable slope coefficients for the value of a firm's outstanding shares at the beginning of the year and its beginning-of-year capital stock yielded an $F$ value of 14.4521. That is well above the 5 percent value of an $F$ distribution with 27 and 187 degrees of freedom.[1]

When an investigator is interested mainly in the fundamental relationship between the outcome variable and a set of primary conditional variables, either for ease of analysis or because of the unavailability of the secondary conditional variables, it would seem reasonable to allow variations in parameters across cross-sectional units and/or over time as a means to take account of the interindividual and/or interperiod heterogeneity. A single-equation model in its most general form can be written as

$$y_{it} = \sum_{k=1}^{K} \beta_{kit} x_{kit} + u_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1. \ldots, T, \end{matrix} \tag{6.1.1}$$

where, in contrast to previous chapters, we no longer treat the intercept differently than other explanatory variables and let $x_{1it} = 1$. However, if all the coefficients are treated as fixed and different for different cross-sectional units in different time periods, there are $NKT$ parameters with only $NT$ observations. Obviously, there is no way we can obtain any meaningful estimates of $\beta_{kit}$. We are thus led to search for an approach that allows the coefficients of interest to differ, but provides some method of modeling the cross-sectional units as a group rather than individually.

One possibility would be to introduce dummy variables into the model that would indicate differences in the coefficients across individual units and/or over time, that is, to develop an approach similar to the least-squares dummy variable approach. In the case in which only cross-sectional differences are present, this approach is equivalent to postulating a separate regression for each cross-sectional unit[2]

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta}_i + u_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{matrix} \tag{6.1.2}$$

where $\boldsymbol{\beta}_i$ and $\mathbf{x}_{it}$ are $K \times 1$ vectors of parameters and explanatory variables.

[1] See Mehta, Narasimham, and Swamy (1978) for another example that using error-components formulation to account for heterogeneity does not always yield economically meaningful results.

[2] Alternatively, we can postulate a separate regression for each time period; so $y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta}_t + u_{it}$.

Alternatively, each regression coefficient can be viewed as a random variable with a probability distribution (e.g., Hurwicz 1950; Klein 1953; Theil and Mennes 1959; Zellner 1966). The random-coefficients specification reduces the number of parameters to be estimated substantially, while still allowing the coefficients to differ from unit to unit and/or from time to time. Depending on the type of assumption about the parameter variation, it can be further classified into one of two categories: stationary and nonstationary random-coefficient models.

Stationary random-coefficient models regard the coefficients as having constant means and variance–covariances. Namely, the $K \times 1$ vector of parameters $\boldsymbol{\beta}_{it}$ is specified as

$$\begin{aligned} \boldsymbol{\beta}_{it} &= \bar{\boldsymbol{\beta}} + \boldsymbol{\xi}_{it}, \quad i = 1, \ldots, N, \\ &\qquad\qquad t = 1, \ldots, T, \end{aligned} \tag{6.1.3}$$

where $\bar{\boldsymbol{\beta}}$ is a $K \times 1$ vector of constants, and $\boldsymbol{\xi}_{it}$ is a $K \times 1$ vector of stationary random variables with 0 means and constant variance–covariances. For this type of model we are interested in (1) estimating the mean coefficient vector $\bar{\boldsymbol{\beta}}$, (2) predicting each individual component $\boldsymbol{\xi}_{it}$, (3) estimating the dispersion of the individual-parameter vector, and (4) testing the hypothesis that the variances of $\boldsymbol{\xi}_{it}$ are 0.

The nonstationary random-coefficient models do not regard the coefficient vector as having constant mean or variance. Changes in coefficients from one observation to the next can be the result of the realization of a nonstationary stochastic process or can be a function of exogenous variables. In this case we are interested in (1) estimating the parameters characterizing the time-evolving process, (2) estimating the initial value and the history of parameter realizations, (3) predicting the future evolutions, and (4) testing the hypothesis of random variation.

Because of the computational complexities, variable-coefficient models have not gained as wide acceptance in empirical work as has the variable-intercept model. However, that does not mean that there is less need for taking account of parameter heterogeneity in pooling the data. In this chapter we survey some of the popular single-equation varying coefficients models. We first discuss models in which the variations of coefficients are independent of the variations of exogenous explanatory variables. Single-equation models with coefficients varying over individuals are discussed in Section 6.2. In Section 6.3 we discuss models with coefficients varying over individuals and time. Section 6.4 concerns models with time-evolving coefficients. Models with coefficients that are functions of other exogenous variables are discussed in Section 6.5. Section 6.6 proposes a mixed fixed and random coefficient model as a unifying framework to various approaches of controlling unobserved heterogeneity. Section 6.7 discusses issues of dynamic models. Section 6.8 provides two examples that use random coefficients model to pool heterogeneous individuals. General models in which the variation of coefficients are correlated with the explanatory

variables are discussed in Section 6.9. For random coefficients models with heteroscedasticity, see Bresson et al. (2011); with cross-correlated residuals (e.g., Bresson and Hsiao 2011); simultaneous-equations models with random coefficients, see Chow (1983), Kelejian (1977), and Raj and Ullah (1981). Further discussion of this subject can also be found in Amemiya (1983), Chow (1983), Hsiao and Pesaran (2008), Judge et al. (1980), and Raj and Ullah (1981).

## 6.2 COEFFICIENTS THAT VARY OVER CROSS-SECTIONAL UNITS

When regression coefficients are viewed as invariant over time, but varying from one unit to another, we can write the model as

$$\begin{aligned} y_{it} &= \sum_{k=1}^{K} \beta_{ki} x_{kit} + u_{it} \\ &= \sum_{k=1}^{K} (\bar{\beta}_k + \alpha_{ki}) x_{kit} + u_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{matrix} \end{aligned} \tag{6.2.1}$$

where $\bar{\boldsymbol{\beta}} = (\bar{\beta}_1, \ldots, \bar{\beta}_K)'$ can be viewed as the common-mean-coefficient vector and $\boldsymbol{\alpha}_i = (\alpha_{1i}, \ldots, \alpha_{Ki})'$ as the individual deviation from the common mean $\bar{\boldsymbol{\beta}}$. If individual observations are heterogeneous or the performance of individual units from the data base is of interest, then $\boldsymbol{\alpha}_i$ are treated as fixed constants. If conditional on $x_{kit}$, individual units can be viewed as random draws from a common population or the population characteristics are of interest, then $\alpha_{ki}$ are generally treated as random variables having 0 means and constant variances and covariances.

### 6.2.1 Fixed-Coefficient Model

#### *6.2.1.1 Complete Heterogeneity*

When $\boldsymbol{\beta}_i$ are treated as fixed and different constants, we can stack the $NT$ observations in the form of the Zellner (1962) seemingly unrelated regression model

$$\begin{aligned} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_N \end{bmatrix} &= \begin{bmatrix} X_1 & & & \mathbf{0} \\ & X_2 & & \\ & & \ddots & \\ \mathbf{0} & & & X_N \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_N \end{bmatrix} + \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_N \end{bmatrix}, \\ &= \tilde{X}\boldsymbol{\beta} + \mathbf{u} \end{aligned} \tag{6.2.2}$$

where $\mathbf{y}_i$ and $\mathbf{u}_i$ are $T \times 1$ vectors of $(y_{i1}, \ldots, y_{iT})'$ and $(u_{i1}, \ldots, u_{iT})'$; $X_i$ is the $T \times K$ matrix of the time-series observations of the $i$th individual's explanatory variables with the $t$th row equal to $\mathbf{x}_{it}'$; $\tilde{X}$ is $NT \times NK$ block

diagonal matrix with the $i$th block being $X_i$; and $\boldsymbol{\beta}$ is an $NK \times 1$ vector, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_N')'$. If the covariances between different cross-sectional units are not 0, $E\mathbf{u}_i\mathbf{u}_j' \neq \mathbf{0}$, the GLS estimator of $(\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_N')$ is more efficient than the single-equation estimator of $\boldsymbol{\beta}_i$ for each cross-sectional unit. If $X_i$ are identical for all $i$ or $E\mathbf{u}_i\mathbf{u}_i' = \sigma_i^2 I$ and $E\mathbf{u}_i\mathbf{u}_j' = \mathbf{0}$ for $i \neq j$, the generalized least-squares (GLS) estimator for $(\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_N')$ is the same as applying the least squares separately to the time-series observations of each cross-sectional unit.

### *6.2.1.2 Group Heterogeneity*

When $N$ is large, it is neither feasible nor desirable to let $\boldsymbol{\beta}_i$ be different for different $i$. An alternative to individual heterogeneity is to assume group heterogeneity in place of complete heterogeneity. In other words, the population is assumed to be composed of $G$ heterogeneous groups. Individuals belonging to a particular group all have the same response function (e.g., Lin and Ng 2012; Su, Shi, and Phillips 2013),

$$y_{it,g} = \mathbf{x}_{it}'\boldsymbol{\beta}_g + u_{it,g}, \quad \text{for } i \in \text{group } g. \tag{6.2.3}$$

If the grouping is known from some external consideration (e.g., Bester and Hansen 2012), estimation of (6.2.3) can proceed following the Zellner (1962) seemingly unrelated regression framework. However, if such external information is not available, two issues arise: (1) how to determine the number of groups, $G$; and (2) how to identify the group to which an individual belongs. Following the idea of Lasso (Least Absolute Shrinkage and Selection Operator; Tibshirani (1996)) under the assumption that the number of groups, $G$, is known, Su, Shi, and Phillips (SSP) (2013) suggest a modified penalized least-squares approach,

$$\text{Min } Q^G = Q + \frac{a}{N}\sum_{i=1}^{N}\prod_{g=1}^{G} \parallel \boldsymbol{\beta}_i - \boldsymbol{\beta}_g \parallel, \tag{6.2.4}$$

to simultaneously classify individuals into groups and estimate $\boldsymbol{\beta}_g$, where $\parallel \cdot \parallel$ denotes the Frobenius norm, $\parallel A \parallel = [tr\ AA']^{1/2}$,

$$Q = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta}_i)^2.$$

and $a$ is a tuning constant. SSP show that minimizing (6.2.4) achieves classification of individuals into groups and consistent estimation of $\boldsymbol{\beta}_g$ in a single step when $N$ and $T$ are large. SSP also propose to select the number of groups, $G$, by minimizing

$$\log \hat{\sigma}_G^2 + cGK, \tag{6.2.5}$$

where

$$\hat{\sigma}_G^2 = \frac{1}{NT}\sum_{g=1}^{G}\sum_{i\epsilon g}\sum_{t=1}^{T}(y_{it} - \mathbf{x}_{it}'\hat{\boldsymbol{\beta}}_{g,G})^2$$

is the estimated average residual sum of squares based on $G$-group estimates of (6.2.3), $\hat{\boldsymbol{\beta}}_{g,G}$, and $c$ is a turning constant.

### 6.2.2 Random-Coefficient Model

#### *6.2.2.1 The Model*

When $\boldsymbol{\beta}_i = \bar{\boldsymbol{\beta}} + \boldsymbol{\alpha}_i$ are treated as random, with common mean $\bar{\boldsymbol{\beta}}$, Swamy (1970) assumed that[3]

$$\begin{aligned} E\boldsymbol{\alpha}_i &= \mathbf{0}, \\ \underset{K\times K}{E\boldsymbol{\alpha}_i\boldsymbol{\alpha}_j'} &= \begin{cases} \Delta & \text{if } i = j, \\ \mathbf{0} & \text{if } i \neq j, \end{cases} \\ E\mathbf{x}_{it}\boldsymbol{\alpha}_j' &= \mathbf{0}, \quad E\boldsymbol{\alpha}_i\mathbf{u}_j' = \mathbf{0}, \\ E\mathbf{u}_i\mathbf{u}_j' &= \begin{cases} \sigma_i^2 I_T & \text{if } i = j, \\ \mathbf{0} & \text{if } i \neq j. \end{cases} \end{aligned} \tag{6.2.6}$$

Stacking all NT observations, we have

$$\mathbf{y} = X\bar{\boldsymbol{\beta}} + \tilde{X}\boldsymbol{\alpha} + \mathbf{u}, \tag{6.2.7}$$

where

$$\underset{NT\times 1}{\mathbf{y}} = (\mathbf{y}_1', \ldots, \mathbf{y}_N')',$$

$$\underset{NT\times K}{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}, \quad \underset{NT\times NK}{\tilde{X}} = \begin{bmatrix} X_1 & & & \mathbf{0} \\ & X_2 & & \\ & & \ddots & \\ \mathbf{0} & & & X_N \end{bmatrix} = \operatorname{diag}(X_1, \ldots, X_N),$$

$\mathbf{u} = (\mathbf{u}_1', \ldots, \mathbf{u}_N')'$, and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_N')'$. The covariance matrix for the composite disturbance term $\tilde{X}\boldsymbol{\alpha} + \mathbf{u}$ is block-diagonal, with the $i$th diagonal block given by

$$\Phi_i = X_i\Delta X_i' + \sigma_i^2 I_T. \tag{6.2.8}$$

[3] See Chamberlain (1992) for an extension of the Mundlak–Chamberlain approach of conditioning the individual effects on the conditioning variables to models with individual-specific slopes that may be correlated with conditioning variables. An instrumental variable estimator is proposed within a finite dimensional, method of moments framework. Also, see section 6.9.

### 6.2.2.2 *Estimation*

Under Swamy's (1970) assumption, the simple regression of $\mathbf{y}$ on $X$ will yield an unbiased and consistent estimator of $\bar{\boldsymbol{\beta}}$ if $(1/NT)\, X'X$ converges to a nonsingular constant matrix. But the estimator is inefficient, and the usual least-squares formula for computing the variance–covariance matrix of the estimator is incorrect, often leading to misleading statistical inferences. Moreover, when the pattern of parameter variation is of interest in its own right, an estimator ignoring parameter variation is incapable of shedding light on this aspect of the economic process.

The best linear unbiased estimator of $\bar{\boldsymbol{\beta}}$ for (6.2.7) is the GLS estimator[4]

$$\begin{aligned}\hat{\bar{\boldsymbol{\beta}}}_{GLS} &= \left(\sum_{i=1}^{N} X_i'\Phi_i^{-1}X_i\right)^{-1}\left(\sum_{i=1}^{N} X_i'\Phi_i^{-1}\mathbf{y}_i\right)\\ &= \sum_{i=1}^{N} W_i\hat{\boldsymbol{\beta}}_i,\end{aligned} \tag{6.2.9}$$

where

$$W_i = \left\{\sum_{i=1}^{N}[\Delta + \sigma_i^2(X_i'X_i)^{-1}]^{-1}\right\}^{-1}[\Delta + \sigma_i^2(X_i'X_i)^{-1}]^{-1},$$

and

$$\hat{\boldsymbol{\beta}}_i = (X_i'X_i)^{-1}X_i'\mathbf{y}_i.$$

The last expression of (6.2.9) shows that the GLS estimator is a matrix-weighted average of the least-squares estimator for each cross-sectional unit, with the weights inversely proportional to their covariance matrices. It also shows that the GLS estimator requires only a matrix inversion of order $K$, and so it is not much more complicated to compute than the simple least-squares estimator.

[4] Repeatedly using the formula that $(A + BDB')^{-1} = A^{-1} - A^{-1}B(B'A^{-1}B + D^{-1})^{-1}B'A^{-1}$ (Rao 1973, their Chapter 1), we have

$$\begin{aligned}X_i'\Phi_i^{-1}X_i &= X_i'[\sigma_i^2 I + X_i\Delta X_i']^{-1}X_i\\ &= X_i'\left\{\frac{1}{\sigma_i^2}I_T - \frac{1}{\sigma_i^2}X_i[X_i'X_i + \sigma_i^2\Delta^{-1}]^{-1}X_i'\right\}X_i\\ &= \frac{1}{\sigma_i^2}\left[X_i'X_i - X_i'X_i\left\{(X_i'X_i)^{-1}\right.\right.\\ &\quad \left.\left. - (X_i'X_i)^{-1}\left[(X_i'X_i)^{-1} + \frac{1}{\sigma_i^2}\Delta\right]^{-1}(X_i'X_i)^{-1}\right\}X_i'X_i\right]\\ &= [\Delta + \sigma_i^2(X_i'X_i)^{-1}]^{-1}.\end{aligned}$$

The covariance matrix for the GLS estimator is

$$\begin{aligned}\text{Var}\,(\hat{\bar{\boldsymbol{\beta}}}_{GLS}) &= \left(\sum_{i=1}^{N} X_i'\Phi_i^{-1}X_i\right)^{-1} \\ &= \left\{\sum_{i=1}^{N}[\Delta + \sigma_i^2(X_i'X_i)^{-1}]^{-1}\right\}^{-1}.\end{aligned} \tag{6.2.10}$$

Swamy (1970) proposed using the least-squares estimators $\hat{\boldsymbol{\beta}}_i = (X_i'X_i)^{-1}X_i'\mathbf{y}_i$ and their residuals $\hat{\mathbf{u}}_i = \mathbf{y}_i - X_i\hat{\boldsymbol{\beta}}_i$ to obtain unbiased estimators of $\sigma_i^2$ and $\Delta$,[5]

$$\begin{aligned}\hat{\sigma}_i^2 &= \frac{\hat{\mathbf{u}}_i'\hat{\mathbf{u}}_i}{T-K} \\ &= \frac{1}{T-K}\mathbf{y}_i'[I - X_i(X_i'X_i)^{-1}X_i']\mathbf{y}_i,\end{aligned} \tag{6.2.11}$$

$$\begin{aligned}\hat{\Delta} = \frac{1}{N-1}\sum_{i=1}^{N}&\left(\hat{\boldsymbol{\beta}}_i - N^{-1}\sum_{i=1}^{N}\hat{\boldsymbol{\beta}}_i\right) \\ &\cdot\left(\hat{\boldsymbol{\beta}}_i - N^{-1}\sum_{i=1}^{N}\hat{\beta}_i\right)' - \frac{1}{N}\sum_{i=1}^{N}\hat{\sigma}_i^2(X_i'X_i)^{-1}.\end{aligned} \tag{6.2.12}$$

Again, just as in the error-component model, the estimator (6.2.12) is not necessarily nonnegative definite. In this situation, Swamy (see also Judge et al. 1980) has suggested replacing (6.2.12) by

$$\hat{\Delta} = \frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{\boldsymbol{\beta}}_i - N^{-1}\sum_{i=1}^{N}\hat{\boldsymbol{\beta}}_i\right)\left(\hat{\boldsymbol{\beta}}_i - N^{-1}\sum_{i=1}^{N}\hat{\boldsymbol{\beta}}_i\right)'. \tag{6.2.13}$$

This estimator, although not unbiased, is non-negative definite and is consistent when both $N$ and $T$ tend to infinity. Alternatively, we can use the Bayes mode estimator suggested by Lindley and Smith (1972) and Smith (1973),

$$\Delta^* = \{R + (N-1)\hat{\Delta}\}/(N + \rho - K - 2), \tag{6.2.14}$$

where $R$ and $\rho$ are prior parameters, assuming that $\Delta^{-1}$ has a Wishart distribution with $\rho$ degrees of freedom and matrix $R$. For instance, we may let $R = \hat{\Delta}$ and $\rho = 2$ as in Hsiao, Pesaran, and Tahmiscioglu (1999).

Swamy (1970) proved that substituting $\hat{\sigma}_i^2$ and $\hat{\Delta}$ for $\sigma_i^2$ and $\Delta$ in (6.2.9) yields an asymptotically normal and efficient estimator of $\bar{\boldsymbol{\beta}}$. The speed of convergence of the GLS estimator is $N^{1/2}$. This can be seen by noting that the

[5] Equation (6.2.9) follows from the relation that $\hat{\boldsymbol{\beta}}_i = \boldsymbol{\beta}_i + (X_i'X_i)^{-1}X'\mathbf{u}_i$ and $E(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta})' = \Delta + \sigma_i^2(X_i'X_i)^{-1}$.

inverse of the covariance matrix for the GLS estimator [equation (6.2.10)] is[6]

$$\begin{aligned}\text{Var}\,(\hat{\bar{\boldsymbol{\beta}}}_{\text{GLS}})^{-1} &= N\Delta^{-1} - \Delta^{-1}\left[\sum_{i=1}^{N}\left(\Delta^{-1} + \frac{1}{\sigma_i^2}X_i'X_i\right)^{-1}\right]\Delta^{-1} \\ &= O(N) - O(N/T).\end{aligned} \tag{6.2.15}$$

Swamy (1970) used the model (6.2.6) and (6.2.7) to reestimate the Grunfeld investment function with the annual data of 11 U.S. corporations. His GLS estimates of the common-mean coefficients of the firms' beginning-of-year value of outstanding shares and capital stock are 0.0843 and 0.1961, with asymptotic standard errors 0.014 and 0.0412, respectively. The estimated dispersion measure of these coefficients is

$$\hat{\Delta} = \begin{bmatrix} 0.0011 & -0.0002 \\ & 0.0187 \end{bmatrix}. \tag{6.2.16}$$

Zellner (1966) has shown that when each $\boldsymbol{\beta}_i$ can be viewed as a random variable with a constant mean, and $\boldsymbol{\beta}_i$ and $\mathbf{x}_i$ are uncorrelated, thereby satisfying Swamy's (1970) assumption, the model will not possess an aggregation bias. In this sense, Swamy's estimate can also be interpreted as an average relationship indicating that in general the value of a firm's outstanding shares is an important variable explaining the investment.

#### *6.2.2.3 Predicting Individual Coefficients*

Sometimes one may wish to predict the individual component $\boldsymbol{\beta}_i$, because it provides information on the behavior of each individual and also because it provides a basis for predicting future values of the dependent variable for a given individual. Swamy (1970, 1971) has shown that the best linear unbiased predictor, conditional on given $\boldsymbol{\beta}_i$, is the least-squares estimator $\hat{\boldsymbol{\beta}}_i$. However, if the sampling properties of the class of predictors are considered in terms of repeated sampling over both time and individuals, Lee and Griffiths (1979) (see also Lindley and Smith 1972 and Section 6.6) have suggested predicting $\boldsymbol{\beta}_i$ by

$$\hat{\boldsymbol{\beta}}_i^* = \hat{\bar{\boldsymbol{\beta}}}_{\text{GLS}} + \Delta X_i'(X_i\Delta X_i' + \sigma_i^2 I_T)^{-1}(\mathbf{y}_i - X_i\hat{\bar{\boldsymbol{\beta}}}_{\text{GLS}}). \tag{6.2.17}$$

This predictor is the best linear unbiased estimator in the sense that $E(\hat{\boldsymbol{\beta}}_i^* - \boldsymbol{\beta}_i) = \mathbf{0}$, where the expectation is an unconditional one.

#### *6.2.2.4 Testing for Coefficient Variation*

An important question in empirical investigation is whether or not the regression coefficients are indeed varying across cross-sectional units. Because the

[6] We use the notation $O(N)$ to denote that the sequence $N^{-1}a_N$ is bounded (Theil 1971, p. 358).

effect of introducing random-coefficient variation is to give the dependent variable a different variance at each observation, models with this feature can be transformed into a particular heteroscedastic formulation, and likelihood-ratio tests can be used to detect departure from the constant-parameter assumption. However, computation of the likelihood-ratio test statistic can be complicated. To avoid the iterative calculations necessary to obtain maximum-likelihood estimates of the parameters in the full model, Breusch and Pagan (1979) have proposed a Lagrangian multiplier test for heteroscedasticity. Their test has the same asymptotic properties as the likelihood-ratio test in standard situations, but it is computationally much simpler. It can be computed simply by repeatedly applying least-square regressions.

Dividing the individual-mean-over-time equation by $\sigma_i^{-1}$, we have

$$\frac{1}{\sigma_i}\bar{y}_i = \frac{1}{\sigma_i}\bar{\mathbf{x}}_i'\bar{\boldsymbol{\beta}} + \omega_i, \quad i = 1., \ldots, N, \tag{6.2.18}$$

where $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$, $\bar{\mathbf{x}}_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{it}$, $\bar{u}_i = \frac{1}{T}\sum_{t=1}^{T} u_{it}$,

$$\omega_i = \frac{1}{\sigma_i}\bar{\mathbf{x}}_i'\boldsymbol{\alpha}_i + \frac{1}{\sigma_i}\bar{u}_i.$$

When the assumption (6.2.6) holds, model (6.2.18) is a model with heteroscedastic variances, $\mathrm{Var}(\omega_i) = (1/T) + (1/\sigma_i^2)\bar{\mathbf{x}}_i'\Delta\bar{\mathbf{x}}_i$, $i = 1, \ldots, N$. Under the null hypothesis that $\Delta = \mathbf{0}$, (6.2.18) has homoscedastic variances, $\mathrm{Var}(\omega_i) = 1/T$, $i = 1, \ldots, N$. Thus, we can generalize the Breusch and Pagan (1979) test of heteroscedasticity to test for random-coefficient variation here.

Following the procedures of Rao (1973, pp. 418–19) it can be shown that the transformed Lagrangian-multiplier statistic[7] for testing the null hypothesis leads to computing one-half the predicted sum of squares in a regression of

$$(T\omega_i^2 - 1) = \frac{1}{\sigma_i^2}\left[\sum_{k=1}^{K}\sum_{k'=1}^{K}\bar{x}_{ki}\bar{x}_{k'i}\sigma^2_{\alpha_{kk'}}\right] + \epsilon_i, \quad i = 1, \ldots, N, \tag{6.2.19}$$

where $\sigma^2_{\alpha_{kk'}} = E(\alpha_{ki}\alpha_{k'i})$.[8] Because $\omega_i$ and $\sigma_i^2$ usually are unknown, we can substitute them by their estimated values $\hat{\omega}_i$ and $\hat{\sigma}_i^2$, where $\hat{\omega}_i$ is the least-squares residual of (6.2.18) and $\hat{\sigma}_i^2$ is given by (6.2.11). When both $N$ and $T$ tend to infinity, the transformed Lagrangian-multiplier statistic has the same

[7] We call this a transformed Lagrangian multiplier test because it is derived by maximizing the log-likelihood function of $y_i/\sigma_i$ rather than maximizing the log-likelihood function of $y_{it}/\sigma_{it}$.

[8] Let

$$(T\hat{\omega}_i^2 - 1) = \frac{1}{\sigma_i^2}\left[\sum_{i=1}^{K}\sum_{i'=1}^{K}\bar{x}_{ki}\bar{x}_{k'i}\hat{\sigma}_{\alpha kk'}\right]$$

be the least-squares predicted value of $(T\hat{\omega}_i^2 - 1)$; then the predicted sum of squares is

$$\sum_{i=1}^{N}(T\hat{\omega}_i^2 - 1)^2.$$

limiting distribution as $\chi^2$ with $[K(K+1)]/2$ degrees of freedom under the null hypothesis of $\Delta = \mathbf{0}$.

The Breusch and Pagan (1979) lagrangian multiplier test can be put into the White (1980) information matrix test framework. Chesher (1984) has shown that the many variants of varying parameters of the same general type of model under consideration can be tested using the statistic

$$D_N(\hat{\boldsymbol{\theta}}_N) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\frac{\partial^2 \log f(y_{it} \mid \mathbf{x}_{it}, \hat{\boldsymbol{\theta}}_N)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} + \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{t=1}^{T}\frac{\partial \log f(y_{it} \mid \mathbf{x}_{it}, \hat{\boldsymbol{\theta}}_N)}{\partial\boldsymbol{\theta}}\right]\left[\sum_{t=1}^{T}\frac{\partial \log f(y_{it} \mid \mathbf{x}_{it}; \hat{\boldsymbol{\theta}}_N)}{\partial\boldsymbol{\theta}'}\right], \tag{6.2.20}$$

where $f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\theta})$ denotes the conditional density of $y_{it}$ given $\mathbf{x}_{it}$ and $\boldsymbol{\theta}$ under the null of no parameter variation, and $\hat{\boldsymbol{\theta}}_N$ denotes the maximum-likelihood estimator of $\boldsymbol{\theta}$. Under the null, $E\big(\frac{\partial^2 \log f(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\big) = -E\big(\frac{\partial \log f(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \cdot \frac{\partial \log f(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\big)$. Therefore, elements of $\sqrt{N}D_N(\hat{\boldsymbol{\theta}}_N)$ are asymptotically jointly normal with mean 0 and covariance matrix given by White (1980) and simplified by Chesher (1983) and Lancaster (1984).

Alternatively, because for given $i$, $\boldsymbol{\alpha}_i$ is fixed, we can test for random variation indirectly by testing whether or not the fixed-coefficient vectors $\boldsymbol{\beta}_i$ are all equal. That is, we form the null hypothesis:

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \ldots = \boldsymbol{\beta}_N = \bar{\boldsymbol{\beta}}.$$

If different cross-sectional units have the same variance, $\sigma_i^2 = \sigma^2, i = 1, \ldots, N$, the conventional analysis-of-covariance (ANCOVA) test for homogeneity discussed in Chapter 2 ($F_3$) can be applied. If $\sigma_i^2$ are assumed different, as postulated by Swamy (1970, 1971), we can apply the modified test statistic

$$F_3^* = \sum_{i=1}^{N}\frac{(\hat{\boldsymbol{\beta}}_i - \hat{\bar{\boldsymbol{\beta}}}^*)' X_i' X_i (\hat{\boldsymbol{\beta}}_i - \hat{\bar{\boldsymbol{\beta}}}^*)}{\hat{\sigma}_i^2}, \tag{6.2.21}$$

where

$$\hat{\bar{\boldsymbol{\beta}}}^* = \left[\sum_{i=1}^{N}\frac{1}{\hat{\sigma}_i^2}X_i'X_i\right]^{-1}\left[\sum_{i=1}^{N}\frac{1}{\hat{\sigma}_i^2}X_i'\mathbf{y}_i\right].$$

Under $H_0$, (6.2.21) is asymptotically $\chi^2$ distributed, with $K(N-1)$ degrees of freedom, as $T$ tends to infinity and $N$ is fixed.

Similarly, one can test for slope homogeneity conditional on individual-specific effects. Let $X_i = (\mathbf{e}_T, \tilde{X}_i)$ and $\boldsymbol{\beta}_i' = (\beta_{1i}, \boldsymbol{\beta}_{2i}')$, where $\tilde{X}_i$ denotes the $T \times (K-1)$ time-varying exogenous variables, $\tilde{\mathbf{x}}_{2,it}$ and $\boldsymbol{\beta}_{2i}$ denotes the

$(K-1)\times 1$ coefficients of $\mathbf{x}_{2,it}$. Then

$$\tilde{F}_3^* = \sum_{i=1}^{N} \left(\hat{\boldsymbol{\beta}}_{2i} - \hat{\tilde{\boldsymbol{\beta}}}_2^*\right)' \left[\frac{1}{\hat{\sigma}_i^2}\tilde{X}_i' Q \tilde{X}_i\right] \left(\hat{\boldsymbol{\beta}}_{2i} - \hat{\tilde{\boldsymbol{\beta}}}_2^*\right), \tag{6.2.22}$$

where $Q = I_T - \frac{1}{T}\mathbf{e}\mathbf{e}'$,

$$\hat{\boldsymbol{\beta}}_{2i} = (\tilde{X}_i' Q \tilde{X}_i)^{-1}(\tilde{X}_i' Q \mathbf{y}_i), \tag{6.2.23}$$

$$\hat{\tilde{\boldsymbol{\beta}}}_2^* = \left(\sum_{i=1}^{N} \frac{1}{\hat{\sigma}_i^2}\tilde{X}_i' Q \tilde{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \frac{1}{\hat{\sigma}_i^2}\tilde{X}_i' Q y_i\right) \tag{6.2.24}$$

and

$$\hat{\sigma}_i^2 = \frac{1}{T-K}(\mathbf{y}_i - \tilde{\mathbf{X}}_i'\hat{\boldsymbol{\beta}}_{2i})' Q (\mathbf{y}_i - \tilde{X}_i'\hat{\boldsymbol{\beta}}_{2i}). \tag{6.2.25}$$

The statistic $\tilde{F}_3^*$ is asymptotically $\chi^2$ distributed with $(K-1)(N-1)$ degrees of freedom when $N$ is fixed and $T \to \infty$. Pesaran and Yamagata (2008) show that when both $N$ and $T$ go to infinity $\frac{1}{\sqrt{N}}F_3^*$ or $\frac{1}{\sqrt{N}}\tilde{F}_3^*$ is asymptotically normally distributed with mean 0 and variance 1 provided $\frac{\sqrt{N}}{T} \to 0$ as $N \to \infty$. Furthermore, they show that if $\hat{\sigma}_i^2$ ((6.2.25)) is replaced by the estimator

$$\tilde{\sigma}_i^2 = \frac{1}{T-1}(\mathbf{y}_i - \tilde{X}_i\hat{\boldsymbol{\beta}}_{2i})' Q (\mathbf{y}_i - \tilde{X}_i\tilde{\boldsymbol{\beta}}_{2i}), \tag{6.2.26}$$

$\frac{1}{\sqrt{N}}F_3^*$ or $\frac{1}{\sqrt{N}}\tilde{F}_3^*$ possesses better finite sample properties than using $\hat{\sigma}_i^2$ in (6.2.21) or (6.2.22).

### *6.2.2.5 Fixed or Random Coefficients*

The question whether $\boldsymbol{\beta}_i$ should be assumed fixed and different or random and different depends on whether $\boldsymbol{\beta}_i$ can be viewed as from a heterogeneous population or random draws from a common population or whether we are making inferences conditional on the individual characteristics or making unconditional inferences on the population characteristics. If $\boldsymbol{\beta}_i$ are heterogeneous or we are making inferences conditional on the individual characteristics, the fixed-coefficient model should be used. If $\boldsymbol{\beta}_i$ can be viewed as random draws from a common population and inference is on the population characteristics, the random-coefficient model should be used. However, extending his work on the variable-intercept model, Mundlak (1978b) has raised the issue of whether or not the variable coefficients are correlated with the explanatory variables. If they are, the assumptions of the Swamy random-coefficient model are unreasonable, and the GLS estimator of the mean coefficient vector will be biased. To correct this bias, Mundlak (1978b) suggested that the inferences of $f(\mathbf{y}_i \mid X_i, \boldsymbol{\beta})$ be viewed as $\int f(\mathbf{y}_i \mid X_i, \bar{\boldsymbol{\beta}}, \boldsymbol{\alpha}_i) f(\boldsymbol{\alpha}_i \mid X_i)\, d\boldsymbol{\alpha}_i$, where $f(\mathbf{y}_i \mid X_i, \bar{\boldsymbol{\beta}}, \boldsymbol{\alpha}_i)$ denotes the conditional density of $y_i$ given $X_i$, $\bar{\boldsymbol{\beta}}$ and $\boldsymbol{\alpha}_i$ and

$f(\boldsymbol{\alpha}_i \mid X_i)$ denotes the conditional density of $\boldsymbol{\alpha}_i$ given $X_i$ which provides auxiliary equations for the coefficient vector $\boldsymbol{\alpha}_i$ as a function of the $i$th individual's observed explanatory variables. Because this framework can be viewed as a special case of a random-coefficients model with the coefficients being functions of other explanatory variables, we shall maintain the assumption that the random coefficients are not correlated with the explanatory variables, and we shall discuss estimation of the random coefficients that are functions of other explanatory variables in Section 6.5.

#### *6.2.2.6 An Example*

To illustrate the specific issues involved in estimating a behavioral equation using temporal cross-sectional observations when the data do not support the hypothesis that the coefficients are the same for all cross-sectional units, we report a study conducted by Barth, Kraft, and Kraft (1979). They used quarterly observations on output prices, wages, materials prices, inventories, and sales for 17 manufacturing industries for the period 1959 (I) to 1971 (II) to estimate a price equation for the U.S. manufacturing sector. Assuming heteroscedastic disturbance, but common intercept and slope coefficients across industries, and using the two-step Aitken estimator, Barth et al. (1979) obtained

$$\begin{array}{lccccc} \hat{y} = & 0.0005 & + & 0.2853x_2 + & 0.0068x_3 + & 0.0024x_4, \\ & (0.0003) & & (0.0304) & (0.005) & (0.0017) \end{array} \tag{6.2.27}$$

where $y_t$ is the quarterly change in output price, $x_2$ is labor costs, $x_3$ is materials input prices, and $x_4$ is a proxy variable for demand, constructed from the ratio of finished inventory to sales. The standard errors of the estimates are in parentheses.

The findings of (6.2.27) are somewhat unsettling. The contribution of materials input costs is extremely small, less than 1 percent. Furthermore, the proxy variable has the wrong sign. As the inventory-to-sales ratio increases, one would expect the resulting inventory buildup to exert a downward pressure on prices.

There are many reasons that (6.2.27) can go wrong. For instance, pricing behavior across industries is likely to vary, because input combinations are different, labor markets are not homogeneous, and demand may be more elastic or inelastic in one industry than another. In fact, a modified one-way ANCOVA test for the common intercept and slope coefficients,

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_N, \quad N = 17,$$

using the statistic (6.2.21), has a value of 449.28. That is well above the $\chi^2$ critical value of 92.841 for the 1 percent significance level with 64 ($(N-1)K$) degrees of freedom.

The rejection of the hypothesis of homogeneous price behavior across industries suggests a need to modify the model to allow for heterogeneous behavior across industries. However, previous studies have found that output prices are affected mainly by unit labor and materials input costs, and secondly, if at all,

by demand factors. Thus, to account for heterogeneous behavior, one can assume that the relationships among variables are proper, but the coefficients are different across industries. But if these coefficients are treated as fixed and different, this will imply a complicated aggregation problem for the price behavior of the U.S. manufacturing sector (e.g., Theil 1954). On the other hand, if the coefficients are treated as random, with common means, there is no aggregation bias (Zellner 1966). The random-coefficient formulation will provide a microeconomic foundation to aggregation, as well as permit the aggregate-price equation to capture more fully the disaggregated industry behavior. Therefore, Barth et al. (1979) used the Swamy random-coefficient formulation, (6.2.6) and (6.2.7), to reestimate the price equation. Their new estimates, with standard errors in parentheses, are

$$\begin{matrix} \hat{y} = & -0.0006 & + & 0.3093x_2 & + & 0.2687x_3 & - & 0.0082x_4. \\ & (0.0005) & & (0.0432) & & (0.0577) & & (0.0101) \end{matrix} \tag{6.2.28}$$

The estimated dispersion of these coefficients is

$$\hat{\Delta} = \begin{bmatrix} \beta_1 & \beta_2 & \beta_3 & \beta_4 \\ 0.0000 & -0.0002 & 0.0000 & -0.0001 \\ & 0.0020 & 0.0003 & 0.0081 \\ & & 0.0320 & 0.0030 \\ & & & 0.0014 \end{bmatrix}. \tag{6.2.29}$$

The results of the Swamy random-coefficient formulation appear more plausible than the previous aggregate price specification [equation (6.2.27), which ignores variation across industries] from several points of view: (1) both labor costs and materials costs are now dominant in determining output prices; (2) the proxy variable for demand has the correct sign, although it plays only a small and insignificant role in the determination of manufacturing prices; and (3) productivity, as captured in the intercept term, appears to be increasing.

This example suggests that one must be careful about drawing conclusions on the basis of aggregate data or pooled estimates that do not allow for individual heterogeneity. Such estimates can be misleading in terms of both the size of coefficients and the significance of variables.

## 6.3 COEFFICIENTS THAT VARY OVER TIME AND CROSS-SECTIONAL UNITS

### 6.3.1 The Model

Just as in the variable-intercept models, it is possible to assume that the coefficient of the explanatory variable has a component specific to an individual unit

and a component specific to a given time period such that

$$y_{it} = \sum_{k=1}^{K} (\bar{\beta}_k + \alpha_{ki} + \lambda_{kt}) x_{kit} + u_{it}, \qquad \begin{array}{l} i = 1, \ldots, N, \\ t = 1, \ldots, T. \end{array} \tag{6.3.1}$$

Stacking all $NT$ observations, we can rewrite (6.3.1) as

$$\mathbf{y} = X\bar{\boldsymbol{\beta}} + \tilde{X}\boldsymbol{\alpha} + \underset{\sim}{\mathbf{X}}\boldsymbol{\lambda} + \mathbf{u}, \tag{6.3.2}$$

where $\mathbf{y}$, $X$, $\tilde{X}$, $\mathbf{u}$, and $\boldsymbol{\alpha}$ are defined in Section 6.2,

$$\underset{NT \times TK}{\underset{\sim}{\mathbf{X}}} = \begin{bmatrix} \underset{\sim}{\mathbf{X}}_1 \\ \underset{\sim}{\mathbf{X}}_2 \\ \vdots \\ \underset{\sim}{\mathbf{X}}_N \end{bmatrix}, \quad \underset{T \times TK}{\underset{\sim}{\mathbf{X}}_i} = \begin{bmatrix} \mathbf{x}_{i1}' & & & \mathbf{0}' \\ & \mathbf{x}_{i2}' & & \\ & & \ddots & \\ \mathbf{0} & & & \mathbf{x}_{iT}' \end{bmatrix},$$

and

$$\underset{KT \times 1}{\boldsymbol{\lambda}} = (\boldsymbol{\lambda}_1', \ldots, \boldsymbol{\lambda}_T')', \; \underset{K \times 1}{\boldsymbol{\lambda}_t} = (\lambda_{1t}, \ldots, \lambda_{Kt})',$$

We can also rewrite (6.3.2) as

$$\begin{aligned} \mathbf{y} = {} & X\bar{\boldsymbol{\beta}} + U_1\boldsymbol{\alpha}_1 + U_2\boldsymbol{\alpha}_2 + \cdots + U_K\boldsymbol{\alpha}_K \\ & + U_{K+1}\boldsymbol{\lambda}_1 + \cdots + U_{2K}\boldsymbol{\lambda}_K + U_{2K+1}\mathbf{u}, \end{aligned} \tag{6.3.3}$$

where

$$\underset{NT \times N}{U_k} = \begin{bmatrix} x_{k11} & & & & \mathbf{0} \\ \vdots & & & & \\ x_{k1T} & & & & \\ & x_{k21} & & & \\ & \vdots & & & \\ & x_{k2T} & & & \\ & & \ddots & & \\ & & & x_{kN1} & \\ \mathbf{0} & & & \vdots & \\ & & & x_{kNT} & \end{bmatrix}, \quad k = 1 \ldots, K,$$

$$
\underset{NT\times T}{U_{K+k}} = \begin{bmatrix}
x_{k11} & & & \mathbf{0} \\
 & x_{k12} & & \\
 & & \vdots & \\
\mathbf{0} & & & x_{k1T} \\
x_{k21} & & & \mathbf{0} \\
 & x_{k22} & & \\
 & & \ddots & \\
\mathbf{0} & & & x_{k2T} \\
x_{kN1} & & & \mathbf{0} \\
 & \ddots & & \\
\mathbf{0} & & & x_{kNT}
\end{bmatrix}, \quad k = 1.\ldots, K, \tag{6.3.4}
$$

$$
U_{2K+1} = I_{NT},
$$

$$
\underset{N\times 1}{\boldsymbol{\alpha}_k} = (\alpha_{k1}, \ldots, \alpha_{kN})', \quad \underset{T\times 1}{\boldsymbol{\lambda}_k} = (\lambda_{k1}, \ldots, \lambda_{kT})'.
$$

When $\boldsymbol{\alpha}_k$ and $\boldsymbol{\lambda}_k$ as well as $\bar{\boldsymbol{\beta}}$ are considered fixed, it is a fixed-effects model; when $\boldsymbol{\alpha}_k$ and $\boldsymbol{\lambda}_k$ are considered random, with $\bar{\boldsymbol{\beta}}$ fixed, equation (6.3.3) corresponds to the mixed analysis-of-variance (ANOVA) model (Hartley and Rao 1967). Thus, model (6.3.1) and its special case, model (6.2.1), fall within the general ANOVA framework.

### 6.3.2 Fixed-Coefficient Model

When $\boldsymbol{\alpha}_k$ and $\boldsymbol{\lambda}_k$ are treated as fixed, as mentioned earlier, (6.3.1) can be viewed as a fixed-effects ANOVA model. However, the matrix of explanatory variables is $NT \times (T + N + 1)K$, but its rank is only $(T + N - 1)K$; so we must impose $2K$ independent linear restrictions on the coefficients $\boldsymbol{\alpha}_k$ and $\boldsymbol{\lambda}_k$ for estimation of $\bar{\boldsymbol{\beta}}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\lambda}$. A natural way of imposing the constraints in this case is to let[9]

$$
\sum_{i=1}^{N} \alpha_{ik} = 0, \tag{6.3.5}
$$

and

$$
\sum_{t=1}^{T} \lambda_{kt} = 0, \quad k = 1.\ldots, K. \tag{6.3.6}
$$

[9] We did not impose similar restrictions in Chapter 6, Section 6.2.1 because we did not separate $\boldsymbol{\beta}$ from $\boldsymbol{\alpha}_i$.

Then the best linear unbiased estimators (BLUEs) of $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\lambda}$ are the solutions of

$$\min\ (\mathbf{y} - X\bar{\boldsymbol{\beta}} - \tilde{X}\boldsymbol{\alpha} - \underline{\mathbf{X}}\boldsymbol{\lambda})'(\mathbf{y} - X\bar{\boldsymbol{\beta}} - \tilde{X}\boldsymbol{\alpha} - \underline{\mathbf{X}}\boldsymbol{\lambda}) \tag{6.3.7}$$

subject to (6.3.5) and (6.3.6).

### 6.3.3 Random-Coefficient Model

When $\boldsymbol{\alpha}_i$ and $\boldsymbol{\lambda}_t$ are treated as random, Hsiao (1974a, 1975) assumes that

$$\begin{aligned} E\underset{K\times K}{\boldsymbol{\alpha}_i\boldsymbol{\alpha}_j'} &= \begin{cases} \Delta & \text{if } i = j, \\ \mathbf{0} & \text{if } i \neq j, \end{cases} \\ E\underset{K\times K}{\boldsymbol{\lambda}_t\boldsymbol{\lambda}_s'} &= \begin{cases} \Lambda & \text{if } t = s, \\ \mathbf{0} & \text{if } t \neq s, \end{cases} \end{aligned} \tag{6.3.8}$$

$$E\boldsymbol{\alpha}_i\boldsymbol{\lambda}_t' = \mathbf{0}, \quad E\boldsymbol{\alpha}_i\mathbf{x}_{it}' = \mathbf{0}, \quad E\boldsymbol{\lambda}_t\mathbf{x}_{it}' = \mathbf{0},$$

and

$$E\mathbf{u}_i\mathbf{u}_j' = \begin{cases} \sigma_u^2 I_T & \text{if } i = j, . \\ \mathbf{0} & \text{if } i \neq j. \end{cases}$$

Then the composite error term,

$$\mathbf{v} = \tilde{X}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\lambda} + \mathbf{u}, \tag{6.3.9}$$

has a variance–covariance matrix

$$\begin{aligned} \Omega = E\mathbf{v}\mathbf{v}' = &\begin{bmatrix} X_1\Delta X_1' & & & \mathbf{0} \\ & X_2\Delta X_2' & & \\ & & \ddots & \\ \mathbf{0} & & & X_N\Delta X_N' \end{bmatrix} \\ &+ \begin{bmatrix} D(X_1\Lambda X_1') & D(X_1\Lambda X_2') & \dots & D(X_1\Lambda X_N') \\ D(X_2\Lambda X_1') & D(X_2\Lambda X_2') & & \\ & & & \ddots \\ D(X_N\Lambda X_1') & & & D(X_N\Lambda X_N') \end{bmatrix} + \sigma_u^2 I_{NT}, \end{aligned} \tag{6.3.10}$$

where

$$\underset{T\times T}{D(X_i\Lambda X_j')} = \begin{bmatrix} \mathbf{x}_{i1}'\Lambda\mathbf{x}_{j1} & & & \mathbf{0} \\ & \mathbf{x}_{i2}'\Lambda\mathbf{x}_{j2} & & \\ & & \ddots & \\ \mathbf{0} & & & \mathbf{x}_{iT}'\Lambda\mathbf{x}_{jT} \end{bmatrix}.$$

We can estimate $\bar{\boldsymbol{\beta}}$ by the least-squares method, but as discussed in Section 6.2.2.2, it is not efficient. Moreover, the conventional formula for the covariance matrix of the least-squares estimator is misleading. If $\Omega$ is known, the BLUE of $\bar{\boldsymbol{\beta}}$ is the GLS estimator,

$$\hat{\bar{\boldsymbol{\beta}}}_{\text{GLS}} = (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}\mathbf{y}). \tag{6.3.11}$$

The variance–covariance matrix of the GLS estimator is

$$\text{Var}\,(\hat{\bar{\boldsymbol{\beta}}}_{\text{GLS}}) = (X'\Omega^{-1}X)^{-1}. \tag{6.3.12}$$

Without knowledge of $\Omega$, we can estimate $\bar{\boldsymbol{\beta}}$ and $\Omega$ simultaneously by the maximum-likelihood method. However, because of the computational difficulty, a natural alternative is to first estimate $\Omega$, and then substitute the estimated $\Omega$ in (6.3.11).

When $\Delta$ and $\Lambda$ are diagonal, it is easy to see from (6.3.3) that $\Omega$ is a linear combination of known matrices with unknown weights. So the problem of estimating the unknown covariance matrix is actually the problem of estimating the variance components. Statistical methods developed for estimating the variance (and covariance) components can be applied here (e.g., Anderson 1969, 1970; Rao 1970, 1972). In this section we shall describe only a method due to Hildreth and Houck (1968).[10]

Consider the time-series equation for the $i$th individual,

$$\mathbf{y}_i = X_i(\bar{\boldsymbol{\beta}} + \boldsymbol{\alpha}_i) + \underline{\mathbf{X}}_i\boldsymbol{\lambda} + \mathbf{u}_i. \tag{6.3.13}$$

We can treat $\boldsymbol{\alpha}_i$ as if it is a vector of constants. Then (6.3.13) is a linear model with heteroscedastic variance. The variance of the error term $r_{it} = \sum_{k=1}^{K}\lambda_{kt}x_{kit} + u_{it}$ is

$$\theta_{it} = E[r_{it}^2] = \sum_{k=1}^{K}\sigma_{\lambda k}^2 x_{kit}^2 + \sigma_u^2. \tag{6.3.14}$$

Let $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iT})'$; then

$$\boldsymbol{\theta}_i = \dot{X}_i\boldsymbol{\sigma}_\lambda^2, \tag{6.3.15}$$

where the first element of $\mathbf{x}_{it} = 1$, $\dot{X}_i$ is $X_i$ with each of its elements squared, and $\boldsymbol{\sigma}_\lambda^2 = (\sigma_{\lambda 1}^2 + \sigma_u^2, \sigma_{\lambda 2}^2, \ldots, \sigma_{\lambda K}^2)'$.

An estimate of $\mathbf{r}_i$ can be obtained as the least-squares residual, $\hat{\mathbf{r}}_i = \mathbf{y}_i - X_i\hat{\boldsymbol{\beta}}_i = M_i\mathbf{y}_i$, where $\hat{\boldsymbol{\beta}}_i = (X_i'X_i)^{-1}X_i'\mathbf{y}_i$ and $M_i = I_T - X_i(X_i'X_i)^{-1}X_i'$. Squaring each element of $\hat{\mathbf{r}}_i$ and denoting it by $\dot{\hat{\mathbf{r}}}_i$, we have

$$E(\dot{\hat{\mathbf{r}}}_i) = \dot{M}_i\theta_i = F_i\sigma_\lambda^2, \tag{6.3.16}$$

where $\dot{M}_i$ is $M_i$ with each of its elements squared, and $F_i = \dot{M}_i\dot{X}_i$.

[10] It has been shown (Hsiao 1975) that the Hildreth–Houck estimator is the minimum-norm quadratic unbiased estimator of Rao (1970).

Repeating the foregoing process for all $i$ gives

$$E(\dot{\hat{\mathbf{r}}}) = F\boldsymbol{\sigma}_{\lambda}^2, \tag{6.3.17}$$

where $\dot{\hat{\mathbf{r}}} = (\dot{\hat{\mathbf{r}}}_1, \ldots, \dot{\hat{\mathbf{r}}}_N)'$, and $F = (F_1', \ldots, F_N')'$. Application of least-squares to (6.3.17) yields a consistent estimator of $\sigma_{\lambda}^2$,

$$\hat{\boldsymbol{\sigma}}_{\lambda}^2 = (F'F)^{-1}F'\dot{\hat{\mathbf{r}}}. \tag{6.3.18}$$

Similarly, we can apply the same procedure with respect to each time period to yield a consistent estimator of $\boldsymbol{\sigma}_{\alpha}^2 = (\sigma_{\alpha_1}^2 + \sigma_u^2, \sigma_{\alpha_2}^2, \ldots, \sigma_{\alpha_K}^2)'$. To obtain separate estimates of $\sigma_u^2$, $\sigma_{\alpha_1}^2$, and $\sigma_{\lambda 1}^2$, we note that $E(\mathbf{x}_{it}'\boldsymbol{\alpha}_i + u_{it})(\mathbf{x}_{it}'\boldsymbol{\lambda}_t + u_{it}) = \sigma_u^2$. So, letting $\hat{s}_{it}$ denote the residual obtained by applying least-squares separately to each time period, we can consistently estimate $\sigma_u^2$ by

$$\hat{\sigma}_u^2 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{r}_{it}\hat{s}_{it}. \tag{6.3.19}$$

Subtracting (6.3.19) from an estimated $\sigma_{\alpha_1}^2 + \sigma_u^2$ and $\sigma_{\lambda_1}^2 + \sigma_u^2$, we obtain consistent estimates of $\sigma_{\alpha_1}^2$ and $\sigma_{\lambda_1}^2$, respectively.

Substituting consistently estimated values of $\boldsymbol{\sigma}_{\alpha}^2$, $\boldsymbol{\sigma}_{\lambda}^2$, and $\sigma_u^2$ into (6.3.11), one can show that when $N$ and $T$ both tend to infinity and $N/T$ tends to a nonzero constant, the two-stage Aitken estimator is asymptotically as efficient as if one knew the true $\Omega$. Also, Kelejian and Stephan (1983) have pointed out that contrary to the conventional regression model, the speed of convergence of $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}$ here is not $(NT)^{1/2}$, but max $(N^{1/2}, T^{1/2})$.

If one is interested in predicting the random components associated with an individual, Lee and Griffiths (1979) have shown that the predictor

$$\hat{\boldsymbol{\alpha}} = (I_N \otimes \Delta)X'\Omega^{-1}(\mathbf{y} - X\hat{\boldsymbol{\beta}}_{\mathrm{GLS}}) \tag{6.3.20}$$

is the BLUE.

To test for the random variation of the coefficients, we can again apply the Breusch and Pagan (1979) Lagrangian-multiplier test for heteroscedasticity. Because for given $i$, $\boldsymbol{\alpha}_i$ is fixed, the error term $\mathbf{x}_{it}'\boldsymbol{\lambda}_t + u_{it}$ will be homoscedastic if the coefficients are not varying over time. Therefore, under the null, one-half the explained sum of squares in a regression[11]

$$\begin{aligned}\frac{\hat{u}_{it}^2}{\hat{\sigma}_u^2} &= \dot{\mathbf{x}}_{it}'\boldsymbol{\sigma}_{\lambda}^2 + \epsilon_{it}, \qquad i = 1, \ldots, N,\\ &\qquad t = 1, \ldots, T,\end{aligned} \tag{6.3.21}$$

[11] Let $(y_{it} - y)$ be the deviation of the sample mean, and let $(\widehat{y_{it} - y})$ be its least-squares prediction. Then the explained sum of squares is $\Sigma(\widehat{y_{it} - y})^2$.

is distributed asymptotically as $\chi^2$ with $K-1$ degrees of freedom, where $\hat{u}_{it} = y_{it} - \hat{\boldsymbol{\beta}}_i' \mathbf{x}_{it}$, $\hat{\sigma}_u^2 = \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \hat{\boldsymbol{\beta}}_i' \mathbf{x}_{it})^2 / NT$, and $\dot{\mathbf{x}}_{it}$ is $\mathbf{x}_{it}$ with each element squared.[12]

Similarly, we can test for random variation across cross-sectional units by regressing

$$\frac{\hat{u}_{it}^{*2}}{\hat{\sigma}_u^{*2}} = \dot{\mathbf{x}}_{it}' \boldsymbol{\sigma}_\alpha^2 + \epsilon_{it}^*, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T, \tag{6.3.22}$$

where $\hat{u}_{it}^* = y_{it} - \hat{\boldsymbol{\beta}}_t' \mathbf{x}_{it}$, $\hat{\sigma}_u^{*2} = \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{u}_{it}^{*2} / NT$, and $\hat{\boldsymbol{\beta}}_t$ is the least-squares estimate of $\boldsymbol{\beta}_t = \boldsymbol{\beta} + \boldsymbol{\lambda}_t$ across cross-sectional units for a given $t$. Under the null hypothesis of no random variation across cross-sectional units, one-half of the explained sum of squares of (6.3.22) is asymptotically $\chi^2$ distributed with $K-1$ degrees of freedom.

We can also test the random variation indirectly by applying the classic ANCOVA test. For details, see Hsiao (1974a).

Swamy and Mehta (1977) have proposed a more general type of time-varying-component model by allowing $E\boldsymbol{\lambda}_t \boldsymbol{\lambda}_t' = \Lambda_t$ to vary over $t$. However, models with the coefficients varying randomly across cross-sectional units and over time have not gained much acceptance in empirical investigations. Part of the reason is because the inversion of $\Omega$ is at least of order max $(NK, TK)$ (Hsiao 1974a). For any panel data of reasonable size, this would be a computationally demanding problem.

## 6.4 COEFFICIENTS THAT EVOLVE OVER TIME

### 6.4.1 The Model

There is a large amount of empirical evidence that parameters of a model change over time. For instance, financial liberalization or changes in monetary policy can cause the relationships between economic variables to alter. If a constant-parameter model is used, misspecification may occur. On the other hand, if a model is too flexible in its treatment of parameter change, over-fitting or imprecise inferences can occur. In this section, we discuss some commonly used time-varying-parameter models that entail a smooth evolution.[13]

In most models with coefficients evolving over time it is assumed that there is no individual heterogeneity (e.g., Zellner, Hong, and Min 1991). At a given $t$, the coefficient vectors $\boldsymbol{\beta}_t$ are identical for all cross-sectional units. For this reason we shall discuss the main issues of time-varying-parameter models assuming that $N = 1$, and then indicate how this analysis can be modified when $N > 1$.

[12] Note here that the first term $\dot{x}_{1it} = 1$. So the null hypothesis is $(\sigma_{\lambda 2}^2, \ldots, \sigma_{\lambda K}^2) = (0, \ldots, 0)$.

[13] This section is largely drawn from the work of Chow (1983, their Chapter 10).

As shown by Chow (1983, Chapter 10), a wide variety of time-varying-parameter models can be put in the general form

$$y_t = \boldsymbol{\beta}_t' \mathbf{x}_t + u_t, \tag{6.4.1}$$

and

$$\boldsymbol{\beta}_t = H\boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t, \quad t = 1\ldots, T, \tag{6.4.2}$$

where $\mathbf{x}_t$ is a $K \times 1$ vector of exogenous variables; $u_t$ is independent normal, with mean 0 and variance $\sigma_u^2$; $\boldsymbol{\eta}_t$ is a $K$-variant independent normal random variable, with mean 0 and covariance matrix $\Psi$; and $\boldsymbol{\eta}$ and $u$ are independent. For instance, when $H = I_K$, it is the random-walk model of Cooley and Prescott (1976). When $H = I_K$ and $\Psi = \mathbf{0}$, this model is reduced to the standard regression model.

The Rosenberg (1972, 1973) return-to-normality model can also be put into this form. The model corresponds to replacing $\boldsymbol{\beta}_t$ and $\boldsymbol{\beta}_{t-1}$ in (6.4.2) by $(\boldsymbol{\beta}_t - \boldsymbol{\beta})$ and $(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta})$ and restricting the absolute value of the characteristic roots of $H$ to $<1$. Although this somewhat changes the formulation, if we define $\boldsymbol{\beta}_t^* = \boldsymbol{\beta}_t - \boldsymbol{\beta}$ and $\bar{\boldsymbol{\beta}}_t = \boldsymbol{\beta}$, the return-to-normality model can be rewritten as

$$\begin{aligned} y_t &= (\mathbf{x}_t', \mathbf{x}_t') \begin{bmatrix} \bar{\boldsymbol{\beta}}_t \\ \boldsymbol{\beta}_t^* \end{bmatrix} + u_t \\ \begin{bmatrix} \bar{\boldsymbol{\beta}}_t \\ \boldsymbol{\beta}_t^* \end{bmatrix} &= \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & H \end{bmatrix} \begin{bmatrix} \bar{\boldsymbol{\beta}}_{t-1} \\ \boldsymbol{\beta}_{t-1}^* \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \boldsymbol{\eta}_t \end{bmatrix}, \end{aligned} \tag{6.4.3}$$

which is a special case of (6.4.1) and (6.4.2).

Similarly, we can allow $\boldsymbol{\beta}_t$ to be stationary, with constant mean $\boldsymbol{\beta}$ (Pagan 1980). Suppose

$$\begin{aligned} y_t &= \mathbf{x}_t'\boldsymbol{\beta} + \mathbf{x}_t'\boldsymbol{\beta}_t^* + u_t, \\ \boldsymbol{\beta}_t^* &= \boldsymbol{\beta}_t - \boldsymbol{\beta} = A^{-1}(\mathcal{L})\boldsymbol{\epsilon}_t, \end{aligned} \tag{6.4.4}$$

where $A(\mathcal{L})$ is a ratio of polynomials of orders $p$ and $q$ in the lag operator $\mathcal{L}(\mathcal{L}\boldsymbol{\epsilon}_t = \boldsymbol{\epsilon}_{t-1})$, and $\boldsymbol{\epsilon}$ is independent normal, so that $\boldsymbol{\beta}_t^*$ follows an autoregressive moving-average (ARMA) $(p, q)$ process. Because an ARMA of order $p$ and $q$ can be written as a first-order autoregressive process, this model can again be put in the form of (6.4.1) and (6.4.2). For example,

$$\boldsymbol{\beta}_t^* = B_1\boldsymbol{\beta}_{t-1}^* + B_2\boldsymbol{\beta}_{t-2}^* + \boldsymbol{\epsilon}_t + B_3\boldsymbol{\epsilon}_{t-1} \tag{6.4.5}$$

can be written as

$$\tilde{\boldsymbol{\beta}}_t^* = \begin{bmatrix} \boldsymbol{\beta}_t^* \\ \boldsymbol{\beta}_{t-1}^* \\ \boldsymbol{\epsilon}_t \end{bmatrix} = \begin{bmatrix} B_1 & B_2 & B_3 \\ I & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{t-1}^* \\ \boldsymbol{\beta}_{t-2}^* \\ \boldsymbol{\epsilon}_{t-1} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_t \\ \mathbf{0} \\ \boldsymbol{\epsilon}_t \end{bmatrix} = H\tilde{\boldsymbol{\beta}}_{t-1}^* + \boldsymbol{\eta}_t. \tag{6.4.6}$$

Thus, we can write Pagan's model in the form

$$y_t = (\mathbf{x}_t', \tilde{\mathbf{x}}_t') \begin{bmatrix} \bar{\boldsymbol{\beta}}_t \\ \tilde{\boldsymbol{\beta}}_t^* \end{bmatrix} + u_t, \tag{6.4.4a}$$

where $\tilde{\mathbf{x}}_t' = (\mathbf{x}_t', \mathbf{0}', \mathbf{0}')$. Equation (6.4.4a) is then formally equivalent to (6.4.3).

The Kalman filter (Kalman 1960) provides a basis for computing the maximum-likelihood estimators and predicting the evolution of the time path of $\boldsymbol{\beta}_t$ for this type of the model. In this section we first consider the problem of estimating $\boldsymbol{\beta}_t$ using information $\mathcal{I}_s$, up to the time $s$, assuming that $\sigma_u^2$, $\Psi$, and $H$ are known. We denote the conditional expectation of $\boldsymbol{\beta}_t$, given $\mathcal{I}_s$, as $E(\boldsymbol{\beta}_t \mid \mathcal{I}_s) = \boldsymbol{\beta}_{t|s}$. The evaluation of $\boldsymbol{\beta}_{t|s}$ is called filtering when $t = s$; it is called smoothing when $s > t$; it is called prediction when $s < t$. We then study the problem of estimating $\sigma_u^2$, $\Psi$, and $H$ by the method of maximum likelihood. Finally, we consider the problem of testing for constancy of the parameters.

### 6.4.2 Predicting $\boldsymbol{\beta}_t$ by the Kalman Filter

Denote $(y_1, \ldots, y_t)$ by $Y_t$. By definition, the conditional mean of $\boldsymbol{\beta}_t$, given $Y_t$, is

$$\begin{aligned} \boldsymbol{\beta}_{t|t} &= E(\boldsymbol{\beta}_t \mid y_t, Y_{t-1}) \\ &= E(\boldsymbol{\beta}_t \mid Y_{t-1}) + L_t[y_t - E(y_t \mid Y_{t-1})] \\ &= \boldsymbol{\beta}_{t|t-1} + L_t[y_t - \mathbf{x}_t'\boldsymbol{\beta}_{t|t-1}]. \end{aligned} \tag{6.4.7}$$

where $y_t - E(y_t \mid Y_{t-1})$ denotes the additional information of $y_t$ not contained in $Y_{t-1}$ and $L_t$ denotes the adjustment factor of $\boldsymbol{\beta}_{t|t-1}$ because of this additional information. If $L_t$ is known, (6.4.7) can be used to update our estimate $\boldsymbol{\beta}_{t|t-1}$ to form $\boldsymbol{\beta}_{t|t}$.

To derive $L_t$, we know from our assumption on $\boldsymbol{\eta}_t$ and $u_t$ that, conditional on $\mathbf{x}_t$, $y_t$ and $\boldsymbol{\beta}_t$ are jointly normally distributed. The normal-distribution theory (Anderson 1985, Chapter 2) states that, conditional on $Y_{t-1}$ (and $X_t$), the mean of $\boldsymbol{\beta}_t$, given $y_t$ is $E(\boldsymbol{\beta}_t \mid Y_{t-1}) + \text{Cov}(\boldsymbol{\beta}_t, y_t \mid Y_{t-1})\text{Var}(y_t \mid Y_{t-1})^{-1}[y_t - E(y_t \mid Y_{t-1})]$. Therefore,

$$L_t = [E(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t-1})(y_t - y_{t|t-1})]\text{Var}(y_t \mid Y_{t-1})^{-1}, \tag{6.4.8}$$

where $y_{t|t-1} = E(y_t \mid Y_{t-1}) = \mathbf{x}_t'\boldsymbol{\beta}_{t|t-1}$. Denoting the covariance matrix $\text{Cov}(\boldsymbol{\beta}_t \mid Y_{t-1}) = E(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t-1})(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t-1})'$ by $\Sigma_{t|t-1}$, we have

$$\begin{aligned} &E(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t-1})(y_t - y_{t|t-1}) \\ &\quad = E\{(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t-1})[(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t-1})'\mathbf{x}_t + u_t]\} = \Sigma_{t|t-1}\mathbf{x}_t, \end{aligned} \tag{6.4.9}$$

and

$$\begin{aligned} \text{Var}(y_t \mid Y_{t-1}) &= E[\mathbf{x}_t'(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t-1}) + u_t][(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t-1})'\mathbf{x}_t + u_t] \\ &= \mathbf{x}_t'\Sigma_{t|t-1}\mathbf{x}_t + \sigma_u^2. \end{aligned} \tag{6.4.10}$$

Hence, (6.4.8) becomes

$$L_t = \Sigma_{t|t-1}\mathbf{x}_t(\mathbf{x}_t'\Sigma_{t|t-1}\mathbf{x}_t + \sigma_u^2)^{-1}. \tag{6.4.11}$$

From (6.4.2) we have

$$\boldsymbol{\beta}_{t|t-1} = H\boldsymbol{\beta}_{t-1|t-1}. \tag{6.4.12}$$

Thus, we can compute $\Sigma_{t|t-1}$ recursively by

$$\begin{aligned}\Sigma_{t|t-1} &= E(\boldsymbol{\beta}_t - H\boldsymbol{\beta}_{t-1|t-1})(\boldsymbol{\beta}_t - H\boldsymbol{\beta}_{t-1|t-1})' \\ &= E[H(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}_{t-1|t-1}) + \boldsymbol{\eta}_t] \\ &\quad \cdot [H(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}_{t-1|t-1}) + \boldsymbol{\eta}_t]' \\ &= H\Sigma_{t-1|t-1}H' + \Psi.\end{aligned} \tag{6.4.13}$$

Next, from (6.4.1) and (6.4.7) we can write

$$\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t} = \boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t-1} - L_t[\mathbf{x}_t'(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t-1}) + u_t]. \tag{6.4.14}$$

Taking the expectation of the product of (6.4.14) and itstranspose, and using (6.4.11), we obtain

$$\begin{aligned}\Sigma_{t|t} &= \Sigma_{t|t-1} - L_t(\mathbf{x}_t'\Sigma_{t|t-1}\mathbf{x}_t + \sigma_u^2)L_t' \\ &= \Sigma_{t|t-1} - \Sigma_{t|t-1}\mathbf{x}_t(\mathbf{x}_t'\Sigma_{t|t-1}\mathbf{x}_t + \sigma_u^2)^{-1}\mathbf{x}_t'\Sigma_{t|t-1}.\end{aligned} \tag{6.4.15}$$

Equations (6.4.13) and (6.4.15) can be used to compute $\Sigma_{t|t}(t = 1, 2, \ldots)$ successively, given $\Sigma_{0|0}$. Having computed $\Sigma_{t|t-1}$, we can use (6.4.11) to compute $L_t$. Given $L_t$, (6.4.7) and (6.4.12) can be used to compute $\boldsymbol{\beta}_{t|t}$ from $\boldsymbol{\beta}_{t-1|t-1}$ if $\boldsymbol{\beta}_{0|0}$ is known.

Similarly, we can predict $\boldsymbol{\beta}_t$ using future observations $y_{t+1}, y_{t+2}, \ldots, y_{t+n}$. We first consider the regression of $\boldsymbol{\beta}_t$ on $y_{t+1}$, conditional on $Y_t$. Analogous to (6.4.7) and (6.4.11) are

$$\boldsymbol{\beta}_{t|t+1} = \boldsymbol{\beta}_{t|t} + F_{t|t+1}(y_{t+1} - y_{t+1|t}) \tag{6.4.16}$$

and

$$F_{t|t+1} = [E(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t})(y_{t+1} - y_{t+1|t})'][\mathrm{Cov}(y_{t+1} \mid Y_t)]^{-1}. \tag{6.4.17}$$

To derive the matrix $F_{t|t+1}$ of regression coefficients, we use (6.4.1) and (6.4.2) to write

$$\begin{aligned}y_{t+1} - y_{t+1|t} &= \mathbf{x}_{t+1}'(\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_{t+1|t}) + u_{t+1} \\ &= \mathbf{x}_{t+1}'H(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t}) + \mathbf{x}_{t+1}'\boldsymbol{\eta}_{t+1} + u_{t+1}.\end{aligned} \tag{6.4.18}$$

Combining (6.4.17), (6.4.18), (6.4.10), and (6.4.11), we have

$$\begin{aligned}F_{t|t+1} &= \Sigma_{t|t}H'\mathbf{x}_{t+1}(\mathbf{x}_{t+1}'\Sigma_{t+1|t}\mathbf{x}_{t+1} + \sigma_u^2)^{-1} \\ &= \Sigma_{t|t}H'\Sigma_{t+1|t}^{-1}L_{t+1}.\end{aligned} \tag{6.4.19}$$

Therefore, from (6.4.19) and (6.4.14), we can rewrite (6.4.16) as

$$\boldsymbol{\beta}_{t|t+1} = \boldsymbol{\beta}_{t|t} + \Sigma_{t|t} H' \Sigma_{t+1|t}^{-1} (\boldsymbol{\beta}_{t+1|t+1} - \boldsymbol{\beta}_{t+1|t}). \tag{6.4.20}$$

Equation (6.4.20) can be generalized to predict $\boldsymbol{\beta}_t$ using future observations $y_{t+1}, \ldots, y_{t+n}$,

$$\boldsymbol{\beta}_{t|t+n} = \boldsymbol{\beta}_{t|t+n-1} + F_t^* (\boldsymbol{\beta}_{t+1|t+n} - \boldsymbol{\beta}_{t+1|t+n-1}), \tag{6.4.21}$$

where $F_t^* = \Sigma_{t|t} H' \Sigma_{t+1|t}^{-1}$. The proof of this is given by Chow (1983, Chapter 10).

When $H$, $\Psi$, and $\sigma_u^2$ are known, (6.4.7) and (6.4.21) trace out the time path of $\boldsymbol{\beta}_t$ and provide the minimum-mean-square-error forecast of the future values of the dependent variable, given the initial values $\boldsymbol{\beta}_{0|0}$ and $\Sigma_{0|0}$. To obtain the initial values of $\boldsymbol{\beta}_{0|0}$ and $\Sigma_{0|0}$, Sant (1977) suggested using the GLS method on the first $K$ observations of $y_t$ and $\mathbf{x}_t$. Noting that

$$\begin{aligned} \boldsymbol{\beta}_t &= H\boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t \\ &= H^2 \boldsymbol{\beta}_{t-2} + \boldsymbol{\eta}_t + H\boldsymbol{\eta}_{t-1} \\ &= H^{t-j} \boldsymbol{\beta}_j + \boldsymbol{\eta}_t + H\boldsymbol{\eta}_{t-1} + \cdots + H^{t-j-1} \boldsymbol{\eta}_j, \end{aligned} \tag{6.4.22}$$

and assuming that $H^{-1}$ exists, we can also write $y_k$ in the form

$$\begin{aligned} y_k &= \mathbf{x}_k' \boldsymbol{\beta}_k + u_k \\ &= \mathbf{x}_k' [H^{-K+k} \boldsymbol{\beta}_K - H^{-K+k} \boldsymbol{\eta}_K - \cdots - H^{-1} \boldsymbol{\eta}_{k+1}] + u_k. \end{aligned}$$

Thus, $(y_1, \ldots, y_K)$ can be written as

$$\begin{aligned} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_K \end{bmatrix} &= \begin{bmatrix} \mathbf{x}_1' H^{-K+1} \\ \mathbf{x}_2' H^{-K+2} \\ \vdots \\ \mathbf{x}_K' \end{bmatrix} \boldsymbol{\beta}_K + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_K \end{bmatrix} \\ &\quad - \begin{bmatrix} \mathbf{x}_1' H^{-1} & \mathbf{x}_1' H^{-2} & \cdots & \mathbf{x}_1' H^{-K+1} \\ \mathbf{0}' & \mathbf{x}_2' H^{-1} & \cdots & \mathbf{x}_2' H^{-K+2} \\ & & \cdots & \cdot \\ & & & \mathbf{x}_{K-1}' H^{-1} \\ & & & \mathbf{0}' \end{bmatrix} \begin{bmatrix} \boldsymbol{\eta}_2 \\ \boldsymbol{\eta}_3 \\ \vdots \\ \boldsymbol{\eta}_K \end{bmatrix}. \end{aligned} \tag{6.4.23}$$

Applying GLS to (6.4.23) gives

$$\begin{aligned} \Sigma_{K|K} = \sigma_u^2 \{ &[H'^{-K+1} \mathbf{x}_1, H'^{-K+2} \mathbf{x}_2, \ldots, \mathbf{x}_K] \\ &\cdot [I_K + A_K (I_{K-1} \otimes P) A_K']^{-1} [H^{-K+1} \mathbf{x}_1, \ldots, \mathbf{x}_K]' \}^{-1} \end{aligned} \tag{6.4.24}$$

and

$$\boldsymbol{\beta}_{K|K} = \frac{1}{\sigma_u^2}\Sigma_{K|K}[H'^{-K+1}\mathbf{x}_1, H'^{-K+2}\mathbf{x}_2, \ldots, \mathbf{x}_K]$$
$$[I_K + A_K(I_{K-1} \otimes P)A_K']^{-1}\begin{bmatrix} y_1 \\ \vdots \\ y_K \end{bmatrix}, \tag{6.4.25}$$

where $P = \sigma_u^{-2}\Psi$, and $A_K$ is the coefficient matrix of $(\boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_K)'$ in (6.4.23). The initial estimators, $\boldsymbol{\beta}_{K|K}$ and $\Sigma_{K|K}$, are functions of $\sigma_u^2$, $\Psi$, and $H$.

### 6.4.3 Maximum-Likelihood Estimation

When $H$, $\Psi$, and $\sigma_u^2$ are unknown, (6.4.7) opens the way for maximum-likelihood estimation without the need for repeated inversions of covariance matrices of large dimensions. To form the likelihood function, we note that

$$y_t - y_{t|t-1} = \mathbf{x}_t'(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t-1}) + u_t = y_t - x_t'\boldsymbol{\beta}_{t|t-1} \tag{6.4.26}$$

is normal and serially uncorrelated. Hence, the joint density of $(y_1, \ldots, y_T)$ can be written as the product of the conditional density of $(y_{K+1}, \ldots, y_T \mid y_1, \ldots, y_K)$ and the marginal density of $(y_1, \ldots, y_K)$. The log-likelihood function of $(y_{K+1}, \ldots, y_T)$, given $(y_1, \ldots, y_K)$, is

$$\log L = -\frac{T-K}{2}\log 2\pi - \frac{1}{2}\sum_{t=K+1}^{T}\log(\mathbf{x}_t'\Sigma_{t|t-1}\mathbf{x}_t + \sigma_u^2)$$
$$-\frac{1}{2}\sum_{t=K+1}^{T}\frac{(y_t - \mathbf{x}_t'\boldsymbol{\beta}_{t|t-1})^2}{\mathbf{x}_t'\Sigma_{t|t-1}\mathbf{x}_t + \sigma_u^2}. \tag{6.4.27}$$

The first $K$ observations are used to compute $\Sigma_{K|K}$ and $\boldsymbol{\beta}_{K|K}$ [equations (6.4.24) and (6.4.25)] as functions of $\sigma_u^2$, $\Psi$, and $H$. Hence, the data $\boldsymbol{\beta}_{t|t-1}$ and $\Sigma_{t|t-1}$ $(t = K+1, \ldots, T)$ required to evaluate $\log L$ are functions of $\sigma_u^2$, $\Psi$, and $H$, as given by (6.4.13), (6.4.15), (6.4.12), and (6.4.11). To find the maximum of (6.4.27), numerical methods will have to be used.

When we estimate the model (6.4.1) and (6.4.2) using panel data, all the derivations in Section 6.4.2 remain valid if we replace $y_t$, $\mathbf{x}_t$, $\mathbf{u}_t$, and $\sigma_u^2$ by the $N \times 1$ vector $\mathbf{y}_t = (y_{1t}, \ldots, y_{Nt})'$, the $N \times K$ matrix $X_t = (\mathbf{x}_{1t}, \ldots, \mathbf{x}_{Nt})'$, the $N \times 1$ vector $\mathbf{u}_t = (u_{1t}, \ldots, u_{Nt})'$, and $\sigma_u^2 I_N$ in appropriate places. The MLE can be carried out in the same way as outlined in this section, except that

the likelihood function (6.4.27) is replaced by

$$\begin{aligned}\text{Log } L = \text{const } &- \frac{1}{2}\sum_t \log \mid X_t'\Sigma_{t|t-1}X_t + \sigma_u^2 I_N \mid \\ &- \frac{1}{2}\sum_t (\mathbf{y}_t - X_t\boldsymbol{\beta}_{t|t-1})' \\ &\cdot (X_t\Sigma_{t|t-1}X_t' + \sigma_u^2 I_N)^{-1}(\mathbf{y}_t - X_t\boldsymbol{\beta}_{t|t-1}).\end{aligned} \tag{6.4.27′}$$

However, we no longer need to use the first $K$ period observations to start the iteration. If $N > K$, we need to use only the first-period cross-sectional data to obtain $\boldsymbol{\beta}_{1|1}$ and $\Sigma_{1|1}$. Additional details with regard to the computation can be found in Harvey (1978) and Harvey and Phillips (1982).

### 6.4.4 Tests for Parameter Constancy

A simple alternative to the null hypothesis of constancy of regression coefficients over time is

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t, \tag{6.4.28}$$

where $\boldsymbol{\eta}_t$ is assumed independently normally distributed, with mean 0 and a diagonal covariance matrix $\Psi$. Regarding $\boldsymbol{\beta}_0$ as fixed, we have

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_0 + \sum_{s=1}^{t} \boldsymbol{\eta}_s. \tag{6.4.29}$$

Thus, the regression model becomes

$$\begin{aligned} y_t = \mathbf{x}_t'\boldsymbol{\beta}_t + u_t &= \mathbf{x}_t'\boldsymbol{\beta}_0 + u_t + \mathbf{x}_t'\left(\sum_{s=1}^{t}\boldsymbol{\eta}_s\right) \\ &= \mathbf{x}_t'\boldsymbol{\beta}_0 + u_t^*, \end{aligned} \tag{6.4.30}$$

where $u_t^* = u_t + \mathbf{x}_t'(\sum_{s=1}^{t}\boldsymbol{\eta}_s)$ has variance

$$Eu_t^{*2} = \sigma_u^2 + t\mathbf{x}_t'\Psi\mathbf{x}_t. \tag{6.4.31}$$

For $\Psi = \text{diag}\{\psi_{kk}\}$, (6.4.31) becomes

$$Eu_t^{*2} = \sigma_u^2 + t\sum_{k=1}^{K} x_{kt}^2\psi_{kk}, \quad t = 1\ldots., T. \tag{6.4.32}$$

The null hypothesis states that $\Psi = \mathbf{0}$. Hence, the Breusch and Pagan (1979) Lagrangian-multiplier test applied here is to regress $\hat{u}_t^2/\hat{\sigma}_u^2$ on $t(1, x_{2t}^2, \ldots, x_{Kt}^2)$, $t = 1, \ldots, T$, where $\hat{u}_t$ is the least-squares residual $\hat{u}_t = y_t - \hat{\boldsymbol{\beta}}'\mathbf{x}_t$, $\hat{\boldsymbol{\beta}} = (\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t')^{-1}(\sum_{t=1}^{T}\mathbf{x}_t y_t)$, and $\hat{\sigma}_u^2 = \sum_{t=1}^{T}\hat{u}_t^2/T$. Under the

null hypothesis, one-half the explained sum of squares of this regression is asymptotically $\chi^2$ distributed, with $K$ degrees of freedom.[14]

When panel data are available, it is possible to test for parameter constancy indirectly using the classic ANCOVA test. By the assumption that the parameter vector $\boldsymbol{\beta}_t$, is constant over cross-sectional units in the same period, an indirect test is to postulate the null hypothesis,

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_T = \boldsymbol{\beta}.$$

If the disturbances of the regression model $y_{it} = \boldsymbol{\beta}_t' \mathbf{x}_{it} + u_{it}$ are independently normally distributed over $i$ and $t$, then the test statistic $F_3'$ from Chapter 2 has an $F$ distribution with $(T-1)K$ and $N(T-K)$ degrees of freedom under the null.

If the null hypothesis is rejected, we can use the information that under mild regularity conditions $\text{plim}_{N\to\infty}\hat{\boldsymbol{\beta}}_t = \boldsymbol{\beta}_t$, $t = 1, \ldots, T$, to investigate the nature of variation in the parameters over time. We can apply the Box–Jenkins (1970) method on $\hat{\boldsymbol{\beta}}_t$ to identify a suitable stochastic process with which to model the parameter variation.

## 6.5 COEFFICIENTS THAT ARE FUNCTIONS OF OTHER EXOGENOUS VARIABLES

Sometimes, instead of assuming that parameters are random draws from a common distribution, an investigation of possible dependence of $\boldsymbol{\beta}_{it}$ on characteristics of the "individuals" or "time" is of considerable interest (e.g., Amemiya 1978b; Hendricks, Koenker, and Poirier 1979; Singh et al. 1976; Swamy and Tinsley 1977; Wachter 1970). A general formulation of stochastic-parameter models with systematic components can be expressed within the context of the linear model. Suppose that

$$\mathbf{y}_i = X_{i1}\boldsymbol{\beta}_1 + X_{i2}\boldsymbol{\beta}_{2i} + \mathbf{u}_i, \quad i = 1\ldots., N, \tag{6.5.1}$$

and

$$\boldsymbol{\beta}_{2i} = Z_i\boldsymbol{\gamma} + \boldsymbol{\eta}_{2i} \tag{6.5.2}$$

where $X_{i1}$ and $X_{i2}$ denote the $T \times K_1$ and $T \times K_2$ matrices of the time-series observations of the first $K_1$ and last $K_2 (= K - K_1)$ exogenous variables for the $i$th individual, $\boldsymbol{\beta}_1$ is a $K_1 \times 1$ vector of fixed constants, $\boldsymbol{\beta}_{2i}$ is a $K_2 \times 1$ vector that varies according to (6.5.2); $Z_i$ and $\boldsymbol{\gamma}$ are a $K_2 \times M$ matrix of known constants and a $M \times 1$ vector of unknown constants, respectively; and $\mathbf{u}_i$ and $\boldsymbol{\eta}_{2i}$ are $T \times 1$ and $K_2 \times 1$ vectors of unobservable random variables that are assumed independent of $X_i$ and $Z_i$. For example, in Wachter (1970), $\mathbf{y}_i$ is a vector of time-series observations on the logarithm of the relative wage rate in the $i$th industry. $X_{i1}$ contains the logarithm of such variables as the

[14] Note that under the alternative, $u_t^*$ is serially correlated. Hence, the Breusch and Pagan test may not be powerful against the alternative.

relative value-added in the $i$th industry and the change in the consumer price, $X_{i2}$ consists of a single vector of time series observations on the logarithm of unemployment, and $Z_i$ contains the degree of concentration and the degree of unionization in the $i$th industry.

For simplicity, we assume that $\mathbf{u}_i$ and $\boldsymbol{\eta}_{2i}$ are uncorrelated with each other and have 0 means. The variance–covariance matrices of $\mathbf{u}_i$ and $\boldsymbol{\eta}_{2i}$ are given by

$$E\mathbf{u}_i\mathbf{u}_j' = \sigma_{ij} I_T \tag{6.5.3}$$

and

$$E\boldsymbol{\eta}_{2i}\boldsymbol{\eta}_{2j}' = \begin{cases} \Lambda & \text{if } i = j, \\ \mathbf{0} & \text{if } i \neq j. \end{cases} \tag{6.5.4}$$

Let $\Sigma = (\sigma_{ij})$. We can write the variance–covariance of $\mathbf{u} = (\mathbf{u}_1', \ldots, \mathbf{u}_N')'$ and $\boldsymbol{\eta}_2 = (\boldsymbol{\eta}_{21}', \ldots, \boldsymbol{\eta}_{2N}')'$ as

$$E\mathbf{u}\mathbf{u}' = \Sigma \otimes I_T \tag{6.5.5}$$

and

$$E\boldsymbol{\eta}_2\boldsymbol{\eta}_2' = \begin{bmatrix} \Lambda & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \Lambda \end{bmatrix} = \tilde{\Lambda}. \tag{6.5.6}$$

Combining (6.5.1) and (6.5.2), we have

$$\mathbf{y} = X_1\boldsymbol{\beta}_1 + \mathbf{W}\boldsymbol{\gamma} + \tilde{X}_2\boldsymbol{\eta}_2 + \mathbf{u}, \tag{6.5.7}$$

where

$$\underset{NT\times 1}{\mathbf{y}} = (y_1', \ldots, \mathbf{y}_N')',$$

$$\underset{NT\times K_1}{X_1} = (X_{11}', \ldots, X_{N1}')',$$

$$\underset{NT\times M}{\mathbf{W}} = (Z_1'X_{12}', Z_2'X_{22}', \ldots, Z_N'X_{N2}')',$$

$$\underset{NT\times NK_2}{\tilde{X}_2} = \begin{bmatrix} X_{12} & & & \mathbf{0} \\ & X_{22} & & \\ & & \ddots & \\ \mathbf{0} & & & X_{N2} \end{bmatrix},$$

and

$$\underset{NK_2\times 1}{\boldsymbol{\eta}_2} = (\boldsymbol{\eta}_{21}', \ldots, \boldsymbol{\eta}_{2N}')'.$$

The BLUE of $\boldsymbol{\beta}_1$ and $\boldsymbol{\gamma}$ of (6.5.7) is the GLS estimator.

$$\begin{aligned}\begin{bmatrix}\hat{\boldsymbol{\beta}}_1\\ \hat{\boldsymbol{\gamma}}\end{bmatrix}_{\text{GLS}} &= \left\{\begin{bmatrix}X_1'\\ W'\end{bmatrix}[\Sigma\otimes I_T+\tilde{X}_2\tilde{\Lambda}\tilde{X}_2']^{-1}(X_1, W)\right\}^{-1}\\ &\quad\cdot\left\{\begin{bmatrix}X_1'\\ W'\end{bmatrix}[\Sigma\otimes I_T+\tilde{X}_2\tilde{\Lambda}\tilde{X}_2']^{-1}\mathbf{y}\right\}.\end{aligned} \tag{6.5.8}$$

If $\Sigma$ is diagonal, the variance–covariance matrix of the stochastic term of (6.5.7) is block-diagonal, with the $i$th diagonal block equal to

$$\Omega_i = X_{i2}\Lambda X_{i2}' + \sigma_{ii}I_T. \tag{6.5.9}$$

The GLS estimator (6.5.8) can be simplified as

$$\begin{aligned}\begin{bmatrix}\hat{\boldsymbol{\beta}}_1\\ \hat{\boldsymbol{\gamma}}\end{bmatrix}_{\text{GLS}} &= \left[\sum_{i=1}^{N}\begin{bmatrix}X_{i1}'\\ Z_i'X_{i2}'\end{bmatrix}\Omega_i^{-1}(X_{i1}, X_{2i}Z_i)\right]^{-1}\\ &\quad\cdot\left[\sum_{i=1}^{N}\begin{bmatrix}X_{i1}'\\ Z_i'X_{i2}'\end{bmatrix}\Omega_i^{-1}\mathbf{y}_i\right].\end{aligned} \tag{6.5.10}$$

Amemiya (1978b) suggested estimating $\Lambda$ and $\sigma_{ij}$ as follows. Let

$$\begin{aligned}\begin{bmatrix}\mathbf{y}_1\\ \vdots\\ \mathbf{y}_N\end{bmatrix} &= \begin{bmatrix}X_{11}\\ \vdots\\ X_{N1}\end{bmatrix}\boldsymbol{\beta}_1+\begin{bmatrix}X_{12}\\ \mathbf{0}\\ \vdots\\ \mathbf{0}\end{bmatrix}\boldsymbol{\beta}_{21}+\begin{bmatrix}\mathbf{0}\\ X_{22}\\ \vdots\\ \mathbf{0}\end{bmatrix}\boldsymbol{\beta}_{22}\\ &\quad+\cdots+\begin{bmatrix}\mathbf{0}\\ \vdots\\ X_{N2}\end{bmatrix}\boldsymbol{\beta}_{2N}+\begin{bmatrix}\mathbf{u}_1\\ \vdots\\ \mathbf{u}_N\end{bmatrix}.\end{aligned} \tag{6.5.11}$$

Apply the least-squares method to (6.5.11). Denote the resulting estimates by $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_{2i}$, $i = 1, \ldots, N$. Then $\sigma_{ij}$ can be estimated by

$$\hat{\sigma}_{ij} = \frac{1}{T}(\mathbf{y}_i - X_{i1}\hat{\boldsymbol{\beta}}_1 - \mathbf{X}_{i2}\hat{\boldsymbol{\beta}}_{2i})'(\mathbf{y}_j - \mathbf{X}_{j1}\hat{\boldsymbol{\beta}}_1 - \mathbf{X}_{j2}\hat{\boldsymbol{\beta}}_{2j}), \tag{6.5.12}$$

and $\boldsymbol{\gamma}$ can be estimated by

$$\hat{\boldsymbol{\gamma}} = \left(\sum_{i=1}^{N}Z_i'Z_i\right)^{-1}\left(\sum_{i=1}^{N}Z_i'\hat{\boldsymbol{\beta}}_{2i}\right). \tag{6.5.13}$$

We then estimate $\Lambda$ by

$$\hat{\Lambda} = \frac{1}{N}\sum_{i=1}^{N}(\hat{\boldsymbol{\beta}}_{2i} - Z_i\hat{\boldsymbol{\gamma}})(\hat{\boldsymbol{\beta}}_{2i} - Z_i\hat{\boldsymbol{\gamma}})'. \tag{6.5.14}$$

Once consistent estimates of $\sigma_{ij}$ and $\Lambda$ are obtained (as both $N$ and $T$ approach infinity), we can substitute them into (6.5.8). The resulting two-stage Aitken estimator of $(\boldsymbol{\beta}_1', \boldsymbol{\gamma}')$ is consistent and asymptotically normally distributed under general conditions. A test of the hypothesis that $\boldsymbol{\gamma} = 0$ can be performed in the usual regression framework using $\hat{\boldsymbol{\gamma}}'_{\text{GLS}}$ Var $(\hat{\boldsymbol{\gamma}}_{\text{GLS}})^{-1}\hat{\boldsymbol{\gamma}}_{\text{GLS}}$, where

$$\text{Var}(\hat{\boldsymbol{\gamma}}_{\text{GLS}}) = [W'\tilde{\Omega}^{-1}W - W'\tilde{\Omega}^{-1}X_1(X_1'\tilde{\Omega}^{-1}X_1)^{-1}X_1'\tilde{\Omega}^{-1}W]^{-1}, \tag{6.5.15}$$

and

$$\tilde{\Omega} = \tilde{X}_2\tilde{\Lambda}\tilde{X}_2' + \Sigma \otimes I_T.$$

## 6.6 A MIXED FIXED- AND RANDOM-COEFFICIENTS MODEL

### 6.6.1 Model Formulation

Many of the previously discussed models can be considered as special cases of a general mixed fixed- and random-coefficients model. For ease of exposition, we shall assume that only time-invariant cross-sectional heterogeneity exists.

Suppose that each cross-sectional unit is postulated to be different, so that

$$y_{it} = \sum_{k=1}^{K} \beta_{ki}x_{kit} + \sum_{\ell=1}^{m} \gamma_{\ell i}w_{\ell it} + u_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{matrix} \tag{6.6.1}$$

where $\mathbf{x}_{it}$ and $\mathbf{w}_{it}$ are each $K \times 1$ and $m \times 1$ vector of explanatory variables that are independent of the error of the equation, $u_{it}$. Stacking the $NT$ observations together, we have

$$\mathbf{y} = X\boldsymbol{\beta} + W\boldsymbol{\gamma} + \mathbf{u}, \tag{6.6.2}$$

where

$$\underset{NT\times NK}{X} = \begin{pmatrix} X_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & X_2 & \ldots & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & & & X_N \end{pmatrix},$$

$$\underset{NT\times Nm}{W} = \begin{pmatrix} W_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & W_2 & & \\ \vdots & & \ddots & \\ \mathbf{0} & & & W_N \end{pmatrix},$$

$$\underset{NT\times 1}{\mathbf{u}} = (\mathbf{u}_1'.\ldots,\mathbf{u}_N'),$$

$$\underset{NK\times 1}{\boldsymbol{\beta}} = (\boldsymbol{\beta}_1',\ldots,\boldsymbol{\beta}_N')' \quad \text{and} \quad \underset{Nm\times 1}{\boldsymbol{\gamma}} = (\boldsymbol{\gamma}_1',\ldots,\boldsymbol{\gamma}_N').$$

Equation (6.6.1), just like (6.6.2), assumes a different behavioral equation relation for each cross-sectional unit. In this situation, the only advantage of pooling is to put the model (6.6.2) into Zellner's (1962) seemingly unrelated regression framework to gain efficiency of the estimates of the individual behavioral equation.

The motivation of a mixed fixed- and random-coefficients model is that though there may be fundamental differences among cross-sectional units, conditioning on these individual specific effects, one may still be able to draw inferences on certain population characteristics through the imposition of a priori constraints on the coefficients of $\mathbf{x}_{it}$ and $\mathbf{w}_{it}$. We assume that there exist two kinds of restrictions, stochastic and fixed restrictions (e.g., Hsiao 1991a, Hsiao et al. 1993) in the form:

A.6.6.1. The coefficients of $\mathbf{x}_{it}$ are assumed to be subject to stochastic restrictions of the form:

$$\boldsymbol{\beta} = A_1\bar{\boldsymbol{\beta}} + \boldsymbol{\alpha}, \tag{6.6.3}$$

where $A_1$ is an $NK \times L$ matrix with known element, $\bar{\boldsymbol{\beta}}$ is an $L \times 1$ vector of constants, and $\boldsymbol{\alpha}$ is assumed to be (normally distributed) random variables with mean $\mathbf{0}$ and nonsingular constant covariance matrix $C$ and is independent of $\mathbf{x}_{it}$.

A.6.6.2. The coefficients of $\mathbf{w}_{it}$ are assumed to be subject to

$$\boldsymbol{\gamma} = A_2\bar{\boldsymbol{\gamma}}, \tag{6.6.4}$$

where $A_2$ is an $Nm \times n$ matrix with known elements, and $\bar{\boldsymbol{\gamma}}$ is an $n \times 1$ vector of constants.

Since $A_2$ is known, we may substitute (6.6.4) into (6.6.2) and write the model as

$$\mathbf{y} = X\boldsymbol{\beta} + \tilde{W}\bar{\boldsymbol{\gamma}} + \mathbf{u} \tag{6.6.5}$$

subject to (6.6.3), where $\tilde{W} = WA_2$.

A.6.6.2 allows for various possible fixed-parameter configurations. For instance, if $\boldsymbol{\gamma}$ is different across cross-sectional units, we can let $A_2 = I_N \otimes I_m$. On the other hand, if we wish to constrain $\boldsymbol{\gamma}_i = \boldsymbol{\gamma}_j$, we can let $A_2 = \mathbf{e}_N \otimes I_m$.

Many of the linear panel data models with unobserved individual specific, but time-invariant heterogeneity can be treated as the special case of the model (6.6.2)–(6.6.4). These include

(1) A common model for all cross-sectional units. If there is no interindividual difference in behavioral patterns, we may let $X = \mathbf{0}$,

$A_2 = \mathbf{e}_N \otimes I_m$, so (6.6.2) becomes

$$y_{it} = \mathbf{w}_{it}'\bar{\boldsymbol{\gamma}} + u_{it}. \tag{6.6.6}$$

(2) Different models for different cross-sectional units. When each individual is considered different, then $X = \mathbf{0}$, $A_2 = I_N \otimes I_m$, and (6.6.2) becomes

$$y_{it} = \mathbf{w}_{it}'\boldsymbol{\gamma}_i + u_{it}. \tag{6.6.7}$$

(3) Variable intercept model (e.g., Kuh 1963, or Chapter 3, Section 3.2). If conditional on the observed exogenous variables, the interindividual differences stay constant through time. Let $X = \mathbf{0}$, and

$$A_2 = (I_N \otimes \mathbf{i}_m \vdots \mathbf{e}_N \otimes I^*_{m-1}),\ \bar{\boldsymbol{\gamma}} = (\gamma_{11}, \ldots, \gamma_{N1}, \bar{\gamma}_2, \ldots, \bar{\gamma}_m)',$$

where we arrange $W_i = (\mathbf{e}_T, \mathbf{w}_{i2}, \ldots, \mathbf{w}_{im}), i = 1, \ldots, N$. $\mathbf{i}_m = (1, 0, \ldots, 0)'$,

$$\underset{m\times(m-1)}{I^*_{m-1}} = (\mathbf{0} \vdots I_{m-1})',$$

then (6.6.2) becomes

$$y_{it} = \gamma_{i1} + \bar{\gamma}_2 w_{it2} + \cdots + \bar{\gamma}_m w_{itm} + u_{it}. \tag{6.6.8}$$

(4) Error components model (e.g., Balestra and Nerlove 1966; Wallace and Hussain 1969; or Chapter 3, Section 3.3). When the effects of the individual-specific, time-invariant omitted variables are treated as random variables just like the assumption on the effects of other omitted variables, we can let $X_i = \mathbf{e}_T$, $\boldsymbol{\alpha}' = (\alpha_1, \ldots, \alpha_N)$, $A_1 = \mathbf{e}_N$, $C = \sigma_\alpha^2 I_N$, $\bar{\beta}$ be an unknown constant, and $\mathbf{w}_{it}$ not contain an intercept term. Then (6.6.2) becomes

$$y_{it} = \bar{\beta} + \bar{\boldsymbol{\gamma}}'\mathbf{w}_{it} + \alpha_i + u_{it} \tag{6.6.9}$$

(5) Random coefficients model (Swamy 1970, or Chapter 6, Section 6.2.2). Let $Z = \mathbf{0}$, $A_1 = \mathbf{e}_N \otimes I_K$, $C = I_N \otimes \Delta$, we have model (6.2.7).

### 6.6.2 A Bayes Solution

The formulation of (6.6.5) subject to (6.6.3) can be viewed from a Bayesian perspective as there exist informative prior on $\boldsymbol{\beta}$ (6.6.3), but not on $\bar{\boldsymbol{\gamma}}$. In the classical sampling approach, inferences are made by typically assuming that the probability law generating the observations, $\mathbf{y}$, $f(\mathbf{y}, \boldsymbol{\theta})$, is known, but not the vector of constant parameters $\boldsymbol{\theta}$. Estimators $\hat{\boldsymbol{\theta}}(\mathbf{y})$ of the parameters $\boldsymbol{\theta}$ are chosen as functions of $\mathbf{y}$ so that their sampling distributions, in repeated experiments, are, in some sense, concentrated as closely as possible about the true values of $\boldsymbol{\theta}$. In the Bayesian approach, a different line is taken. First, all quantities, including the parameters, are considered random variables. Second, all probability statements are conditional, so that in making a probability statement

it is necessary to refer to the conditioning event as well as the event whose probability is being discussed. Therefore, as part of the model, a prior distribution of the parameter $\boldsymbol{\theta}$, $p(\boldsymbol{\theta})$, is introduced. The prior is supposed to express a state of knowledge (or ignorance) about $\boldsymbol{\theta}$ before the data are obtained. Given the probability model $f(y;\boldsymbol{\theta})$, the prior distribution, and the data $\mathbf{y}$, the probability distribution of $\boldsymbol{\theta}$ is revised to $p(\boldsymbol{\theta} \mid \mathbf{y})$, which is called the posterior distribution of $\boldsymbol{\theta}$, according to Bayes' theorem (e.g., Intriligator, Bodkin, and Hsiao 1996).[15]

$$P(\boldsymbol{\theta} \mid \mathbf{y}) \propto P(\boldsymbol{\theta}) f(\mathbf{y} \mid \boldsymbol{\theta}), \tag{6.6.10}$$

where the sign "$\propto$" denoting "is proportional to," with the factor of proportionality being a normalizing constant.

Under the assumption that

A.6.6.3. $\mathbf{u} \sim N(\mathbf{0}, \Omega)$,

we may write the model (6.6.5) as

A.1 Conditional on $X$, $\tilde{W}$, $\boldsymbol{\beta}$, and $\bar{\boldsymbol{\gamma}}$

$$\mathbf{y} \sim N(X\boldsymbol{\beta} + \tilde{W}\bar{\boldsymbol{\gamma}}, \Omega). \tag{6.6.11}$$

A.2 The prior distributions of $\boldsymbol{\beta}$ and $\bar{\boldsymbol{\gamma}}$ are independent,

$$P(\boldsymbol{\beta}, \bar{\boldsymbol{\gamma}}) = P(\boldsymbol{\beta}) \cdot P(\bar{\boldsymbol{\gamma}}). \tag{6.6.12}$$

A.3 $P(\boldsymbol{\beta}) \sim N(A_1\bar{\boldsymbol{\beta}}, C)$.

A.4 There is no information about $\bar{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{\gamma}}$; therefore $P(\bar{\boldsymbol{\beta}})$ and $P(\bar{\boldsymbol{\gamma}})$ are independent and

$$P(\bar{\boldsymbol{\beta}}) \propto \text{ constant},$$

$$P(\bar{\boldsymbol{\gamma}}) \propto \text{ constant}.$$

Conditional on $\Omega$ and $C$, repeatedly applying the formulas in Appendix 6, yields (Hsiao et al. 1993)

(1) The posterior distribution of $\bar{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{\gamma}}$ given $\mathbf{y}$ is

$$N\left(\begin{pmatrix}\bar{\beta}^* \\ \bar{\boldsymbol{\gamma}}^*\end{pmatrix}, D_1\right), \tag{6.6.13}$$

where

$$D_1 = \left[\begin{pmatrix}A_1'X' \\ \tilde{W}'\end{pmatrix}(\Omega + XCX')^{-1}(XA_1, \tilde{W})\right]^{-1}, \tag{6.6.14}$$

and

$$\begin{pmatrix}\bar{\boldsymbol{\beta}}^* \\ \bar{\boldsymbol{\gamma}}^*\end{pmatrix} = D_1\begin{bmatrix}A_1'X' \\ \tilde{W}'\end{bmatrix}(\Omega + XCX')^{-1}\mathbf{y} \tag{6.6.15}$$

[15] According to Bayes' theorem, the probability of B given A, written as $P(B \mid A)$, equals $P(B \mid A) = \frac{P(A|B)P(B)}{P(A)}$ which is proportional to $P(A \mid B)P(B)$.

(2) The posterior distribution of $\boldsymbol{\beta}$ given $\bar{\boldsymbol{\beta}}$ and $\mathbf{y}$ is $N(\boldsymbol{\beta}^*, D_2)$, where

$$D_2 = \{X'[\Omega^{-1} - \Omega^{-1}\tilde{W}(\tilde{W}'\Omega^{-1}\tilde{W})^{-1}\tilde{W}'\Omega^{-1}]X + C^{-1}\}^{-1}, \tag{6.6.16}$$

$$\boldsymbol{\beta}^* = D_2\{X'[\Omega^{-1} - \Omega^{-1}\tilde{W}(\tilde{W}'\Omega^{-1}\tilde{W})^{-1}\tilde{W}'\Omega^{-1}]\mathbf{y} + C^{-1}A_1\bar{\boldsymbol{\beta}}\}. \tag{6.6.17}$$

(3) The (unconditional) posterior distribution of $\boldsymbol{\beta}$ is $N(\boldsymbol{\beta}^{**}, D_3)$, where

$$\begin{aligned} D_3 = \{&X'[\Omega^{-1} - \Omega^{-1}\tilde{W}(\tilde{W}'\Omega^{-1}\tilde{W})^{-1}\tilde{W}'\Omega^{-1}]X + C^{-1} \\ &- C^{-1}A_1(A_1'C^{-1}A_1)^{-1}A_1'C^{-1}\}^{-1}, \end{aligned} \tag{6.6.18}$$

$$\begin{aligned} \boldsymbol{\beta}^{**} &= D_3\{X'[\Omega^{-1} - \Omega^{-1}\tilde{W}(\tilde{W}'\Omega^{-1}\tilde{W})^{-1}\tilde{W}'\Omega^{-1}]\mathbf{y}\} \qquad (6.6.19) \\ &= D_2\{X'[\Omega^{-1} - \Omega^{-1}\tilde{W}(\tilde{W}'\Omega^{-1}\tilde{W})^{-1}\tilde{W}'\Omega^{-1}]X\hat{\boldsymbol{\beta}} + C^{-1}A_1\bar{\boldsymbol{\beta}}^*\}, \end{aligned}$$

where $\hat{\boldsymbol{\beta}}$ is the GLS estimate of (6.6.5),

$$\begin{aligned} \hat{\boldsymbol{\beta}} = &\{X'[\Omega^{-1} - \Omega^{-1}\tilde{W}(\tilde{W}'\Omega^{-1}\tilde{W})^{-1}\tilde{W}'\Omega^{-1}]X\}^{-1} \\ &\cdot \{X'[\Omega^{-1} - \Omega^{-1}\tilde{W}(\tilde{W}'\Omega^{-1}\tilde{W})^{-1}\tilde{W}'\Omega^{-1}]\mathbf{y}\}. \end{aligned} \tag{6.6.20}$$

Given a quadratic loss function of the error of the estimation, a Bayes point estimate is the posterior mean. The posterior mean of $\bar{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{\gamma}}$ (6.6.15) is the GLS estimator of the model (6.6.5) after substituting the restriction (6.6.3),

$$\mathbf{y} = XA_1\bar{\boldsymbol{\beta}} + \tilde{W}\bar{\boldsymbol{\gamma}} + \mathbf{v}, \tag{6.6.21}$$

where $\mathbf{v} = X\boldsymbol{\alpha} + \mathbf{u}$. However, the posterior mean of $\boldsymbol{\beta}$ is not the GLS estimator of (6.6.5). It is the weighted average between the GLS estimator of $\boldsymbol{\beta}$ and the overall mean $\bar{\boldsymbol{\beta}}$ (6.6.17) or $\bar{\boldsymbol{\beta}}^*$ (6.6.19), with the weights proportional to the inverse of the precision of respective estimates. The reason is that although both (6.6.2) and (6.6.5) allow the coefficients to be different across cross-sectional units, (6.6.3) has imposed additional prior information that $\boldsymbol{\beta}$ are randomly distributed with mean $A_1\bar{\boldsymbol{\beta}}$. For (6.6.2), the best linear predictor for an individual outcome is to substitute the best linear unbiased estimator of the individual coefficients into the individual equation. For model of (6.6.5) and (6.6.3), because the expected $\boldsymbol{\beta}_i$ is the same across $i$ and the actual difference can be attributed to a chance outcome, additional information about $\boldsymbol{\beta}_i$ may be obtained by examining the behavior of others, hence (6.6.17) or (6.6.19).

In the special case of error components model (6.6.9), $X = I_N \otimes \mathbf{e}_T$. Under the assumption that $\mathbf{w}_{it}$ contains an intercept term (i.e., $\bar{\beta} = 0$) and $u_{it}$ is i.i.d., the Bayes estimator ((6.6.15)) of $\bar{\boldsymbol{\gamma}}$ is simply the GLS estimator of (6.6.21), $\bar{\boldsymbol{\gamma}}^*$. The Bayes estimator of $\alpha_i$ ((6.6.17)) is

$$\alpha_i^{**} = \left(\frac{T\sigma_\alpha^2}{T\sigma_\alpha^2 + \sigma_u^2}\right)\hat{\bar{v}}_i, \tag{6.6.22}$$

where $\hat{\bar{v}}_i = \frac{1}{T}\sum_{\tau=1}^{T}\hat{v}_{it}$ and $\hat{v}_{it} = y_{it} - \bar{\boldsymbol{\gamma}}^{*}\mathbf{w}_{it}$. Substituting $\bar{\boldsymbol{\gamma}}^{*}$, and $\alpha_i^{**}$ for the unknown $\bar{\boldsymbol{\gamma}}$, and $\alpha_i$ in (6.6.9), Wansbeek and Kapteyn (1978) and Taub (1979) show that

$$\hat{y}_{i,T+S} = \bar{\boldsymbol{\gamma}}^{*'}\mathbf{w}_{i,t+s} + \alpha_i^{**} \tag{6.6.23}$$

is the best linear predictor (BLUP) for the $i$th individual $s$ periods ahead.[16]

### 6.6.3 Random or Fixed Differences?

#### *6.6.3.1 An Example of the Contrast between Individual and Pooled Parameter Estimates*

In a classical framework, it makes no sense to predict the independently drawn random variable $\boldsymbol{\beta}_i$ (or $\boldsymbol{\alpha}_i$). However, in panel data, we actually operate with two dimensions – a cross-sectional dimension and a time series dimension. Even though $\boldsymbol{\beta}_i$ is an independently distributed random variable across $i$, once a particular $\boldsymbol{\beta}_i$ is drawn, it stays constant over time. Therefore, it makes sense to predict $\boldsymbol{\beta}_i$. The classical predictor of $\boldsymbol{\beta}_i$ is the GLS estimator of the model (6.6.5). The Bayes predictor (6.6.19) is the weighted average between the GLS estimator of $\boldsymbol{\beta}$ for the model (6.6.5) and the overall mean $A_1\bar{\boldsymbol{\beta}}$ if $\bar{\boldsymbol{\beta}}$ is known or $A_1\bar{\boldsymbol{\beta}}^{*}$ if $\bar{\boldsymbol{\beta}}$ is unknown with the weights proportional to the inverse of the precisions of respective estimates. The Bayes estimator of the individual coefficients, $\boldsymbol{\beta}_i$, "shrinks" the GLS estimator of $\boldsymbol{\beta}_i$ toward the grand mean $\bar{\boldsymbol{\beta}}$ or $\bar{\boldsymbol{\beta}}^{*}$. The reason for doing so stems from de Finneti's (1964) exchangeability assumption. When there are not enough time series observations to allow for precise estimation of individual $\boldsymbol{\beta}_i$ (i.e., $T$ is small), additional information about $\boldsymbol{\beta}_i$ may be obtained by examining the behavior of others because the expected response is assumed the same and the actual differences in response among individuals are the work of a chance mechanism.

Table 6.1 presents the Canadian route specific estimates of the demand for customer-dialed long distance service over 920 miles (long-haul) based on quarterly data from 1980.I to 1989.IV (Hsiao, Appelbe, and Dineen 1993). Some of the point-to-point individual route estimates (unconstrained model) of the price and income coefficients have the wrong signs (Table 6.1, column 2), perhaps because of multicollinearity. However, when one invokes the representative consumer argument by assuming that consumers respond in more or less the same way to price and income changes, thus assuming the coefficients of these variables across routes are considered random draws from a common population with constant mean and variance–covariance matrix, but also allows the route-specific effects to exist by assuming that the coefficients of the intercept and seasonal dummies are fixed and different for different routes, all the

[16] When $u_{it}$ is serially correlated, see Baltagi and Li (1992). For the asymptotic mean square error when the coefficients and error components parameters are estimated, see Baillie and Baltagi (1999).

Table 6.1. *Long-haul regression coefficients*[a]

| | Price coefficient, | |
|---|---|---|
| Route | unconstrained | Mixed coefficients |
| 1 | −0.0712(−0.15) | −0.2875(N/A) |
| 2 | 0.1694(0.44) | −0.0220(N/A) |
| 3 | −1.0142(−5.22) | −0.7743(N/A) |
| 4 | −0.4874(−2.29) | −0.1686(N/A) |
| 5 | −0.3190(−2.71) | −0.2925(N/A) |
| 6 | 0.0365(0.20) | −0.0568(N/A) |
| 7 | −0.3996(−3.92) | −0.3881(N/A) |
| 8 | −0.1033(−0.95) | −0.2504(N/A) |
| 9 | −0.3965(−4.22) | −0.2821(N/A) |
| 10 | −0.6187(−4.82) | −0.5934(N/A) |
| Average | N/A | −0.3116 |
| Route | Income coefficient | |
| 1 | 1.4301(3.07) | 0.4740(N/A) |
| 2 | −0.348(−0.09) | 0.2679(N/A) |
| 3 | 0.3698(1.95) | 0.3394(N/A) |
| 4 | 0.2497(0.70) | 0.3145(N/A) |
| 5 | 0.5556(2.71) | 0.3501(N/A) |
| 6 | 0.1119(0.95) | 0.1344(N/A) |
| 7 | 0.9197(8.10) | 0.5342(N/A) |
| 8 | 0.3886(3.88) | 0.5255(N/A) |
| 9 | 0.6688(6.16) | 0.5648(N/A) |
| 10 | 0.1928(2.39) | 0.2574(N/A) |
| Average | N/A | 0.3762 |

[a] $t$-statistics in parentheses.
*Source:* Hsiao, Appelbe, and Dineen (1993, Table 3).

estimated route specific price and income coefficients have the correct signs (Table 6.1, column 3).

### *6.6.3.2 An Example of Prediction Comparison*

When homogeneity is rejected by the data, whether to treat unobserved heterogeneity as fixed or random has paramount importance in panel data modeling. For instance, in a study of Ontario, Canada regional electricity demand, Hsiao et al. (1989) estimate a model of the form

$$y_{it} = \gamma_i y_{i,t-1} + \boldsymbol{\delta}_i' \mathbf{d}_{it} + \boldsymbol{\beta}_i' \mathbf{x}_{it} + u_{it}, \tag{6.6.24}$$

where $y_{it}$ denotes the logarithm of monthly kilowatt-hour or kilowatt demand for region $i$ at time $t$; $\mathbf{d}_{it}$ denotes 12 monthly dummies; and $\mathbf{x}_{it}$ denotes climatic

Table 6.2. *Root-mean-square prediction error of log kilowatt-hours (one-period-ahead forecast)*

| | Root Mean Square Error | | | |
|---|---|---|---|---|
| Municipality | Region-specific | Pooled | Random coefficients | Mixed |
| Hamilton | 0.0865 | 0.0535 | 0.0825 | 0.0830 |
| Kitchener–Waterloo | 0.0406 | 0.0382 | 0.0409 | 0.0395 |
| London | 0.0466 | 0.0494 | 0.0467 | 0.0464 |
| Ottawa | 0.0697 | 0.0523 | 0.0669 | 0.0680 |
| St. Catharines | 0.0796 | 0.0724 | 0.0680 | 0.0802 |
| Sudbury | 0.0454 | 0.0857 | 0.0454 | 0.0460 |
| Thunder Bay | 0.0468 | 0.0615 | 0.0477 | 0.0473 |
| Toronto | 0.0362 | 0.0497 | 0.0631 | 0.0359 |
| Windsor | 0.0506 | 0.0650 | 0.0501 | 0.0438 |
| Unweighted average | 0.0558 | 0.0586 | 0.0568 | 0.0545 |
| Weighted average[a] | 0.0499 | 0.0525 | 0.0628 | 0.0487 |

[a] The weight is kilowatt-hours of demand in the municipality in June 1985.
*Source:* Hsiao et al. (1989, p. 584).

factor and the logarithm of income, own price, and price of its close substitutes, all measured in real terms. Four different specifications are considered:

1. The coefficients $\boldsymbol{\theta}_i' = (\gamma_i, \boldsymbol{\delta}_i', \boldsymbol{\beta}_i')$ are fixed and different for different region.
2. The coefficients $\boldsymbol{\theta}_i = \boldsymbol{\theta}' = (\gamma, \boldsymbol{\delta}', \boldsymbol{\beta}')$ for all $i$.
3. The coefficient vectors $\boldsymbol{\theta}_i$ are randomly distributed with common mean $\boldsymbol{\theta}$ and covariance matrix $\Delta$.
4. The coefficients $\boldsymbol{\beta}_i$ are randomly distributed with common mean $\bar{\boldsymbol{\beta}}$ and covariance matrix $\Delta_{11}$, and the coefficients $\gamma_i$ and $\boldsymbol{\delta}_i$ are fixed and different for different $i$.

Monthly data for Hamilton, Kitchener-Waterloo, London, Ottawa, St. Catherines, Sudbury, Thunder Bay, Toronto, and Windsor from January 1967 to December 1982 are used to estimate these four different specifications. Comparisons of the one-period ahead root mean square prediction error

$$\sqrt{\sum_{t=T+1}^{T+f} (y_{it} - \hat{y}_{it})^2 / f}$$

from January 1983 to December 1986 are summarized in Tables 6.2 and 6.3. As one can see from these tables, the simple pooling (model 2) and random-coefficients (model 3) formulations on average yield less precise prediction for regional demand. The mixed fixed- and random-coefficients model (model 4) performs the best. It is interesting to note that combining information across

Table 6.3. *Root-mean-square prediction error of log kilowatts (one-period-ahead forecast)*

| | Root Mean Square Error | | | |
|---|---|---|---|---|
| Municipality | Regional specific | Pooled | Random coefficients | Mixed |
| Hamilton | 0.0783 | 0.0474 | 0.0893 | 0.0768 |
| Kitchener–Waterloo | 0.0873 | 0.0440 | 0.0843 | 0.0803 |
| London | 0.0588 | 0.0747 | 0.0639 | 0.0586 |
| Ottawa | 0.0824 | 0.0648 | 0.0846 | 0.0768 |
| St. Catharines | 0.0531 | 0.0547 | 0.0511 | 0.0534 |
| Sudbury | 0.0607 | 0.0943 | 0.0608 | 0.0614 |
| Thunder Bay | 0.0524 | 0.0597 | 0.0521 | 0.0530 |
| Toronto | 0.0429 | 0.0628 | 0.0609 | 0.0421 |
| Windsor | 0.0550 | 0.0868 | 0.0595 | 0.0543 |
| Unweighted average | 0.0634 | 0.0655 | 0.0674 | 0.0619 |
| Weighted average[a] | 0.0558 | 0.0623 | 0.0673 | 0.0540 |

[a] The weight is kilowatt-hours of demand in the municipality in June 1985.
*Source:* Hsiao et al. (1989, p. 584).

regions together with a proper account of regional-specific factors is capable of yielding better predictions for regional demand than the approach of simply using regional-specific data (model 1).

### *6.6.3.3 Model Selection*

The preceding example demonstrates that the way in which individual heterogeneity is taken into account makes a difference in the accuracy of inference. The various estimation methods discussed so far presuppose that we know which coefficients should be treated as fixed (and different) and which coefficients should be treated as random. In practice, we have very little prior information on selecting the appropriate specifications. Various statistical tests have been suggested to select an appropriate formulation (e.g., Breusch and Pagan 1979; Hausman 1978 or Chapter 6, Section 6.2.2.4). However, all these tests essentially exploit the implication of certain formulation in a specific framework. They are indirect in nature. The distribution of a test statistic is derived under a specific null, but the alternative is composite. The rejection of a null does not automatically imply the acceptance of a specific alternative. It would appear more appropriate to treat the fixed coefficients, random coefficients, or various forms of mixed fixed- and random-coefficients models as different models and use model selection criteria to select an appropriate specification (Hsiao and Sun 2000). For instance, well known model selection criterion such as Akaike (1973) information criteria or Schwarz (1978) Bayesian information criteria that selects the model $H_j$ among $j = 1.\ldots, J$ different specifications

if it yields the smallest value of

$$-2\log f(\mathbf{y} \mid H_j) + 2m_j, \quad j = 1\ldots., J, \tag{6.6.25}$$

or

$$-2\log f(\mathbf{y} \mid H_j) + m_j \log NT, \quad j = 1\ldots., J, \tag{6.6.26}$$

can be used, where $\log f(\mathbf{y} \mid H_j)$ and $m_j$ denote the log-likelihood values of $\mathbf{y}$ and the number of unknown parameters of model $H_j$. Alternatively, Hsiao (1995) and Min and Zellner (1993) suggest selecting the model that yields the highest predictive density. In this framework, time series observations are divided into two periods, 1 to $T_1$, denoted by $\mathbf{y}^1$, and $T_1 + 1$ to $T$, denoted by $\mathbf{y}^2$. The first $T_1$ observations are used to obtain the probability distribution of the parameters associated with $H_j$, say $\boldsymbol{\theta}^j$, $P(\boldsymbol{\theta}^j \mid \mathbf{y}^1)$. The predictive density is then evaluated as

$$\int f(\mathbf{y}^2 \mid \boldsymbol{\theta}^j) p(\boldsymbol{\theta}^j \mid \mathbf{y}^1) d\boldsymbol{\theta}^j, \tag{6.6.27}$$

where $f(\mathbf{y}^2 \mid \boldsymbol{\theta}^j)$ is the density of $\mathbf{y}^2$ conditional on $\boldsymbol{\theta}^j$. Given the sensitivity of Bayesian approach to the choice of prior distribution the advantage of using (6.6.27) is that the choice of a model does not have to depend on the prior. One can use the noninformative (or diffuse) prior to derive $P(\boldsymbol{\theta}^j \mid \mathbf{y}^1)$. It is also consistent with the theme that "a severe test for an economic theory, the only test and the ultimate test is its ability to predict" (Klein 1988; p. 21; see also Friedman 1953).

When $\mathbf{y}^2$ contains only a limited number of observations, the choice of model in terms of predictive density may become heavily sample dependent. If too many observations are put in $\mathbf{y}^2$, then a great deal of sample information is not utilized to estimate unknown parameters. One compromise is to modify (6.6.27) by recursively updating the estimates,

$$\begin{aligned} &\int f(\mathbf{y}_T \mid \boldsymbol{\theta}^j, \mathbf{y}^{T-1}) P(\boldsymbol{\theta}^j \mid \mathbf{y}^{T-1}) d\boldsymbol{\theta}^j \\ &\cdot \int f(\mathbf{y}_{T-1} \mid \boldsymbol{\theta}^j, \mathbf{y}^{T-2}) P(\boldsymbol{\theta}^j \mid \mathbf{y}^{T-2}) d\boldsymbol{\theta}^j \\ &\ldots \int f(\mathbf{y}_{T_1+1} \mid \boldsymbol{\theta}^j, \mathbf{y}^1) P(\boldsymbol{\theta}^j \mid \mathbf{y}^1) d\boldsymbol{\theta}^j, \end{aligned} \tag{6.6.28}$$

where $P(\boldsymbol{\theta}^j \mid y^T)$ denotes the posterior distribution of $\boldsymbol{\theta}$ given observations from 1 to $T$. While the formula may look formidable, it turns out that the Bayes updating formula is fairly straightforward to compute. For instance, consider the model (6.6.5). Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \bar{\boldsymbol{\gamma}})$ and $\boldsymbol{\theta}_t$ and $V_t$ denote the posterior mean and variance of $\boldsymbol{\theta}$ based on the first $t$-observations; then

$$\boldsymbol{\theta}_t = V_{t-1}(Q_t' \Omega^{-1} \mathbf{y}_t + V_{t-1}^{-1} \boldsymbol{\theta}_{t-1}), \tag{6.6.29}$$

$$V_t = (Q_t' \Omega^{-1} Q_t + V_{t-1}^{-1})^{-1}, \quad t = T_1 + 1, \ldots, T, \tag{6.6.30}$$

and

$$P(\mathbf{y}_{t+1} \mid \mathbf{y}^t) = \int P(\mathbf{y}_{t+1} \mid \theta, \mathbf{y}^t) P(\boldsymbol{\theta} \mid \mathbf{y}^t) d\boldsymbol{\theta}$$
$$\sim N(Q_{t+1}\boldsymbol{\theta}_t, \Omega + Q_{t+1} V_t Q'_{t+1}), \tag{6.6.31}$$

where $\mathbf{y}'_t = (y_{1t}, y_{2t}, \ldots, y_{Nt})$, $Q_t = (\mathbf{x}'_t, \mathbf{w}'_t)$, $\mathbf{x}_t = (\mathbf{x}_{1t}, \ldots, \mathbf{x}_{Nt})$, $\mathbf{w}_t = (\mathbf{w}_{1t}, \ldots, \mathbf{w}_{Nt})$, $\Omega = E\mathbf{u}_t\mathbf{u}'_t$, and $\mathbf{u}'_t = (u_{1t}, \ldots, u_{Nt})$ (Hsiao et al. 1993).

Hsiao and Sun (2000) have conducted limited Monte Carlo studies to evaluate the performance of these model selection criteria in selecting the random, fixed, and mixed random–fixed coefficients specification. They all appear to have a very high percentage in selecting the correct specification.

## 6.7 DYNAMIC RANDOM-COEFFICIENTS MODELS

For ease of exposition and without loss of the essentials, instead of considering generalizing (6.6.5) into the dynamic model, in this section we consider the generalization of random coefficients model (6.2.1) to the dynamic model of the form[17]

$$y_{it} = \gamma_i y_{i,t-1} + \boldsymbol{\beta}'_i \mathbf{x}_{it} + u_{it}, \quad |\gamma_i| < 1, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{matrix} \tag{6.7.1}$$

where $\mathbf{x}_{it}$ is a $K \times 1$ vector of exogenous variables, and the error term $u_{it}$ is assumed to be independently, identically distributed (i.i.d.) over $t$ with mean 0 and variance $\sigma^2_{u_i}$ and is independent across $i$. The coefficients $\boldsymbol{\theta}_i = (\gamma_i, \boldsymbol{\beta}'_i)'$ are assumed to be independently distributed across $i$ with mean $\bar{\boldsymbol{\theta}} = (\bar{\gamma}, \bar{\boldsymbol{\beta}}')'$ and covariance matrix $\Delta$. Let

$$\boldsymbol{\theta}_i = \bar{\boldsymbol{\theta}} + \boldsymbol{\alpha}_i, \tag{6.7.2}$$

where $\boldsymbol{\alpha}_i = (\alpha_{i1}, \boldsymbol{\alpha}'_{i2})$; we have

$$E\boldsymbol{\alpha}_i = \mathbf{0}, \quad E\boldsymbol{\alpha}_i\boldsymbol{\alpha}'_j = \Delta \text{ if } i = j \text{ and } \mathbf{0} \text{ otherwise}, \tag{6.7.3}$$

[17] We are concerned only with the estimation of the short-run adjustment coefficient $\gamma$. For a discussion of estimating the long-run coefficient, see Pesaran and Smith (1995); Pesaran and Zhao (1999); Pesaran, Shin, and Smith (1999); and Phillips and Moon (1999, 2000).

and[18]

$$E\boldsymbol{\alpha}_i \mathbf{x}'_{jt} = \mathbf{0}. \tag{6.7.4}$$

Stacking the $T$ time series observations of the ith individuals in matrix form yields

$$\underset{T\times 1}{\mathbf{y}_i} = Q_i\boldsymbol{\theta}_i + \mathbf{u}_i, \quad i = 1, \ldots, N. \tag{6.7.5}$$

where $\mathbf{y}_i = (y_{i1}\ldots., y_{iT})'$, $Q_i = (\mathbf{y}_{i,-1}, X_i)$, $\mathbf{y}_{i,-1} = (y_{i0}, \ldots, y_{i,T-1})'$, $X_i = (\mathbf{x}_{i1}\ldots., \mathbf{x}_{iT})'$, $\mathbf{u}_i = (u_{i1}\ldots., u_{iT})'$, and for ease of exposition, we assume that $y_{i0}$ are observable.[19]

We note that because $y_{i,t-1}$ despends on $\gamma_i$, $EQ_i\boldsymbol{\alpha}'_i \neq \mathbf{0}$, that is, the independence between the explanatory variables and $\boldsymbol{\alpha}_i$ (6.2.6) is violated. Substituting $\boldsymbol{\theta}_i = \bar{\boldsymbol{\theta}} + \boldsymbol{\alpha}_i$ into (6.7.5) yields

$$\mathbf{y}_i = Q_i\bar{\boldsymbol{\theta}} + \mathbf{v}_i, \quad i = 1\ldots., N, \tag{6.7.6}$$

where

$$\mathbf{v}_i = Q_i\boldsymbol{\alpha}_i + \mathbf{u}_i. \tag{6.7.7}$$

[18] The strict exogeneity condition (6.7.4) of $\mathbf{x}_{it}$ is crucial in the identification of dynamic random-coefficients model. Chamberlain (1993) has given an example of the lack of identification of $\gamma$ in a model of the form

$$y_{it} = \gamma y_{i,t-1} + \boldsymbol{\beta}_i x_{it} + \alpha_i + u_{it},$$

where $x_{it}$ takes either 0 or 1. Because $E(\alpha_i \mid \mathbf{x}_i, \mathbf{y}_{i,-1})$ is unrestricted, the only moments that are relevant for the identification of $\gamma$ are

$$E(\Delta y_{it} - \gamma\Delta y_{i,t-1} \mid \mathbf{x}_i^{t-1}, \mathbf{y}_i^{t-2}) = E(\beta_i \Delta x_{it} \mid \mathbf{x}_i^{t-1}, \mathbf{y}_i^{t-2}), \quad t = 2, \ldots, T,$$

where $\mathbf{x}_i^t = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{it})$, $\mathbf{y}_i^t = (y_{i0}, \ldots, y_{it})$. Let $\mathbf{w}_i^t = (\mathbf{x}_i^t, \mathbf{y}_i^t)$, the above expression is equivalent to the following two conditions:

$$\begin{aligned} &D(\Delta y_{it} - \gamma\Delta y_{i,t-1} \mid \mathbf{w}_i^{t-2}, \mathbf{x}_{i,t-1} = 0) \\ &\quad = E(\beta_i \mid \mathbf{w}_i^{t-2}, \mathbf{x}_{i,t-1} = 0) P_r(x_{it} = 1 \mid \mathbf{x}_i^{t-2}, x_{i,t-1} = 0), \end{aligned}$$

and

$$\begin{aligned} &E(\Delta y_{it} - \gamma\Delta y_{i,t-1} \mid \mathbf{x}_i^{t-2}, x_{i,t-1} = 1) \\ &\quad = -E(\boldsymbol{\beta}_i \mid \mathbf{w}_i^{t-2}, x_{i,t-1} = 1) P_r(x_{it} = 0 \mid \mathbf{w}_i^{t-2}, x_{i,t-1} = 1) \end{aligned}$$

If $E(\beta_i \mid \mathbf{w}_i^{t-2}, x_{i,t-1} = 0)$ and $E(\beta_i \mid \mathbf{w}_i^{t-2}, x_{i,t-1} = 1)$ are unrestricted and $T$ is fixed, the autoregressive parameter $\gamma$ cannot be identified from the above two equations.

[19] We assume that $T(>3)$ is large enough to identify $\gamma$ and $\beta$. For an example of lack of identification when $T = 3$ and $y_{it}$ is binary, see Chamberlain (1993) or Arellano and Honoré (2001); see also Chapter 7.

Since

$$y_{i,t-1} = \sum_{j=0}^{\infty} (\bar{\gamma} + \alpha_{i1})^j \mathbf{x}'_{i,t-j-1}(\bar{\boldsymbol{\beta}} + \boldsymbol{\alpha}_{i2}) + \sum_{j=0}^{\infty} (\bar{\gamma} + \alpha_{i1})^j u_{i,t-j-1}, \tag{6.7.8}$$

it follows that $E(\mathbf{v}_i \mid Q_i) \neq \mathbf{0}$. Therefore, contrary to the static case, the least-squares estimator of the common mean, $\bar{\boldsymbol{\theta}}$ is inconsistent.

Equations (6.7.7) and (6.7.8) also demonstrate that the covariance matrix of $\mathbf{v}_i$, $V$, is not easily derivable. Thus, the procedure of premultiplying (6.7.6) by $V^{-1/2}$ to transform the model into the one with serially uncorrelated error is not implementable. Neither does the instrumental variable method appear implementable because the instruments that are uncorrelated with $\mathbf{v}_i$ are most likely uncorrelated with $Q_i$ as well.

Pesaran and Smith (1995) have noted that as $T \to \infty$, the least-squares regression of $\mathbf{y}_i$ on $Q_i$ yields a consistent estimator of $\boldsymbol{\theta}_i$, $\hat{\boldsymbol{\theta}}_i$. They suggest a mean group estimator of $\bar{\boldsymbol{\theta}}$ by taking the average of $\hat{\boldsymbol{\theta}}_i$ across $i$,

$$\hat{\bar{\boldsymbol{\theta}}} = \frac{1}{N} \sum_{i=1}^{N} \hat{\boldsymbol{\theta}}_i. \tag{6.7.9}$$

The mean group estimator (6.7.9) is consistent and asymptotically normally distributed so long as $\sqrt{N}/T \to 0$ as both $N$ and $T \to \infty$ (Hsiao, Pesaran, and Tahmiscioglu 1999).

However, panels with large $T$ are typically the exception in economics. Nevertheless, under the assumption that $y_{i0}$ are fixed and known and $\boldsymbol{\alpha}_i$ and $u_{it}$ are independently normally distributed, we can implement the Bayes estimator of $\bar{\boldsymbol{\theta}}$ conditional on $\sigma_i^2$ and $\Delta$ using the formula (6.6.13) just as in the mixed model case discussed in Section 6.6. The Bayes estimator condition on $\Delta$ and $\sigma_i^2$ is equal to

$$\hat{\bar{\boldsymbol{\theta}}}_B = \left\{ \sum_{i=1}^{N} \left[ \sigma_i^2 (Q_i' Q_i)^{-1} + \Delta \right]^{-1} \right\}^{-1} \sum_{i=1}^{N} \left[ \sigma_i^2 (Q_i' Q_i)^{-1} + \Delta \right]^{-1} \hat{\boldsymbol{\theta}}_i, \tag{6.7.10}$$

which is a weighted average of the least-squares estimator of individual units, with the weights being inversely proportional to individual variances. When $T \to \infty$, $N \to \infty$ and $\sqrt{N}/T^{3/2} \to 0$, the Bayes estimator is asymptotically equivalent to the mean group estimator (6.7.9).

In practice, the variance components, $\sigma_i^2$ and $\Delta$, are rarely known, so the Bayes estimator (6.7.10) is rarely feasible. One approach is to substitute the consistently estimated $\sigma_i^2$ and $\Delta$, say (6.2.11) and (6.2.12), into the formula (6.7.10), and treat them as if they were known. For ease of reference, we shall

call (6.7.10) with known $\sigma_i^2$ and $\Delta$ the infeasible Bayes estimator. We shall call the estimator obtained by substituting $\sigma_i^2$ and $\Delta$ in (6.7.10) by their consistent estimates, say (6.2.11) and (6.2.12), the empirical Bayes estimator.

The other approach is to follow Lindley and Smith (1972) by assuming that the prior distributions of $\sigma_i^2$ and $\Delta$ are independent and are distributed as

$$P\left(\Delta^{-1}, \sigma_1^2, \ldots, \sigma_N^2\right) = W(\Delta^{-1} \mid (\rho R)^{-1}, \rho) \prod_{i=1}^{N} \sigma_i^{-1}, \tag{6.7.11}$$

where $W$ represents the Wishart distribution with scale matrix $(\rho R)$ and degrees of freedom $\rho$ (e.g., Anderson 1985). Incorporating this prior into the model (6.7.1)–(6.7.2), we can obtain the marginal posterior densities of the parameters of interest by integrating out $\sigma_i^2$ and $\Delta$ from the joint posterior density. However, the required integrations do not yield closed form solutions. Hsiao, Pesaran, and Tahmiscioglu (1999) have suggested using Gibbs sampler to calculate marginal densities.

The Gibbs sampler is an iterative Markov Chain Monte Carlo method that requires only the knowledge of the full conditional densities of the parameter vector (e.g., Gelfand and Smith 1990). Starting from some arbitrary initial values, say $(\boldsymbol{\theta}_1^{(0)}, \boldsymbol{\theta}_2^{(0)}, \ldots, \boldsymbol{\theta}_k^{(0)})$ for a parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1 \ldots, \boldsymbol{\theta}_k)$, it samples alternatively from the conditional density of each component of the parameter vector conditional on the values of other components sampled in the latest iteration. That is:

(1) Sample $\boldsymbol{\theta}_1^{(j+1)}$ from $P(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2^{(j)}, \boldsymbol{\theta}_3^{(j)}, \ldots, \boldsymbol{\theta}_k^{(j)}, \mathbf{y})$
(2) Sample $\boldsymbol{\theta}_2^{(j+1)}$ from $P(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1^{(j+1)}, \boldsymbol{\theta}_3^{(j)}, \ldots, \boldsymbol{\theta}_k^{(j)}, \mathbf{y})$
$\vdots$
($k$) Sample $\boldsymbol{\theta}_k^{(j+1)}$ from $P(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_1^{(j+1)}, \ldots, \boldsymbol{\theta}_{k-1}^{(j+1)}, \mathbf{y})$

The vectors $\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(k)}$ form a Markov Chain, with transition probability from stage $\boldsymbol{\theta}^{(j)}$ to the next stage $\boldsymbol{\theta}^{(j+1)}$ being

$$K(\boldsymbol{\theta}^{(j)}, \boldsymbol{\theta}^{(j+1)}) = P\left(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2^{(j)}, \ldots, \boldsymbol{\theta}_k^{(j)}, \mathbf{y}\right) P\left(\boldsymbol{\theta}_2 \mid \boldsymbol{\theta}_1^{(j+1)}, \boldsymbol{\theta}_3^{(j)}, \ldots, \boldsymbol{\theta}_k^{(j)}, y\right)$$
$$\ldots P\left(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}_1^{(j+1)}, \ldots, \boldsymbol{\theta}_{k-1}^{(j+1)}, \mathbf{y}\right).$$

As the number of iterations $j$ approaches infinity, the sampled values in effect can be regarded as drawing from true joint and marginal posterior densities. Moreover, the ergodic averages of functions of the sample values will be consistent estimations of their expected values.

Under the assumption that the prior of $\bar{\theta}$ is $N(\bar{\boldsymbol{\theta}}^*, \Psi)$, the relevant conditional distributions that are needed to implement the Gibbs sampler for (6.7.1)–(6.7.2)

are easily obtained from

$$P\left(\boldsymbol{\theta}_i \mid \mathbf{y}, \bar{\boldsymbol{\theta}}, \Delta^{-1}, \sigma_1^2, \ldots, \sigma_N^2\right)$$
$$\sim N\left\{A_i(\sigma_i^{-2} Q_i' \mathbf{y}_i + \Delta^{-1}\bar{\theta}), A_i\right\}, \quad i = 1, \ldots, N,$$
$$P(\bar{\boldsymbol{\theta}} \mid \mathbf{y}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N, \Delta^{-1}, \sigma_1^2, \ldots, \sigma_N^2) \sim N\left\{D(N\Delta^{-1}\hat{\bar{\boldsymbol{\theta}}} + \Psi^{-1}\boldsymbol{\theta}^*), B\right\}$$
$$P\left(\Delta^{-1} \mid \mathbf{y}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N, \bar{\boldsymbol{\theta}}, \sigma_1^2, \ldots, \sigma_N^2\right)$$
$$\sim W\left[\left(\sum_{i=1}^{N}(\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta}_i - \bar{\boldsymbol{\theta}})' + \rho R\right)^{-1}, \rho + N\right],$$
$$P\left(\sigma_i^2 \mid \mathbf{y}_i, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N, \bar{\boldsymbol{\theta}}, \Delta^{-1}\right)$$
$$\sim IG[T/2, (\mathbf{y}_i - Q_i\boldsymbol{\theta}_i)'(\mathbf{y}_i - Q_i\boldsymbol{\theta}_i)/2], i = 1\ldots., N,$$

where $A_i = (\sigma_i^{-2} Q_i' Q_i + \Delta^{-1})^{-1}$, $D = (N\Delta^{-1} + \Psi^{-1})^{-1}$, $\hat{\bar{\boldsymbol{\theta}}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\theta}_i$, and $IG$ denotes the inverse gamma distribution.

Hsiao, Pesaran, and Tahmiscioglu (1999) have conducted Monte Carlo experiments to study the finite sample properties of (6.7.10), referred as infeasible Bayes estimator; the Bayes estimator obtained through the Gibbs sampler, referred as hierarchical Bayes estimator; the empirical Bayes estimator; the group mean estimator (6.7.8); the bias corrected group mean estimator obtained by directly correcting the finite $T$ bias of the least squares estimator, $\hat{\boldsymbol{\theta}}_i$, using the formula of Kiviet (1995); Kiviet and Phillips (1993); and then taking the average; and the pooled least-squares estimator. Table 6.4 presents the bias of the different estimators of $\bar{\gamma}$ for $N = 50$ and $T = 5$ or 20. The infeasible Bayes estimator performs very well. It has small bias even for $T = 5$. For $T = 5$, its bias falls within the range of 3 to 17 percent. For $T = 20$, the bias is at most about 2 percent. The hierarchical Bayes estimator also performs well,[20] followed by the empirical Bayes estimator when $T$ is small but improves quickly as $T$ increases. The empirical Bayes estimator gives very good results even for $T = 5$ in some cases but the bias also appears to be quite high in certain other cases. As $T$ gets larger its bias decreases considerably. The mean group and the bias corrected mean group estimator both have large bias when $T$ is small, with the bias-corrected mean group estimator performing slightly better. However, the performance of both improve as $T$ increases, and both are still much better than the least-squares estimator. The least-squares estimator yields significant bias and its bias persists as $T$ increases.

The Bayes estimator is derived under the assumption that the initial observations, $y_{i0}$, are fixed constants. As discussed in Chapter 4 or Anderson and Hsiao (1981, 1982), this assumption is clearly unjustifiable for a panel with

[20] The $\Psi^{-1} = 0$, $\rho = 2$ and $R$ equal to the Swamy estimate of $\Delta$ are used to implement the hierarchical Bayes estimator.

Table 6.4. *Bias of the short-run coefficient* $\bar{\gamma}$

| | | | Bias | | | | | |
|---|---|---|---|---|---|---|---|---|
| $T$ | $\bar{\gamma}$ | | Pooled OLS | Mean group | Bias-corrected mean group | Infeasible Bayes | Empirical Bayes | Hierarchical Bayes |
| 5 | 1 | 0.3 | 0.36859 | −0.23613 | −0.14068 | 0.05120 | −0.12054 | −0.02500 |
| | 2 | 0.3 | 0.41116 | −0.23564 | −0.14007 | 0.04740 | −0.11151 | −0.01500 |
| | 3 | 0.6 | 1.28029 | −0.17924 | −0.10969 | 0.05751 | −0.02874 | 0.02884 |
| | 4 | 0.6 | 1.29490 | −0.18339 | −0.10830 | 0.06879 | −0.00704 | 0.06465 |
| | 5 | 0.3 | 0.06347 | −0.26087 | −0.15550 | 0.01016 | −0.18724 | −0.10068 |
| | 6 | 0.3 | 0.08352 | −0.26039 | −0.15486 | 0.01141 | −0.18073 | −0.09544 |
| | 7 | 0.6 | 0.54756 | −0.28781 | −0.17283 | 0.05441 | −0.12731 | −0.02997 |
| | 8 | 0.6 | 0.57606 | −0.28198 | −0.16935 | 0.06258 | −0.10366 | −0.01012 |
| 20 | 9 | 0.3 | 0.44268 | −0.07174 | −0.01365 | 0.00340 | −0.00238 | 0.00621 |
| | 10 | 0.3 | 0.49006 | −0.06910 | −0.01230 | 0.00498 | −0.00106 | 0.00694 |
| | 11 | 0.35 | 0.25755 | −0.06847 | −0.01209 | −0.00172 | −0.01004 | −0.00011 |
| | 12 | 0.35 | 0.25869 | −0.06644 | −0.01189 | −0.00229 | −0.00842 | 0.00116 |
| | 13 | 0.3 | 0.07199 | −0.07966 | −0.01508 | −0.00054 | −0.01637 | −0.00494 |
| | 14 | 0.3 | 0.09342 | −0.07659 | −0.01282 | 0.00244 | −0.01262 | −0.00107 |
| | 15 | 0.55 | 0.26997 | −0.09700 | −0.02224 | −0.00062 | −0.01630 | 0.00011 |
| | 16 | 0.55 | 0.29863 | −0.09448 | −0.02174 | −0.00053 | −0.01352 | 0.00198 |

*Source:* Hsiao, Pesaran, and Tahmiscioglu (1999).

finite $T$. However, contrary to the sampling approach where the correct modeling of initial observations is quite important, the Bayesian approach appears to perform fairly well in the estimation of the mean coefficients for dynamic random-coefficients models even when the initial observations are treated as fixed constants. The Monte Carlo study also cautions against the practice of justifying the use of certain estimators based on their asymptotic properties. Both the mean group and the corrected mean group estimators perform poorly in panels with very small $T$. The hierarchical Bayes estimator appears preferable to the other consistent estimators unless the time dimension of the panel is sufficiently large.

## 6.8 TWO EXAMPLES

### 6.8.1 Liquidity Constraints and Firm Investment Expenditure

The effects of financial constraints on company investment have been subject to intensive debate by economists. At one extreme, Jorgenson (1971) claims that "the evidence clearly favors the Modigliani-Miller theory (1958, 61). Internal liquidity is not an important determinant of the investment, given the level of output and external funds." At the other extreme, Stiglitz and Weiss (1981) argue that because of imperfections in the capital markets, costs of internal and external funds generally will diverge, and internal and external funds generally will not be perfect substitutes for each other. Fazzari, Hubbard, and Petersen (1988), Bond and Meghir (1994), etc. tested for the importance of internal finance by studying the effects of cash flow across different groups of companies like identifying groups of firms according to company retention practices. If the null hypothesis of perfect capital market is correct, then no difference should be found in the coefficient of cash flow variable across groups. However, these authors find that cash flow coefficient is large for companies with low dividend payout rates.

However, there is no sound theoretical basis for assuming that only low dividend payout companies are subject to financial constraints. The finding that larger companies have larger cash flow coefficients is inconsistent with both the transaction costs and asymmetric information explanations of liquidity constraints. Whether firm heterogeneity can be captured by grouping firms according to some indicators remains open to question.

Hsiao and Tahmiscioglu (1997) use COMPUSTAT annual industrial files of 561 firms in manufacturing sector for the period 1971–1992 to estimate the following five different investment expenditure models with and without using liquidity models:

$$\left(\frac{I}{K}\right)_{it} = \alpha_i^* + \gamma_i \left(\frac{I}{K}\right)_{i,t-1} + \beta_{i1}\left(\frac{LIQ}{K}\right)_{i,t-1} + \epsilon_{it}, \tag{6.8.1}$$

$$\left(\frac{I}{K}\right)_{it} = \alpha_i^* + \gamma_i \left(\frac{I}{K}\right)_{i,t-1} + \beta_{i1}\left(\frac{LIQ}{K}\right)_{i,t-1} + \beta_{i2} q_{it} + \epsilon_{it}, \tag{6.8.2}$$

$$\left(\frac{I}{K}\right)_{it} = \alpha_i^* + \gamma_i \left(\frac{I}{K}\right)_{i,t-1} + \beta_{i1} \left(\frac{LIQ}{K}\right)_{i,t-1} + \beta_{i2} \left(\frac{S}{K}\right)_{i,t-1} + \epsilon_{it}, \tag{6.8.3}$$

$$\left(\frac{I}{K}\right)_{it} = \alpha_i^* + \gamma_i \left(\frac{I}{K}\right)_{i,t-1} + \beta_{i2} q_{it} + \epsilon_{it}, \tag{6.8.4}$$

and

$$\left(\frac{I}{K}\right)_{it} = \alpha_i^* + \gamma_i \left(\frac{I}{K}\right)_{i,t-1} + \beta_{i2} \left(\frac{S}{K}\right)_{i,t-1} + \epsilon_{it}. \tag{6.8.5}$$

where $I_{it}$ is firm $i$'s capital investment at time $t$, $LIQ_{it}$ is a liquidity variable (defined as cash flow minus dividends); $S_{it}$ is sales, $q_{it}$ is Tobin's $q$ (Brainard and Tobin 1968; Tobin 1969), defined as the ratio of the market value of the firm to the replacement value of capital; and $K_{it}$ is the beginning-of-period capital stock. The coefficient $\beta_{i1}$ measures the short-run impact of liquidity variable on firm $i$'s investment in each of these three specifications. Models 4 and 5 ((6.8.4) and (6.8.5)) are two popular variants of investment equations that do not use the liquidity variable as an explanatory variable – the Tobin $q$ model (e.g., Hayashi 1982; Summers 1981) and the sales capacity model (e.g., Kuh 1963). The sale variable can be regarded as a proxy for future demand for the firm's output. The $q$ theory relates investment to marginal $q$, which is defined as the ratio of the market value of new investment goods to their replacement cost. If a firm has unexploited profit opportunities, then an increase of its capital stock of \$1 will increase its market value by more than \$1. Therefore, firm managers can be expected to increase investment until marginal $q$ equals 1. Thus, investment will be an increasing function of marginal $q$. Because marginal $q$ is unobservable, it is common in empirical work to substitute it with average or Tobin's $q$.

Tables 6.5 and 6.6 present some summary information from the firm by firm regressions of these five models. Table 6.5 shows the percentage of significant coefficients at the 5 percent significance level for a one-tailed test. Table 6.6 shows the first and third quartiles of the estimated coefficients. The estimated coefficients vary widely from firm to firm. The $F$-test of slope homogeneity across firms while allowing for firm-specific intercepts is also rejected (see Table 6.5).

The approach of relating the variation of $\beta_{i1}$ to firm characteristics such as dividend payout rate, company size, sales growth, capital intensity, standard deviation of retained earnings, debt-to-equity ratio, measures of liquidity stocks from the balance sheet, number of shareholders, and industry dummies is unsuccessful. These variables as a whole do not explain the variation of estimated $\beta_{i1}$ well. The maximum $\bar{R}^2$ is only 0.113. Many of the estimated coefficients are not significant under various specifications. Neither can one substitute functions of the form (6.5.2) into (6.8.1)–(6.8.5) and estimate the coefficients directly because of perfect multicollinearity. So Hsiao and Tahmiscioglu (1997) classify firms into reasonably homogeneous groups using the

Table 6.5. *Individual firm regressions (percentage of firms with significant coefficients)*

| | Percentage of firms | | | | |
|---|---|---|---|---|---|
| | Model 1 | 2 | 3 | 4 | 5 |
| Coefficient for: | | | | | |
| $(\text{LIQ}/K)_{t-1}$ | 46 | 36 | 31 | | |
| $q$ | | 31 | | 38 | |
| $(S/K)_{t-1}$ | | | 27 | | 44 |
| Percentage of firms with significant autocorrelation | 14 | 12 | 13 | 20 | 15 |
| Actual $F$ | 2.47 | 2.98 | 2.01 | 2.66 | 2.11 |
| Critical $F$ | 1.08 | 1.08 | 1.08 | 1.06 | 1.06 |

*Note:* The number of firms is 561. The significance level is 5 percent for a one-tailed test. Actual $F$ is the $F$ statistic for testing the equality of slope coefficients across firms. For the $F$ test, the 5 percent significance level is chosen. To detect serial correlation, Durbin's $t$-test at the 5 percent significance level is used.

*Source:* Hsiao and Tahmiscioglu (1997, Table 1).

capital intensity ratio of 0.55 as a cut-off point. Capital intensity is defined as the minimum value of the ratio of capital stock to sales over the sample period. It is the most statistically significant and most stable variable under different specifications.

Table 6.7 presents the variable intercept estimates for the groups of less and more capital intensive firms. The liquidity variable is highly significant in all three variants of the liquidity model. There are also significant differences in the coefficients of the liquidity variable across the two groups. However, Table 6.7 also shows that the null hypothesis of the equality of slope coefficients conditioning on the firm-specific effects is strongly rejected for all specifications

Table 6.6. *Coefficient heterogeneity: slope estimates at first and third quartiles across a sample of 561 firms*

| | Slope estimates | | | |
|---|---|---|---|---|
| Model | $(I/K)_{i,t-1}$ | $(\text{LIQ}/K)_{i,t-1}$ | $q_{it}$ | $(S/K)_{i,t-1}$ |
| 1 | .026, .405 | .127, .529 | | |
| 2 | −.028, .359 | .062, .464 | 0, .039 | |
| 3 | .100, .295 | .020, .488 | | −.005, .057 |
| 4 | .110, .459 | | .007, .048 | |
| 5 | −.935, .367 | | | .012, .077 |

*Source:* Hsiao and Tahmiscioglu (1997, Table 2).

Table 6.7. *Variable intercept estimation of models for less- and more-capital-intensive firms*

| Variable | Variable intercept estimate | | | | | |
|---|---|---|---|---|---|---|
| | Less-capital-intensive firms | | | More-capital-intensive firms | | |
| $(I/K)_{i,t-1}$ | .265 | .198 | .248 | .392 | .363 | .364 |
| | (.011) | (.012) | (.011) | (.022) | (.023) | (.022) |
| $(\text{LIQ}/K)_{i,t-1}$ | .161 | .110 | .119 | .308 | .253 | .278 |
| | (.007) | (.007) | (.007) | (.024) | (.027) | (.025) |
| $(S/K)_{i,t-1}$ | | .023 | | | .025 | |
| | | (.001) | | | (.006) | |
| $q_{it}$ | | | .011 | | | .009 |
| | | | (.0006) | | | (.002) |
| Actual $F$ | 2.04 | 1.84 | 2.22 | 2.50 | 2.19 | 2.10 |
| Critical $F$ | 1.09 | 1.07 | 1.07 | 1.20 | 1.17 | 1.17 |
| Numerator d.f. | 834 | 1,251 | 1,251 | 170 | 255 | 255 |
| Denominator d.f. | 6,592 | 6,174 | 6,174 | 1,368 | 1,282 | 1,282 |
| Number of firms | 418 | 418 | 418 | 86 | 86 | 86 |

*Note:* The dependent variable is $(I/K)_{it}$. Less-capital-intensive firms are those with minimum $(K/S)$ between 0.15 and 0.55 over the sample period. For more-capital-intensive firms, the minimum $(K/S)$ is greater than 0.55. The regressions include company-specific intercepts. Actual $F$ is the $F$ statistic for testing the homogeneity of slope coefficients. For the $F$ test, a 5 percent significance level is chosen. The estimation period is 1974–1992. Standard errors are in parentheses.
*Source:* Hsiao and Tahmiscioglu (1997, Table 5).

for both groups. In other words, using the capital intensity ratio of 0.55 as a cut-off point, there is still substantial heterogeneity within the groups.

As neither there appears to have a set of explanatory variables that adequately explains the variation of $\beta_{i1}$, nor can homogeneity be achieved by classifying firms into groups, one is left with either treating $\boldsymbol{\beta}_i$ as fixed and different or treating $\boldsymbol{\beta}_i$ as random draws from a common distribution. Within the random-effects framework, individual differences are viewed as random draws from a population with constant mean and variance. Therefore, it is appropriate to pool the data and try to draw some generalization about the population. On the other hand, if individual differences reflect fundamental heterogeneity or if individual response coefficients depend on the values of the included explanatory variables, estimation of the model parameters based on the conventional random effects formulation can be misleading. To avoid this bias, heterogeneity among individuals must be treated as fixed. In other words, one must investigate investment behavior firm by firm, and there is no advantage of pooling. Without pooling, the shortage of degrees of freedom and multicollinearity can render the resulting estimates meaningless and make drawing general conclusions difficult.

Table 6.8 presents the estimates of the mixed fixed- and random-coefficients model of the form (6.6.24) by assuming that conditional on company-specific effects, the remaining slope coefficients are randomly distributed around a

Table 6.8. *Estimation of mixed fixed- and random-coefficient models for less- and more-capital-intensive firms*

| | Estimate | | | | | |
|---|---|---|---|---|---|---|
| Variable | Less-capital-intensive firms | | | More-capital-intensive firms | | |
| $(I/K)_{i,t-1}$ | .230 | .183 | .121 | .321 | .302 | .236 |
| | (.018) | (.017) | (.019) | (.036) | (.037) | (.041) |
| $(\text{LIQ}/K)_{i,t-1}$ | .306 | .252 | .239 | .488 | .449 | .416 |
| | (.021) | (.023) | (.027) | (.065) | (.067) | (.079) |
| $(S/K)_{i,t-1}$ | | .024 | | | .038 | |
| | | (.003) | | | (.015) | |
| $q_{it}$ | | .019 | | | .022 | |
| | | (.003) | | | (.008) | |
| Number of firms | 418 | 418 | 418 | 86 | 86 | 86 |

*Note:* The dependent variable is $(I/K)_{it}$. The regressions include fixed firm-specific effects. The estimation period is 1974–1992. Standard errors are in parentheses.
*Source:* Hsiao and Tahmiscioglu (1997, Table 7).

certain mean within each of less and more capital-intensive groups. To evaluate the appropriateness of these specifications, Table 6.9 presents the comparison of the recursive predictive density of the mixed fixed- and random-coefficients models and the fixed-coefficients model assuming that each company has different coefficients for the three variants of the liquidity model by dividing

Table 6.9. *Least-squares estimation of aggreate money demand function*

| Dependent variable | Sample period | Variable | Parameter estimate | Standard error |
|---|---|---|---|---|
| M2 | 1980.IV–2000.IV | Intercept | 1.30462 | 0.28975 |
| | | Real GDP | −0.15425 | 0.04538 |
| | | RM2(-1) | 1.07022 | 0.02790 |
| | | Bond rate | −0.00186 | 0.00069 |
| | 1992.I–2000.IV | Intercept | −0.16272 | 0.85081 |
| | | Real GDP | 0.00847 | 0.06772 |
| | | RM2(-1) | 1.00295 | 0.02248 |
| | | Bond rate | −0.00250 | 0.00140 |
| M1 | 1980.IV–2000.IV | Intercept | 0.46907 | 0.21852 |
| | | Real GDP | −0.01857 | 0.01700 |
| | | RM(-1) | 0.98964 | 0.01249 |
| | | Bond rate | −0.00566 | 0.00135 |
| | 1992.I–2000.IV | Intercept | −0.68783 | 2.10228 |
| | | Real GDP | 0.08414 | 0.14898 |
| | | RM1(-1) | 0.96038 | 0.019999 |
| | | Bond rate | −0.01005 | 0.00283 |

*Source:* Hsiao, Shen and Fujiki (2005, Table 5).

the sample into pre- and post-1989 periods. The numbers reported in Table 6.8 are the logarithms of (6.6.28). The results indicate that the mixed fixed- and random-coefficients model is favored over the fixed-coefficients model for both groups. Similar comparisons between the liquidity model, Tobin's $q$, and sales accelerator models also favor liquidity as an important explanatory variable.

Table 6.8 shows that the estimated liquidity coefficients are highly significant and there are significant differences between different classes of companies. The mean coefficient of the liquidity variable turns out to be 60 to 80 percent larger for the more capital-intensive group than for the less capital-intensive group. The implied long-run relationships between the liquidity variable and the fixed investment variable are also statistically significant. For instance, for model (6.8.1), a 10% increase in liquidity capital ratio leads to a 4% increase in fixed investment capital ratio in the long run for the less capital-intensive group compared to a 7% increase in the ratio for the more capital-intensive group. The mixed model also yields substantially larger coefficient estimates of the liquidity variable than those obtained from the variable intercept model. If the coefficients are indeed randomly distributed and the explanatory variables are positively autocorrelated, then this is precisely what one would expect from the within-estimates (Pesaran and Smith 1995).

In short, there are substantial differences across firms in their investment behavior. When these differences are ignored by constraining the parameters to be identical across firms, the impact of liquidity variable on firm investment is seriously underestimated. The mixed fixed- and random-coefficients model appears to fit the data well. The mixed model allows pooling and allows some general conclusions to be drawn about a group of firms. The estimation results and prediction tests appear to show that financial constraints are the most important factor affecting actual investment expenditure, at least for a subset U.S. manufacturing companies.

### 6.8.2 Aggregate versus Disaggregate Analysis

A model is a simplification of the real world. The purpose is to capture the essential factors that affect the outcomes. One of the tools for reducing the real-world detail is through "suitable" aggregation. However, for aggregation not to distort the fundamental behavioral relations among economic agents, certain "homogeneity" conditions must hold between the micro-units. Many economists have shown that if micro-units are heterogeneous, aggregation can lead to very different relations among macro-variables from those of the micro-relations (e.g., Lewbel 1992, 1994; Pesaran 2003; Stoker 1993; Theil 1954; Trivedi 1985).

For instance, consider the simple dynamic equation,

$$y_{it} = \gamma_i y_{i,t-1} + \mathbf{x}'_{it}\boldsymbol{\beta}_i + \alpha_i + u_{it}, \ | \gamma_i | < 1, \quad i = 1, \ldots, N, \quad (6.8.6)$$

where the error $u_{it}$ is covariance stationary. Equation (6.8.6) implies a long-run relation between $y_{it}$ and $\mathbf{x}_{it}$,

$$y_{it} - \mathbf{x}_{it}'\mathbf{b}_i - \eta_i = v_{it} \tag{6.8.7}$$

where $\mathbf{b}_i = (1-\gamma_i)^{-1}\boldsymbol{\beta}_i$, $\eta_i = (1-\gamma_i)^{-1}\alpha_i$, $v_{it} = (1-\gamma_i)^{-1}u_{it}$.

Let $y_t = \sum_{i=1}^N y_{it}$ and $\mathbf{x}_t = \sum_{i=1}^N \mathbf{x}_{it}$; then a similar long-run relation between $\mathbf{y}_t$ and $\mathbf{x}_t$,

$$y_t - \mathbf{x}_t'\mathbf{b} - c = v_t, \tag{6.8.8}$$

holds for a stationary $v_t$ if and only if either of the following conditions hold (Hsiao, Shen, and Fujiki 2005):

(1) $\frac{1}{1-\gamma_i}\boldsymbol{\beta}_i = \frac{1}{1-\gamma_j}\boldsymbol{\beta}_j$ for all $i$ and $j$; or

(2) if $\frac{1}{1-\gamma_i}\boldsymbol{\beta}_i \neq \frac{1}{1-\gamma_j}\boldsymbol{\beta}_j$, then $\mathbf{x}_t' = (\mathbf{x}_{1t}', \ldots, \mathbf{x}_{Nt}')$ must lie on the null space of $D$ for all $t$, where $D' = \left(\frac{1}{1-\gamma_1}\boldsymbol{\beta}_1' - \mathbf{b}', \ldots, \frac{1}{1-\gamma_N}\boldsymbol{\beta}_N' - \mathbf{b}'\right)$.

Panel data provide information on micro-units. They can be used to check if either of these two suitable aggregation conditions hold. For instance, Hsiao, Shen, and Fujiki (2005) find that the estimated aggregate relations between (real) money demand, (real) GDP, and (5-year) bond rate are unstable and sensitive to the time period dovered (see Table 6.9). Depending on the sample period covered, the estimated relations are either of wrong sign or statistically insignificant. They find that the estimated long-run income elasticities are 75.23 for M1 and 11.04 for M2, respectively, an incredible magnitude.

Hsiao, Shen, and Fujiki (2005) attribute the "incredible" results from aggregate data analysis to the "heterogeneity" among the 47 prefectures of Japan. When micro-relations are "heterogeneous," one way is to estimate each micro-relation separately. However, there may not have enough time series observations to obtain reliable micro-relations. Moreover, policymakers are interested in average relations, not individual relations. A random coefficienet framework is a convenient formulation that take account indivdiual heterogeneity while still allowing the estimation of average relation. Table 6.10 provides a random coefficient model estimates of the mean relation between (real) money demand and (real) GDP and (5-year) bond rate for teh 40 Japanese perfectures. The estimated short-run income elasticity for M1 and M2 is 0.88 and 0.47, respectively. The long-run income elasticity is 2.56 for M1 and 1.01 for M2. These results appear to be consistent with economic theory and the broadly observed facts about Japan. The average growth rate for M2 in the 1980s is about 9.34 percent. The inflation rate is 1.98 percent. The real M2 growth rate is 7.36 percent. The real growth rate of GDP during this period is 4.13 percent. Taking account the impact of 5-year bond rate fell from 9.332 percent at 1980.I to 5.767 at 1989.IV, the results are indeed very close to the estimated long-run income elasticites based on disaggregate data analysis.

If "heterogeneity" is indeed present in micro-units, then shall we predict the aggregate outcome based on the summation of estimated micro-relations

Table 6.10. *Random-coefficient estimates of Japan Prefecture money demand equation*

| | M1 | | M2 | |
|---|---|---|---|---|
| | Coefficient | Standard error | Coefficient | Standard error |
| Lagged money | 0.656 | 0.034 | 0.533 | 0.069 |
| Income | 0.881 | 0.114 | 0.473 | 0.064 |
| Bond Rate | −0.0476 | 0.006 | −0.009 | 0.003 |
| Constant | −2.125 | 0.038 | 0.043 | 0.239 |
| | Variance–covariance matrix of $M1(\gamma_i, \boldsymbol{\beta}_i')$ | | | |
| | 0.015 | | | |
| | −0.001 | 0.177 | | |
| | 0.001 | −0.059 | 0.0005 | |
| | −0.024 | −0.588 | −0.023 | 2.017 |
| | Variance–covariance matrix of $M2$ equation $(\gamma_i, \boldsymbol{\beta}_i')$. | | | |
| | 0.068 | | | |
| | −0.031 | 0.062 | | |
| | 0.002 | 0.0003 | 0.0014 | |
| | −0.13 | −0.107 | −0.009 | 0.8385 |

*Source:* Hsiao, Shen and Fujiki (2005, Table 1).

or shall we predict the aggregate outcomes based on the estimated aggregate relations? Unfortunately, there is not much work on this specific issue. In choosing between whether to predict aggregate variables using aggregate ($H_a$) or disaggregate ($H_d$) equations, Griliches and Grunfeld (1960) suggest using the criterion of:

$$\text{Choose } H_d \text{ if } \mathbf{e}_d'\mathbf{e}_d < \mathbf{e}_a'\mathbf{e}_a; \quad \text{otherwise choose } H_a$$

where $\mathbf{e}_d$ and $\mathbf{e}_a$ are the estimates of the errors in predicting aggregate outcomes under $H_d$ and $H_a$, respectively. Hsiao, Shen, and Fujiki (2005) provide a simulation comparison on artificially generated time series data for each prefecture based on the observed stylized facts. Table 6.11 presents the

Table 6.11. *Error sum of squares (ESS) and predicted error sum of squares (PES) for disaggregate and aggregate data*

| | M1 | | M2 | |
|---|---|---|---|---|
| | Aggregate data | Disaggregate data | Aggregate data | Disaggregate data |
| EES | $3.78 \times 10^9$ | $1.35 \times 10^6$ | $3.59 \times 10^{43}$ | $7.45 \times 10^{42}$ |
| PES | $2.51 \times 10^{10}$ | $5.75 \times 10^7$ | $9.55 \times 10^{45}$ | $2.04 \times 10^{43}$ |

*Source:* Hsiao, Shen and Fujiki (2005, Table VIII).

within-sample fit comparisons in the first row and the post-sample prediction comparison in the second row. Both criteria unambiguously favour predicting aggregate outcomes by summing the outcomes from the disaggregate equations.

## 6.9 CORRELATED RANDOM-COEFFICIENTS MODELS

### 6.9.1 Introduction

Standard random-coefficients models assume the variation of coefficients are independent of the variation of regressors (e.g., Chapter 6, Section 6.2.1; Hsiao and Pesaran 2009). In recent years, a great deal of attention has been devoted to the correlated random coefficients model (e.g., Card 1995; Heckman and Vytlacil 1998; Heckman, Urzua, and Vytlacil, 2006; Heckman, Schmierer, and Urzua 2010). This type of model is motivated by the measurement of treatment effect of a policy. For instance, in the study of return to schooling, it is plausible that there are unmeasured ability or motivation factors that affect the return to schooling and are also correlated with the level of schooling (e.g., Card 1995; Heckman and Vytlacil 1998). As a matter of fact, Li and Tobias (2011) find strong evidence that the amount of schooling attained is determined, in part, by the individual's own return to education. Specifically a one percentage increase in the return to schooling is associated with roughly 0.2 more years of education.

A common formulation for a correlated random-coefficients model is to let

$$\boldsymbol{\beta}_i = \bar{\boldsymbol{\beta}} + \boldsymbol{\alpha}_i. \tag{6.9.1}$$

Substituting (6.9.1) into the regression model (6.1.2) yields

$$y_{it} = \mathbf{x}_{it}'\bar{\boldsymbol{\beta}} + \mathbf{x}_{it}'\boldsymbol{\alpha}_i + u_{it}, \tag{6.9.2}$$

where

$$E\boldsymbol{\alpha}_i = \mathbf{0}, \tag{6.9.3}$$

$$E\boldsymbol{\alpha}_i\boldsymbol{\alpha}_j' = \begin{cases} \Delta, & \text{if } i = j. \\ \mathbf{0}, & \text{if } i \neq j. \end{cases} \tag{6.9.4}$$

and

$$E(u_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}_i) = 0, \tag{6.9.5}$$

However, we now assume

$$E\mathbf{x}_{it}\boldsymbol{\alpha}_i' \neq \mathbf{0}. \tag{6.9.6}$$

Let $v_{it} = \mathbf{x}_{it}'\alpha_i + u_{it}$; then

$$E(v_{it} \mid \mathbf{x}_{it}) \neq 0. \tag{6.9.7}$$

### 6.9.2 Identification with Cross-Sectional Data

If only cross-sectional observations of $(y, \mathbf{x})$ are available, it is not possible to identify $\bar{\boldsymbol{\beta}}$. Nor does the existence of instruments $\mathbf{z}_1$ such that

$$\operatorname{cov}(\mathbf{z}_1, \mathbf{x}) \neq \mathbf{0}, \tag{6.9.8}$$

$$\operatorname{cov}(\mathbf{z}_1, u) = \mathbf{0} \tag{6.9.9}$$

alone is sufficient to identify $\bar{\boldsymbol{\beta}}$ because the instrumental variable estimator

$$\begin{aligned}
\hat{\bar{\boldsymbol{\beta}}}_{iv} &= \left[\left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{z}_{1i}'\right)\left(\sum_{i=1}^{N} \mathbf{z}_{1i}\mathbf{z}_{1i}'\right)^{-1}\left(\sum_{i=1}^{N} \mathbf{z}_{1i}\mathbf{x}_i\right)\right]^{-1} \\
&\quad \cdot \left[\left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{z}_{1i}'\right)\left(\sum_{i=1}^{N} \mathbf{z}_{1i}\mathbf{z}_{1i}'\right)^{-1}\left(\sum_{i=1}^{N} \mathbf{z}_{1i} y_i\right)\right] \\
&= \bar{\boldsymbol{\beta}} + \left[\left(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{z}_{1i}'\right)\left(\sum_{i=1}^{N} \mathbf{z}_{1i}\mathbf{z}_{1i}'\right)^{-1}\left(\sum_{i=1}^{N} \mathbf{z}_{1i}\mathbf{x}_i'\right)\right]^{-1} \\
&\quad \cdot \left[\left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \mathbf{z}_{1i}'\right)\left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_{1i}\mathbf{z}_{1i}'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_{1i}\mathbf{x}_i'\boldsymbol{\alpha}_i + \frac{1}{N}\sum_{i=1}^{N} \mathbf{z}_{1i}u_i\right)\right].
\end{aligned} \tag{6.9.10}$$

Although under (6.9.9) $\operatorname{plim} \frac{1}{N}\sum_{i=1}^{n} \mathbf{z}_{1i}u_i = \mathbf{0}$,

$$\begin{aligned}
&\operatorname{plim} \frac{1}{N}\sum_{i=1}^{n} \mathbf{z}_{1i}\mathbf{x}_i'\boldsymbol{\alpha}_i \\
&\quad = E[\mathbf{z}_1 E(\mathbf{x}'\boldsymbol{\alpha} \mid \mathbf{z}_1)] \\
&\quad = E[\mathbf{z}_1 E(\mathbf{x}' \mid \mathbf{z}_1) E(\boldsymbol{\alpha} \mid \mathbf{x}, \mathbf{z}_1)],
\end{aligned} \tag{6.9.11}$$

which is not equal 0 given (6.9.8) and the assumption that $E(\boldsymbol{\alpha} \mid \mathbf{x}) \neq \mathbf{0}$.

To identify $\bar{\boldsymbol{\beta}}$, the variation of $\mathbf{x}_i$ and $\boldsymbol{\beta}_i$ need to be independent conditional on $\mathbf{z}_{1i}$. In other words, we need exclusion restrictions. Heckman and Vytlacil (1998) consider estimating $\bar{\boldsymbol{\beta}}$ assuming the existence of instruments $\mathbf{z}_i = (\mathbf{z}_{1i}, \mathbf{z}_{2i})$ such that

$$\mathbf{x}_i = \Pi \mathbf{z}_{1i} + \mathbf{v}_i \tag{6.9.12}$$

$$\boldsymbol{\beta}_i = \Phi \mathbf{z}_{2i} + \boldsymbol{\eta}_i \tag{6.9.13}$$

where $\mathbf{z}_{1i}$ and $\mathbf{z}_{2i}$ are $m_1 \times 1$ and $m_2 \times 1$ vectors of instruments that satisfy

$$E(u_i \mid \mathbf{z}_i) = 0, \tag{6.9.14}$$

$$E(\boldsymbol{\eta}_i \mid \mathbf{z}_i) = \mathbf{0}, \tag{6.9.15}$$

$$E(\mathbf{v}_i \mid \mathbf{z}_i) = \mathbf{0}, \tag{6.9.16}$$

and $\mathbf{z}_2$ contains elements that are not in $\mathbf{z}_1$.

Then, under (6.9.13), if $\Phi$ is known, an estimator for $\bar{\boldsymbol{\beta}}$ can be obtained from the relation,

$$E(\boldsymbol{\beta}) = \bar{\boldsymbol{\beta}} = \Phi E(\mathbf{z}_2). \tag{6.9.17}$$

Substituting (6.9.12) and (6.9.13) into (6.9.14) yields

$$\begin{aligned} y_i &= \mathbf{x}_i'\boldsymbol{\beta}_i + u_i \\ &= (\mathbf{z}_{1i}'\Pi' + \mathbf{v}_i')(\Phi\mathbf{z}_{2i} + \boldsymbol{\eta}_i) + u_i \\ &= (\mathbf{z}_{1i}'\boldsymbol{\pi}_1\mathbf{z}_{2i}')\boldsymbol{\phi}_1 \ + \ (\mathbf{z}_{1i}'\boldsymbol{\pi}_2\mathbf{z}_{2i}')\boldsymbol{\phi}_2 \\ &\quad + \cdots + (\mathbf{z}_{1i}'\boldsymbol{\pi}_K\mathbf{z}_{2i}')\boldsymbol{\phi}_K \ + \ E(\mathbf{v}_i'\boldsymbol{\eta}_i \mid \mathbf{z}_i) + \epsilon_i^*, \end{aligned} \tag{6.9.18}$$

where $\boldsymbol{\pi}_k'$ and $\boldsymbol{\phi}_k'$ denote the $k$th row of $\Pi$ and $\Phi$, respectively,

$$\boldsymbol{\epsilon}_i^* = \mathbf{v}_i'\Phi\mathbf{z}_{2i} + \boldsymbol{\eta}_i'\Pi\mathbf{z}_{1i} + [\mathbf{v}_i'\boldsymbol{\eta}_i - E(\mathbf{v}_i'\boldsymbol{\eta}_i \mid \mathbf{z}_i)] + u_i. \tag{6.9.19}$$

Under (6.9.12)–(6.9.16), $E(\boldsymbol{\epsilon}_i^* \mid \mathbf{z}_i) = 0$.

Therefore, a consistent estimator of $\Phi$ exists if

$$E(\mathbf{v}_i\boldsymbol{\eta}_i' \mid \mathbf{z}_{1i}, \mathbf{z}_{2i}) = \Sigma_{v\eta} \tag{6.9.20}$$

is not a function of $\mathbf{z}_{1i}$ and $\mathbf{z}_{2i}$, and

$$\operatorname{rank}\left[\frac{1}{N}\sum_{i=1}^{n}\left(\mathbf{z}_{2i}\mathbf{z}_{1i}' \otimes \hat{\Pi}\right)\left(\hat{\Pi}' \otimes \mathbf{z}_{1i}\mathbf{z}_{2i}'\right)\right] = Km_2. \tag{6.9.21}$$

In other words, the necessary condition for the identification of $E(\boldsymbol{\beta}_i) = \bar{\boldsymbol{\beta}}$ for the correlated random-coefficients model (6.9.1)–(6.9.7) when only cross-sectional data are available are that there exist $m_1$ instruments for $\mathbf{x}_i$ and nonzero $m_2$ instruments for $\boldsymbol{\beta}_i$ that satisfy (6.9.12)–(6.9.16) with $m_1^2 > Km_2, m_2 > 0$, and either (6.9.20) holds or $E(\mathbf{v}_i'\boldsymbol{\eta}_i \mid \mathbf{z}_i)$ is known.

The requirements that there exist nonzero $\mathbf{z}_1$ and $\mathbf{z}_2$ with $m_1^2 \geq Km_2$, and (6.9.20) holds are stronger than the usual requirement for the existence of an instrumental variable estimator. As a matter of fact, the necessary condition requires the existence of both $\mathbf{z}_1$ and $\mathbf{z}_2$ (i.e., $m_1 > 0, m_2 > 0$). Neither is (6.9.20) an innocuous assumption. To the best of my knowledge, the conditional covariance independent of $\mathbf{z}$ holds only if $\mathbf{v}$, $\boldsymbol{\eta}$, and $\mathbf{z}$ are joint normal.

### 6.9.3 Estimation of the Mean Effects with Panel Data

When only cross-sectional data are available, the identification conditions of average effects for a correlated random-coefficients model are very stringent and may not be satisfied for many data sets. The instrumental variable approach requires the estimation of a large number of parameters $[(m_1 + m_2)K]$. Multicollinearity and shortages of degrees of freedom could lead to very unreliable estimates. On the other hand, panel data, by blending interindividual differences with intraindividual dynamics can offer several alternatives to get around the difficulties of the correlations between the coefficients and the regressors without the prior conjecture of the existence of certain instruments that satisfy the exclusion restrictions. For ease of exposition, we suppose there are $T$ time series observations of $(y_{it}, \mathbf{x}_{it})$ for each individual $i$. Let $(\mathbf{y}_i', \mathbf{x}_i')$ be the stacked $T$ time series observations of $y_{it}$ and $\mathbf{x}_{it}'$ for each $i$.

#### *6.9.3.1 Group Mean Estimator*

We note that condition on $\mathbf{x}_i, \boldsymbol{\beta}_i$ is a fixed constant. Under (6.9.5), the least-squares estimator of the equation

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta}_i + u_{it}, \quad t = 1, \ldots, T, \tag{6.9.22}$$

yields an unbiased estimator of $\boldsymbol{\beta}_i$, $\hat{\boldsymbol{\beta}}_i$, for each $i$ with covariance matrix $\sigma_i^2(X_i'X_i)^{-1}$ if $u_{it}$ is independently distributed over $t$, where $X_i$ denotes the $T \times K$ stacked $(\mathbf{x}_{it}')$. If $u_{it}$ is independently distributed over $i$, taking the simple average of $\hat{\boldsymbol{\beta}}_i$ as in Hsiao, Pesaran, and Tahmiscioglu (1999),

$$\hat{\bar{\boldsymbol{\beta}}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\boldsymbol{\beta}}_i. \tag{6.9.23}$$

yields a consistent estimator of $\bar{\boldsymbol{\beta}}$ as $N \to \infty$. If $T > K$, the estimator (6.9.23) is consistent and asymptotically normally distributed as $N \to \infty$, and $\sqrt{N}(\hat{\bar{\boldsymbol{\beta}}} - \bar{\boldsymbol{\beta}})$ is asymptotically normally distributed with mean $\mathbf{0}$ and covariance matrix

$$\text{Asy Cov}\,(\hat{\bar{\boldsymbol{\beta}}}) = \left[\Delta + \frac{1}{N}\sum_{i=1}^{N}\sigma_i^2(X_i'X_i)^{-1}\right], \tag{6.9.24}$$

if $u_{it}$ is independently distributed over $i$ and $t$ with variance $\sigma_i^2$.

#### *6.9.3.2 Conventional Fixed-Effects Estimator*

The estimator (6.9.23) is simple to implement. However, if $T < K$, we cannot estimate $\boldsymbol{\beta}_i$ using the $i$th individual's time series observations $(y_i, \mathbf{x}_i')$. Nevertheless, the conventional fixed-effects estimator can still allow us to obtain consistent estimator of $\bar{\boldsymbol{\beta}}$ in a number of situations.

Let $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$ and $\bar{\mathbf{x}}_i = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_{it}$. The conventional fixed-effects estimator first takes the deviation of each observation from its time series mean,

and then regresses $(y_{it} - \bar{y}_i)$ on $(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)$ (e.g., Chapter 3). Model (6.9.2) leads to

$$(y_{it} - \bar{y}_i) = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\bar{\boldsymbol{\beta}} + (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\boldsymbol{\alpha}_i + (u_{it} - \bar{u}_i), \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{matrix} \quad (6.9.25)$$

where $\bar{u}_i = \frac{1}{T}\sum_{t=1}^{T} u_{it}$.

The fixed-effects estimator (6.9.25) will converge to $\bar{\boldsymbol{\beta}}$ provided

$$\text{plim } \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\boldsymbol{\alpha}_i = \mathbf{0} \quad (6.9.26)$$

and

$$\text{plim } \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(u_{it} - \bar{u}_i) = \mathbf{0} \quad (6.9.27)$$

Under (6.9.5), (6.9.27) holds. Under the assumption that $\mathbf{x}_{it}$ and $\boldsymbol{\alpha}_i$ are linearly related with finite variance, then $\mathbf{x}_{it} - \bar{\mathbf{x}}_i$ does not involve $\boldsymbol{\alpha}_i$, and hence (6.9.26) holds by a law of large numbers. Hence the conventional fixed-effects estimator is $\sqrt{N}$ consistent and asymptotically normally distributed as $N \to \infty$. The asymptotic covariance matrix of the conventional fixed-effects estimator (3.2.5) can be approximated using the Newey–West heteroscedasticity-autocorrelation consistent formula (Vogelsang 2012).

When $(\mathbf{x}_{it}, \boldsymbol{\alpha}_i)$ jointly have an elliptical distribution[21] (e.g., Fang and Zhang 1990; Gupta and Varga 1993), $\mathbf{x}_{it}$ and $\boldsymbol{\alpha}_i$ are linearly related. Another case that the fixed-effects estimator can be consistent is that $(\mathbf{x}_{it}, \boldsymbol{\alpha}_i)$ are jointly symmetrically distributed; then $\frac{1}{NT}\sum_{i=1}^{N}(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\boldsymbol{\alpha}_i$ will converge to 0 even though $\mathbf{x}_{it}$ have a mean different from 0.

### *6.9.3.3 Panel Pooled Least-Squares Estimator*

The conventional fixed-effects (FE) estimator (3.2.5) can yield a consistent estimator of $\bar{\boldsymbol{\beta}}$, under certain conditions. However, if $\boldsymbol{\alpha}$ and $\mathbf{x}_{it}$ are not linearly related, it is inconsistent. Moreover, if $\mathbf{x}_{it}$ contains time-invariant variables, then the mean effects of time-invariant variables cannot be identified by the conventional fixed-effects estimator. Furthermore, the FE estimator only makes use of within- (group) variation. Because in general the between-group variation is much larger than within-group variation, the FE estimator could also mean a loss of efficiency. To get around these limitations on the FE estimator as well as allowing the case that $\boldsymbol{\alpha}_i$ and $\mathbf{x}_{it}$ are not linearly related, Hsiao, Li, Liang, and Xie (2012) suggest a modified specification to obtain the estimate of the mean effects.

[21] Many commonly assumed distributions such as uniform, normal, Student's $t$, double exponential, etc. belong to the family of elliptical distributions.

To illustrate the basic idea, we first assume that $E(\alpha_i \mid \mathbf{x}_i)$ is a linear function of $\mathbf{x}_i$. We will show later that a similar procedure can be applied if $E(\boldsymbol{\alpha}_i \mid \mathbf{x}_i)$ is a function of a higher order of $\mathbf{x}_i$.

A6.9.1: $(\mathbf{x}_i', \boldsymbol{\alpha}_i')$ are independently, identically distributed across $i$ with

$$E(\boldsymbol{\alpha}_i \mid \mathbf{x}_i) = \mathbf{a} + B\mathbf{x}_i, \tag{6.9.28}$$

where $\mathbf{a}$ and $B$ are the $K \times 1$ and $K \times TK$ constant vector and matrix, respectively.

From (6.9.3) and (6.9.28), we have

$$E_x[E(\boldsymbol{\alpha}_i \mid \mathbf{x}_i)] = \mathbf{a} + BE(\mathbf{x}_i) = \mathbf{0}. \tag{6.9.29}$$

It follows that

$$E(\boldsymbol{\alpha}_i \mid \mathbf{x}_i) = B(\mathbf{x}_i - E\mathbf{x}_i) \tag{6.9.30}$$

Substituting

$$\boldsymbol{\alpha}_i = E(\boldsymbol{\alpha}_i \mid \mathbf{x}_i) + \boldsymbol{\omega}_i \tag{6.9.31}$$

and (6.9.30) into (6.9.1) yields

$$y_{it} = \mathbf{x}_{it}'\bar{\boldsymbol{\beta}} + \mathbf{x}_{it}'B(\mathbf{x}_i - E\mathbf{x}_i) + v_{it}^*, \tag{6.9.32}$$

where

$$v_{it}^* = \mathbf{x}_{it}'\boldsymbol{\omega}_i + u_{it}. \tag{6.9.33}$$

By construction, $E(v_{it}^* \mid \mathbf{x}_i) = 0$. Therefore, the least-squares regression of

$$\mathbf{y}_i = X_i\bar{\boldsymbol{\beta}} + X_i \otimes (\mathbf{x}_i - \bar{\mathbf{x}})' \operatorname{vec}(B') + \mathbf{v}_i \tag{6.9.34}$$

yields $\sqrt{N}$ consistent and asymptotically normally distributed estimator of $\bar{\boldsymbol{\beta}}$ when $N \to \infty$, where $\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$, and $\mathbf{v}_i^* = (v_{i1}^*, \ldots, v_{iT}^*)'$ as long as

$$\frac{1}{N}\sum_{i=1}^{N}\begin{bmatrix} X_i'X_i & X_i'X_i \otimes (\mathbf{x}_i' - \bar{\mathbf{x}}') \\ (\mathbf{x}_i - \bar{\mathbf{x}}) \otimes X_i'X_i & (\mathbf{x}_i - \bar{\mathbf{x}}) \otimes X_i'X_i \otimes (\mathbf{x}_i - \bar{\mathbf{x}})' \end{bmatrix} \tag{6.9.35}$$

is a full rank matrix.

However, because

$$E\mathbf{v}_i\mathbf{v}_i' = X_i\Delta^*X_i' + \sigma_i^2 I_T, \tag{6.9.36}$$

where $\Delta^* = E(\boldsymbol{\omega}_i\boldsymbol{\omega}_i')$, a more efficient estimator of $\bar{\boldsymbol{\beta}}$ will be a generalized least-squares estimator (GLS) if $\Delta^*$ and $\sigma_i^2$ are known. If $\Delta^*$ and $\sigma_i^2$ are unknown, we can apply the feasible GLS (FGLS) through a two-step procedure.

Similar reasoning can be applied if $E(\boldsymbol{\alpha}_i \mid \mathbf{x}_i)$ is a higher order polynomial of $\mathbf{x}_i$, say

$$E(\boldsymbol{\alpha}_i \mid \mathbf{x}_i) = \mathbf{a} + B\mathbf{x}_i + C\mathbf{x}_i \otimes \mathbf{x}_i. \tag{6.9.37}$$

then from $E_x[E(\boldsymbol{\alpha}_i \mid \mathbf{x}_i)] = 0$, it follows that then the least-squares regression of

$$\begin{aligned} y_{it} = \mathbf{x}_{it}'\bar{\boldsymbol{\beta}} &+ \mathbf{x}_{it}' \otimes (\mathbf{x}_i - E\mathbf{x}_i)' \operatorname{vec}(B') \\ &+ \mathbf{x}_{it}' \otimes [(\mathbf{x}_i \otimes \mathbf{x}_i) - E(\mathbf{x}_i \otimes \mathbf{x}_i)]' \operatorname{vec}(C') \\ &+ v_{it}, \end{aligned} \tag{6.9.38}$$

is consistent when $N \to \infty$, where $v_{it} = \mathbf{x}_{it}'\boldsymbol{\omega}_i + u_{it}$ and $\boldsymbol{\omega}_i = \boldsymbol{\alpha}_i - E(\boldsymbol{\alpha}_i \mid \mathbf{x}_i)$.

LS (or FGLS) regression of (6.9.38) not only requires the estimation of a large number of parameters, but it could also raise the issue of multicollinearity. However, if $\mathbf{x}_{it}$ is stationary, a more parsimonious approximation would be to follow Mundlak (1978a) to replace (6.9.37) by

$$E(\boldsymbol{\alpha}_i \mid \mathbf{x}_i) = \mathbf{a} + \tilde{B}\bar{\mathbf{x}}_i + \tilde{C}(\bar{\mathbf{x}}_i \otimes \bar{\mathbf{x}}_i), \tag{6.9.39}$$

where $\bar{\mathbf{x}}_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{it}$ and regressing

$$\begin{aligned} y_{it} = \mathbf{x}_{it}'\bar{\boldsymbol{\beta}} &+ \mathbf{x}_{it}' \otimes (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})' \operatorname{vec}(\tilde{B}') \\ &+ \mathbf{x}_{it}' \otimes [(\bar{\mathbf{x}}_i \otimes \bar{\mathbf{x}}_i) - (\bar{\mathbf{x}} \otimes \bar{\mathbf{x}})]' \operatorname{vec}(\tilde{C}') \\ &+ v_{it}, \end{aligned} \tag{6.9.40}$$

where $\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$, $(\bar{\mathbf{x}} \otimes \bar{\mathbf{x}}) = \frac{1}{N}\sum_{i=1}^{N}[\bar{\mathbf{x}}_i \otimes \bar{\mathbf{x}}_i]$.

**Remark 6.9.1:** When $\mathbf{x}_{it}$ contains an intercept term, (6.9.35) is not a full rank matrix. Let $\mathbf{x}_{it}' = (1, \tilde{\mathbf{x}}_{it})$, where $\tilde{\mathbf{x}}_{it}$ denotes the $(1 \times (K-1))$ time-varying explanatory variables. Let $\boldsymbol{\alpha}_i' = (\alpha_{1i}, \tilde{\boldsymbol{\alpha}}_i')$ and $\bar{\boldsymbol{\beta}}' = (\bar{\boldsymbol{\beta}}_1, \tilde{\bar{\boldsymbol{\beta}}})$ be the corresponding partitions, Rewrite (6.9.28) in the form

$$E\begin{pmatrix} \alpha_{1i} & | & \\ & | & \mathbf{x}_{it} \\ \tilde{\boldsymbol{\alpha}}_i & | & \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{b}}_1' \\ \tilde{B} \end{pmatrix} (\tilde{\mathbf{x}}_{it} - E\tilde{\mathbf{x}}_{it}). \tag{6.9.41}$$

Then

$$\begin{aligned} E(\mathbf{y}_i \mid X_i) = X_i\bar{\boldsymbol{\beta}} &+ \mathbf{e}(\mathbf{x}_i - E\bar{\tilde{\mathbf{x}}}_i)'\tilde{\mathbf{b}}_1^* \\ &+ \tilde{X}_i \otimes [\tilde{X}_i - E(\tilde{X}_i)]' \operatorname{vec}(\hat{B}'), \end{aligned} \tag{6.9.42}$$

where $\tilde{\mathbf{b}}_1^* = T\mathbf{b}_1$, $\tilde{X}_i$ is the $T \times (K-1)$ stacked $\tilde{\mathbf{x}}_{it}'$ and $\mathbf{e}$ is a $(T \times 1)$ vector of 1's.

Therefore a consistent estimator of $\bar{\boldsymbol{\beta}}' = (\bar{\boldsymbol{\beta}}_1, \tilde{\bar{\boldsymbol{\beta}}})$ can be obtained by the least-squares regression of

$$\begin{aligned} y_{it} = \bar{\beta}_1 + \tilde{\mathbf{x}}_{it}'\tilde{\bar{\boldsymbol{\beta}}} &+ (\bar{\tilde{\mathbf{x}}}_i - \bar{\tilde{\mathbf{x}}})'\tilde{\mathbf{b}}_1^* \\ &+ (\tilde{\mathbf{x}}_{it}' \otimes (\tilde{\mathbf{x}}_{it} - \bar{\tilde{\mathbf{x}}})') \operatorname{vec}(\tilde{B}') + v_{it}, \end{aligned} \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T. \end{matrix} \tag{6.9.43}$$

*6.9.3.4 Semiparametric Estimates*

The application of group mean estimator (6.9.23) requires precise $T > K$. If $T \leq K$ (Hsiao, Li, Liang, and Xie 2012) suggest a semiparametric estimator if there exists a $q$-dimensional random variable $\mathbf{z}_i (q < T \leq K)$ such that conditional on $\mathbf{z}_i$, $\boldsymbol{\beta}_i$ and $\mathbf{x}_{it}$ are conditionally independent, $(\boldsymbol{\beta}_i \perp \mathbf{x}_{it} \mid \mathbf{z}_i)$, and $E[\mathbf{x}_{it}\mathbf{x}_{it}' \mid \mathbf{z}_i]$ is a full rank matrix. Then $E(\boldsymbol{\alpha}_i \mid \mathbf{x}_i, \mathbf{z}_i) = E(\boldsymbol{\alpha}_i \mid \mathbf{z}_i) \equiv \mathbf{g}(\mathbf{z}_i)$. The random variable $\mathbf{z}_i$ can contain components of $\mathbf{x}_i$; it could be a function of $\mathbf{x}_i$; say, the propensity score (Rosenbaum and Rubin 1983); or an instrument for $\mathbf{x}_{it}$ and $\boldsymbol{\beta}_i$ (e.g., Heckman and Vytlacil 2001, 2005, 2007; Heckman, Schmierer, and Urzua 2011); or simply the time series mean of the $i$th individual's $\mathbf{x}_{it}$, $\bar{\mathbf{x}}_i$ (Mundlak 1978a).

$$\begin{aligned} y_{it} &= \mathbf{x}_{it}'\bar{\boldsymbol{\beta}} + \mathbf{x}_{it}'\mathbf{g}(\mathbf{z}_i) + \epsilon_{it} \\ &= \mathbf{x}_{it}'\boldsymbol{\theta}(\mathbf{z}_i) + \epsilon_{it}, \end{aligned} \tag{6.9.44}$$

where $\boldsymbol{\theta}(\mathbf{z}_i) = \bar{\boldsymbol{\beta}} + \mathbf{g}(\mathbf{z}_i)$ and $E\mathbf{g}(\mathbf{z}_i) = \mathbf{0}$. Given $\boldsymbol{\theta}(\mathbf{z}_i)$ and $E(\boldsymbol{\theta}(\mathbf{z}_i)) = \bar{\boldsymbol{\beta}}$, a consistent estimator of $\bar{\boldsymbol{\beta}}$ is

$$\hat{\bar{\boldsymbol{\beta}}}_{semi} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\theta}(\mathbf{z}_i). \tag{6.9.45}$$

Hsiao, Li, Liang, and Xie (2012) suggest two types of semiparametric estimators for $\boldsymbol{\theta}(\mathbf{z})$: local constant and local polynomial estimation methods. The local constant estimator of $\boldsymbol{\theta}(\mathbf{z})$ is given by

$$\hat{\boldsymbol{\theta}}_{LC}(\mathbf{z}) = \left(\sum_{j=1}^{N}\sum_{s=1}^{T} x_{js}x_{js}' K_{h,z_{jz}}\right)^{-1} \sum_{j=1}^{N}\sum_{s=1}^{T} x_{js}y_{js}K_{h,z_{jz}}, \tag{6.9.46}$$

where $K_{h,z_{jz}} = \Pi_{l=1}^{q} k(\frac{z_{jl}-z_l}{h_l})$ is the product kernel, $k(\cdot)$ is the univariate kernel function, and $z_{jl}$ and $z_l$ are the $l$th-component of $\mathbf{z}_j$ and $\mathbf{z}$, respectively.

The local polynomial estimation minimizes the kernel weighted sum of squared errors

$$\sum_{j=1}^{N}\sum_{s=1}^{T}\left[y_{js} - \sum_{0\leq|k|\leq p} x_{js}' b_k(\mathbf{z})(\mathbf{z}_j - \mathbf{z})^k\right]^2 K_{h,z_{jz}}, \tag{6.9.47}$$

with respect to each $b_k(\mathbf{z})$ which gives an estimate of $\hat{b}_k(\mathbf{z})$, and $k\hat{b}_k(\mathbf{z})$ estimates $D^k\boldsymbol{\theta}(\mathbf{z})$. Thus, $\hat{\boldsymbol{\theta}}_{LP} = \hat{\mathbf{b}}_0(\mathbf{z})$ is the $p$th order local polynomial estimator of $\boldsymbol{\theta}(\mathbf{z})$.

Simple Monte Carlo studies conducted by Hsiao et al. (2012) show that if the exact order of $E(\boldsymbol{\alpha}_i \mid \mathbf{x}_i)$ is known, the panel pooled least-squares estimator performs well. If $E(\boldsymbol{\alpha}_i \mid \mathbf{x}_i)$ is unknown, the group mean estimator ((6.9.23) or (6.9.45)) semiparametric estimator is robust to a variety of joint distribution of $(\boldsymbol{\alpha}_i, \mathbf{x}_{it})$, but not the conventional fixed-effects estimator.

## APPENDIX 6A: COMBINATION OF TWO NORMAL DISTRIBUTIONS

Suppose that conditional on $X$, $\boldsymbol{\beta}$, $\mathbf{y} \sim N(X\boldsymbol{\beta}, \Omega)$ and $\boldsymbol{\beta} \sim N(A\bar{\boldsymbol{\beta}}, C)$. Then the posterior of $\boldsymbol{\beta}$ and $\bar{\boldsymbol{\beta}}$ given $\mathbf{y}$ is

$$P(\boldsymbol{\beta}, \bar{\boldsymbol{\beta}} \mid \mathbf{y}) \propto \exp\left[-\frac{1}{2}\{(\mathbf{y} - X\boldsymbol{\beta})'\Omega^{-1}(\mathbf{y} - X\boldsymbol{\beta}) + (\boldsymbol{\beta} - A\bar{\boldsymbol{\beta}})'C^{-1}(\boldsymbol{\beta} - A\bar{\boldsymbol{\beta}})\}\right], \tag{6A.1}$$

where "$\propto$" denotes "proportionality." Using the identity (e.g., Rao 1971, p. 33)

$$(D + BFB')^{-1} = D^{-1} - D^{-1}B(B'D^{-1}B + F^{-1})^{-1}B'D^{-1}. \tag{6A.2}$$

and

$$(D + F)^{-1} = D^{-1} - D^{-1}(D^{-1} + F^{-1})^{-1}D^{-1}, \tag{6A.3}$$

we can complete the squares of

$$\begin{aligned}(\boldsymbol{\beta} - A\bar{\boldsymbol{\beta}})'C^{-1}(\boldsymbol{\beta} - A\bar{\boldsymbol{\beta}}) + (\mathbf{y} - X\boldsymbol{\beta})'\Omega^{-1}(\mathbf{y} - X\boldsymbol{\beta}) \\ = \boldsymbol{\beta}'C^{-1}\boldsymbol{\beta} + \bar{\boldsymbol{\beta}}'A'C^{-1}A\bar{\boldsymbol{\beta}} - 2\boldsymbol{\beta}'C^{-1}A\bar{\boldsymbol{\beta}} \\ + \mathbf{y}'\Omega^{-1}\mathbf{y} + \boldsymbol{\beta}'X'\Omega^{-1}X\boldsymbol{\beta} - 2\boldsymbol{\beta}'X'\Omega^{-1}\mathbf{y}.\end{aligned} \tag{6A.4}$$

Let

$$\begin{aligned}Q_1 = [\boldsymbol{\beta} - (X'\Omega^{-1}X + C^{-1})^{-1}(X\Omega^{-1}\mathbf{y} + C^{-1}A\bar{\boldsymbol{\beta}})]'(C^{-1} + X'\Omega^{-1}X) \\ \cdot [\boldsymbol{\beta} - (X'\Omega^{-1}X + C^{-1})^{-1}(X'\Omega^{-1}\mathbf{y} + C^{-1}A\bar{\boldsymbol{\beta}})].\end{aligned} \tag{6A.5}$$

then

$$\begin{aligned}&\boldsymbol{\beta}'C^{-1}\boldsymbol{\beta} + \boldsymbol{\beta}'X'\Omega^{-1}X\boldsymbol{\beta} - 2\boldsymbol{\beta}'C^{-1}A\bar{\boldsymbol{\beta}} - 2\boldsymbol{\beta}'X'\Omega^{-1}\mathbf{y} \\ &= Q_1 - (X'\Omega^{-1}\mathbf{y} + C^{-1}A\bar{\boldsymbol{\beta}})'(X'\Omega^{-1}X + C^{-1})^{-1}(X'\Omega^{-1}\mathbf{y} + C^{-1}A\bar{\boldsymbol{\beta}}).\end{aligned} \tag{6A.6}$$

Substituting (6A.6) into (6A.4) yields

$$\begin{aligned}&Q_1 + \mathbf{y}'[\Omega^{-1} - \Omega^{-1}X(X'\Omega^{-1}X + C^{-1})^{-1}X'\Omega^{-1}]\mathbf{y} \\ &\quad + \bar{\boldsymbol{\beta}}'A'[C^{-1} - C^{-1}(X'\Omega^{-1}X + C^{-1})^{-1}C^{-1}]A\bar{\boldsymbol{\beta}} \\ &\quad - 2\bar{\boldsymbol{\beta}}'A'C^{-1}(X'\Omega^{-1}X + C^{-1})^{-1}X'\Omega^{-1}\mathbf{y} \\ &\quad = Q_1 + \mathbf{y}'(XCX' + \Omega)^{-1}\mathbf{y} + \bar{\boldsymbol{\beta}}'A'X'(XCX' + \Omega)^{-1}XA\bar{\boldsymbol{\beta}} \\ &\quad\quad - 2\bar{\boldsymbol{\beta}}'A'X'(XCX' + \Omega)^{-1}\mathbf{y} \\ &\quad = Q_1 + Q_2 + Q_3\end{aligned} \tag{6A.7}$$

where

$$Q_2 = \{\bar{\boldsymbol{\beta}} - [A'X'(XCX' + \Omega)^{-1}XA]^{-1}[A'X'(XCX' + \Omega)^{-1}y]\}'$$
$$\cdot [A'X'(XCX' + \Omega)^{-1}XA] \tag{6A.8}$$
$$\cdot \{\bar{\boldsymbol{\beta}} - [A'X'(XCX' + \Omega)^{-1}XA]^{-1}[A'X'(XCX' + \Omega)^{-1}\mathbf{y}]\},$$

$$Q_3 = \mathbf{y}'\{(XCX' + \Omega)^{-1} - (XCX' + \Omega)^{-1}XA[A'X(XCX' + \Omega)^{-1}XA]^{-1}$$
$$\cdot A'X'(XCX' + \Omega)^{-1}\}\ \mathbf{y} \tag{6A.9}$$

Since $Q_3$ is a constant independent of $\boldsymbol{\beta}$ and $\bar{\boldsymbol{\beta}}$, we can write $P(\boldsymbol{\beta}, \bar{\boldsymbol{\beta}} \mid \mathbf{y})$ in the form of $P(\boldsymbol{\beta} \mid \bar{\boldsymbol{\beta}}, \mathbf{y})P(\bar{\boldsymbol{\beta}} \mid \mathbf{y})$, which becomes

$$P\{\boldsymbol{\beta}, \bar{\boldsymbol{\beta}} \mid \mathbf{y}\} \propto \exp\left\{-\frac{1}{2}Q_1\right\} \exp\left\{-\frac{1}{2}Q_2\right\} \tag{6A.10}$$

where $\exp\{-\frac{1}{2}Q_1\}$ is proportional to $P(\boldsymbol{\beta} \mid \bar{\boldsymbol{\beta}}, \mathbf{y})$ and $\exp\{-\frac{1}{2}Q_2\}$ is proportional to $P(\bar{\boldsymbol{\beta}} \mid \mathbf{y})$. That is, $P(\boldsymbol{\beta} \mid \bar{\boldsymbol{\beta}}, \mathbf{y})$ is $N\{(X'\Omega^{-1}X + C^{-1})^{-1}(X'\Omega^{-1}\mathbf{y} + C^{-1}A\bar{\boldsymbol{\beta}}), (C^{-1} + X'\Omega^{-1}X)^{-1}\}$ and $P(\bar{\boldsymbol{\beta}} \mid \mathbf{y})$ is $N\{[A'X'(XCX' + \Omega)^{-1}XA]^{-1}[A'X'(XCX' + \Omega)^{-1}\mathbf{y}]^{-1}, [A'X'(XCX' + \Omega)^{-1}XA]^{-1}\}$.

Alternatively, we may complete the square of the left side of (6A.4) with the aim of writing $P(\boldsymbol{\beta}, \bar{\boldsymbol{\beta}} \mid \mathbf{y})$ in the form of $P(\bar{\boldsymbol{\beta}} \mid \boldsymbol{\beta}, \mathbf{y})P(\boldsymbol{\beta} \mid \mathbf{y})$,

$$Q_4 + \boldsymbol{\beta}'[X'\Omega^{-1}X + C^{-1} - C^{-1}A(A'CA)^{-1}A'C^{-1}]\boldsymbol{\beta}$$
$$- 2\boldsymbol{\beta}'X'\Omega^{-1}\mathbf{y} + \mathbf{y}'\Omega^{-1}\mathbf{y} \tag{6A.11}$$
$$= Q_4 + Q_5 + Q_3,$$

where

$$Q_4 = [\bar{\boldsymbol{\beta}} - (A'C^{-1}A)^{-1}A'C^{-1}\boldsymbol{\beta}]'(A'C^{-1}A)$$
$$\cdot [\bar{\boldsymbol{\beta}} - (A'C^{-1}A)^{-1}A'C^{-1}\boldsymbol{\beta}], \tag{6A.12}$$

$$Q_5 = [\boldsymbol{\beta} - D^{-1}X'\Omega^{-1}\mathbf{y}]'D[\boldsymbol{\beta} - D^{-1}X'\Omega^{-1}\mathbf{y}]. \tag{6A.13}$$

and

$$D = X'\Omega^{-1}X + C^{-1} - C^{-1}A(A'C^{-1}A)^{-1}A'C^{-1}. \tag{6A.14}$$

Therefore, $P(\bar{\boldsymbol{\beta}} \mid \boldsymbol{\beta}, \mathbf{y}) \sim N\{(A'C^{-1}A)^{-1}C^{-1}\boldsymbol{\beta}, (A'C^{-1}A)^{-1}\}$ and $P(\boldsymbol{\beta} \mid \mathbf{y}) \sim N\{D^{-1}X'\Omega^{-1}\ \mathbf{y}, D^{-1}\}$.

CHAPTER 7

# Discrete Data

## 7.1 INTRODUCTION

In this chapter we consider situations in which an analyst has at his disposal a random sample of $N$ individuals, having recorded histories indicating the presence or absence of an event in each of $T$ equally spaced discrete time periods. Statistical models in which the endogenous random variables take only discrete values are known as discrete, categorical, qualitative-choice, or quantal-response models. The literature, both applied and theoretical, on this subject is vast. Amemiya (1981), Maddala (1983), and McFadden (1976, 1984) have provided excellent surveys. Thus, the focus of this chapter is only on controlling for unobserved characteristics of individual units to avoid specification bias. In Section 7.2, we briefly review some popular parametric specifications for cross-sectional data. Sections 7.3 and 7.4 discuss inference of panel parametric and semiparametric static models with heterogeneity, respectively. Section 7.5 discusses dynamic models. Section 7.6 discusses alternative approaches to identify state dependence.

## 7.2 SOME DISCRETE-RESPONSE MODELS FOR CROSS-SECTIONAL DATA

In this section we briefly review some widely used discrete-response models for cross-sectional data. We suppose there are observations for $K + 1$ variables $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, N$, where the dependent variable $y_i$ can take only two values, which for convenience and without any loss of generality will be the value of 1 if an event occurs and 0 if it does not. Examples of this include purchases of durables in a given year, participation in the labor force, the decision to enter college, and the decision to marry.

The discrete outcome of $y_i$ can be viewed as the observed counterpart of a latent continuous random variable crossing a threshold. Suppose that the continuous latent random variable, $y_i^*$, is a linear function of a vector of explanatory variable, $\mathbf{x}_i$,

$$y_i^* = \beta' \mathbf{x}_i + v_i, \tag{7.2.1}$$

where the error term $v_i$ is independent of $\mathbf{x}_i$ with mean 0. Suppose, instead of observing $y_i^*$, we observe $y_i$, where

$$y_i = \begin{cases} 1, & \text{if} \quad y_i^* > 0, \\ 0, & \text{if} \quad y_i^* \leq 0. \end{cases} \tag{7.2.2}$$

The expected value of $y_i$ is then the probability that the event will occur,

$$\begin{aligned} E(y_i \mid \mathbf{x}_i) &= 1 \cdot Pr(v_i > -\beta'\mathbf{x}_i) + 0 \cdot Pr(v_i \leq -\beta'\mathbf{x}_i) \\ &= Pr(v_i > -\beta'\mathbf{x}_i) \\ &= Pr(y_i = 1 \mid \mathbf{x}_i). \end{aligned} \tag{7.2.3}$$

When the probability law of generating $v_i$ follows a two-point distribution $(1 - \beta'\mathbf{x}_i)$ and $(-\beta'\mathbf{x}_i)$, with probabilities $\beta'\mathbf{x}_i$ and $(1 - \beta'\mathbf{x}_i)$, respectively, we have the linear-probability model

$$y_i = \beta'\mathbf{x}_i + v_i \tag{7.2.4}$$

with $Ev_i = \beta'\mathbf{x}_i(1 - \beta'\mathbf{x}_i) + (1 - \beta'\mathbf{x}_i)(-\beta'\mathbf{x}_i) = 0$. When the probability density function of $v_i$ is a standard normal density function, $\frac{1}{\sqrt{2\pi}} \exp(-\frac{v^2}{2}) = \phi(v)$, we have the probit model,

$$\begin{aligned} Pr(y_i = 1 \mid \mathbf{x}_i) &= \int_{-\beta'\mathbf{x}_i}^{\infty} \phi(v_i) dv_i \\ &= \int_{-\infty}^{\beta'\mathbf{x}_i} \phi(v_i) dv_i = \Phi(\beta'\mathbf{x}_i). \end{aligned} \tag{7.2.5}$$

When the probability density function is a standard logistic,

$$\frac{\exp(v_i)}{(1 + \exp(v_i))^2} = [(1 + \exp(v_i))(1 + \exp(-v_i))]^{-1}$$

we have the logit model

$$Pr(y_i = 1 \mid \mathbf{x}_i) = \int_{-\beta'\mathbf{x}_i}^{\infty} \frac{\exp(v_i)}{(1 + \exp(v_i))^2} dv_i = \frac{\exp(\beta'\mathbf{x}_i)}{1 + \exp(\beta'\mathbf{x}_i)}. \tag{7.2.6}$$

Letting $F(\beta'\mathbf{x}_i) = E(y_i \mid \mathbf{x}_i)$, the three commonly used parametric models for the binary choice may be summarized with a single index $w$ as:

*Linear-Probability Model*

$$F(w) = w. \tag{7.2.7}$$

*Probit model*

$$F(w) = \int_{-\infty}^{w} \frac{1}{\sqrt{2\pi}} e^{-\frac{v^2}{2}} dv = \Phi(w) \tag{7.2.8}$$

*Logit model*

$$F(w) = \frac{e^w}{1 + e^w}. \tag{7.2.9}$$

The linear-probability model is a special case of the linear regression model with heteroscadastic variance, $\beta'\mathbf{x}_i(1-\beta'\mathbf{x}_i)$. It can be estimated by least-squares or weighted least-squares (Goldberger 1964). But it has an obvious defect in that $\beta'\mathbf{x}_i$ is not constrained to lie between 0 and 1 as a probability should, whereas the probit and logit models do. The probability functions used for the probit and logit models are the standard normal distribution and the logistic distribution, respectively. We use cumulative standard normal because in the dichotomy case, the probability an event occurs depends only on $(\frac{1}{\sigma})\beta'\mathbf{x}$, where $\sigma$ denotes the standard deviation of a normal density. There is no way to identify the variance of a normal density. The logit probability density function is symmetric around 0 and has a variance of $\pi^2/3$. Because they are distribution functions, the probit and logit models are bounded between 0 and 1.

The cumulative normal distribution and the logistic distribution are very close to each other, except that the logistic distribution has sightly heavier tails (Cox, 1970). Moreover, the cumulative normal distribution $\Phi$ is reasonably well approximated by a linear function for the range of probabilities between 0.3 and 0.7. Amemiya (1981) has suggested an approximate conversion rule for the coefficients of these three models. Let the coefficients for the linear-probability, probit, and logit models be denoted as $\hat{\beta}_{LP}$, $\hat{\beta}_{\Phi}$, $\hat{\beta}_{L}$, respectively. Then

$$\hat{\beta}_L \simeq 1.6\,\hat{\beta}_{\Phi},$$

$$\hat{\beta}_{LP} \simeq 0.4\,\hat{\beta}_{\Phi} \text{ except for the constant term,} \tag{7.2.10}$$

and

$$\hat{\beta}_{LP} \simeq 0.4\,\hat{\beta}_{\Phi} + 0.5 \text{ for the constant term.}$$

For random sample of $N$ individuals, the likelihood function for these three models can be written in general form as

$$L = \prod_{i=1}^{N} F(\beta'\mathbf{x}_i)^{y_i}[1-F(\beta'\mathbf{x}_i)]^{1-y_i}. \tag{7.2.11}$$

Differentiating the logarithm of the likelihood function yields the vector of first derivatives and the matrix of second-order derivatives as

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \frac{y_i - F(\boldsymbol{\beta}'\mathbf{x}_i)}{F(\boldsymbol{\beta}'\mathbf{x}_i)[1-F(\boldsymbol{\beta}'\mathbf{x}_i)]} F'(\boldsymbol{\beta}'\mathbf{x}_i)\mathbf{x}_i \tag{7.2.12}$$

$$\frac{\partial \log L}{\partial \beta \partial \beta'} = \sum_{i=1}^{N} \left\{ -\left[ \frac{y_i}{F^2(\beta'\mathbf{x}_i)} + \frac{1-y_i}{[1-F(\beta'\mathbf{x}_i)]^2} \right] [F'(\beta'\mathbf{x}_i)]^2 + \left[ \frac{y_i - F(\beta'\mathbf{x}_i)}{F(\beta'\mathbf{x}_i)[1-F(\beta'\mathbf{x}_i)]} \right] F''(\beta'\mathbf{x}_i) \right\} \mathbf{x}_i\mathbf{x}_i' \tag{7.2.13}$$

where $F'(\beta'\mathbf{x}_i)$ and $F''(\beta'\mathbf{x}_i)$ denote the first and second derivatives of $F(\beta'\mathbf{x}_i)$ with respect to $\beta'\mathbf{x}_i$. If the likelihood function (7.2.11) is concave, as in the models discussed here (e.g., Amemiya 1985, p. 273), a Newton–Raphson method,

$$\hat{\beta}^{(j)} = \hat{\beta}^{(j-1)} - \left(\frac{\partial^2 \log L}{\partial\beta\partial\beta'}\right)^{-1}_{\beta=\hat{\beta}^{(j-1)}} \left(\frac{\partial \log L}{\partial\beta}\right)_{\beta=\hat{\beta}^{(j-1)}}, \tag{7.2.14}$$

or a method of scoring,

$$\hat{\beta}^{(j)} = \hat{\beta}^{(j-1)} - \left[E\frac{\partial^2 \log L}{\partial\beta\partial\beta'}\right]^{-1}_{\beta=\hat{\beta}^{(j-1)}} \left(\frac{\partial \log L}{\partial\beta}\right)_{\beta=\hat{\beta}^{(j-1)}}, \tag{7.2.15}$$

can be used to find the maximum-likelihood estimator of $\beta$, where $\hat{\beta}^{(j)}$ denotes the $j$th iterative solution.

In the case in which there are repeated observations of $y$ for a specific value of $\mathbf{x}$, the proportion of $y = 1$ for individuals with the same characteristic, $\mathbf{x}$, is a consistent estimator of $p = F(\beta'\mathbf{x})$. Taking the inverse of this function yields $F^{-1}(p) = \beta'\mathbf{x}$. Substituting $\hat{p}$ for $p$, we have $F^{-1}(\hat{p}) = \beta'\mathbf{x} + \zeta$, where $\zeta$ denotes the approximation error of using $F^{-1}(\hat{p})$ for $F^{-1}(p)$. Since $\zeta$ has a nonscalar covariance matrix, we can apply the weighted-least-squares method to estimate $\beta$. The resulting estimator, which is generally referred to as the minimum-chi-square estimator, has the same asymptotic efficiency as the maximum-likelihood estimator (MLE) and computationally may be simpler than the MLE. Moreover, in finite samples, the minimum-chi-square estimator may even have a smaller mean squared error than the MLE (e.g., Amemiya 1974, 1976, 1980a; Berkson 1944, 1955, 1957, 1980; Ferguson 1958; Neyman 1949). However, despite its statistical attractiveness, the minimum-chi-square method is probably less useful than the maximum-likelihood method in analyzing survey data than it is in the laboratory setting. Application of the minimum-chi-square method requires repeated observations for each value of the vector of explanatory variables. In survey data, most explanatory variables are continuous. The survey sample size has to be extremely large for the possible configurations of explanatory variables. Furthermore, if the proportion of $y = 1$ is 0 or 1 for a given $\mathbf{x}$, the minimum-chi-square method for that value of $\mathbf{x}$ is not defined, but those observations can still be utilized to obtain the MLE. For this reason, we shall confine our attention to the maximum-likelihood method.[1]

When the dependent variable $y_i$ can assume more than two values, say $y_i$ takes $m_i + 1$ possible values, $0, 1, \ldots, m_i$, we can introduce $(m_i + 1)$ binary variables with

$$y_{ij} = \begin{cases} 1, & \text{if } y_i = j, \\ 0, & \text{otherwise,} \end{cases} \tag{7.2.16}$$

$$i = 1, \ldots, N, \quad j = 0, 1, \ldots, m_i.$$

[1] For a survey of the minimum-chi-square method, see Hsiao (1985b).

Let Prob $(y_i = j) =$ Prob $(y_{ij} = 1) = F_{ij}$. If the sample is randomly drawn, the likelihood function takes the form

$$L = \prod_{i=1}^{N} \prod_{j=1}^{m_i} F_{ij}^{y_{ij}}, \tag{7.2.17}$$

which is similar to the binary case (7.2.11). The complication is in the specification of $F_{ij}$. Once $F_{ij}$ is specified, general results concerning the methods of estimation and inference of the dichotomous case also apply here.

If there is a natural ordering of the outcomes, say

$$y_i = \begin{cases} 0, & \text{if the price of a home bought } < \$49,999, \\ 1, & \text{if the price of a home bought } = \$50,000\text{–}\$99,999, \\ 2, & \text{if the price of a home bought } > \$100,000. \end{cases}$$

one can use a single latent response function

$$y_i^* = \mathbf{x}_i \boldsymbol{\beta} + v_i \tag{7.2.18}$$

to characterize the ordered outcomes with

$$y_i = \begin{cases} 0 & \text{if} \quad y_i^* < c_1, \\ 1 & \text{if} \quad c_1 < y_i^* < c_2, \\ 2 & \text{if} \quad c_2 < y_i^*. \end{cases} \tag{7.2.19}$$

If the outcomes are unordered, for instance,

$$y_i = \begin{cases} 1, & \text{if mode of transport is car,} \\ 2, & \text{if mode of transport is bus,} \\ 3, & \text{if mode of transport is train,} \end{cases}$$

then we will have to use a multivariate probability distribution to characterize the outcomes. One way to postulate unordered outcomes is to assume that the $j$th alternative is chosen because it yields higher utility than the utility of other alternatives. Let the $i$th individual's utility of choosing $j$th alternative be

$$y_{ij}^* = \mathbf{x}_i' \beta_j + v_{ij},\ j = 0, 1, \ldots, m_i. \tag{7.2.20}$$

Then

$$\begin{aligned} \text{Prob}(y_i = j \mid \mathbf{x}_i) &= \text{Prob}(y_{ij}^* > y_{i\ell}^*, \quad \forall\, \ell \neq j \mid \mathbf{x}_i) \\ &= F_{ij}. \end{aligned} \tag{7.2.21}$$

The probability $F_{ij}$ is derived from the joint distribution of $(v_{io}, \ldots, v_{im})$. If $(v_{io}, \ldots, v_{im})$ follows a multivariate normal distribution, then (7.2.21) yields a multivariate probit. If the errors $v_{ij}$ are independently, identically distributed with type I extreme value distribution, (7.2.21) yields a conditional logit model (McFadden 1974). However, contrary to the univariate case, the similarity between the probit and logit specifications no longer holds. In general, they will lead to different inferences. The advantage of multivariate probit model is that it allows the choice among alternatives to have arbitrary correlation. The disadvantage is that the evaluation of Prob$(y_i = j)$ involves multiple

integrations that can be computationally infeasible. The advantage of the conditional logit model is that the evaluation of Prob($y_i = j$) does not involve multiple integration. The disadvantage is that the relative odds between two alternatives are independent of the presence or absence of the other alternatives-the so-called *independence of irrelevant alternatives*. If the errors among alternatives are not independently distributed, this can lead to grossly false predictions of the outcomes. For discussion of model specification tests, see Hausman and McFadden (1984), Hsiao (1992b), Lee (1982, 1987), Small and Hsiao (1985).

Because in many cases, a multi-response model can be transformed into a dichotomous model characterized by the $\sum_{i=1}^{N}(m_i + 1)$ binary variables as in (7.2.16),[2] for ease of exposition, we shall concentrate only on the dichotomous model.[3]

When there is no information about the probability laws of generating $v_i$, a semi-parametric approach can be used to estimate $\beta$ subject to certain normalization rule (e.g., Klein and Spady 1993; Manski 1985; Powell, Stock, and Stoker 1989). However, whether an investigator takes a parametric or semi-parametric approach, the cross-sectional model assumes that the error term $v_i$ in the latent response function (7.2.1) is independently, identically distributed and is independent of $\mathbf{x}_i$. In other words, conditional on $\mathbf{x}_i$, everyone has the same probability that an event will occur. It does not allow the possibility that the average behavior given $\mathbf{x}$ can be different from individual probabilities, that is, that it does not allow $Pr(y_i = 1 \mid \mathbf{x}) \neq Pr(y_j = 1 \mid \mathbf{x})$. The availability of panel data provides the possibility to distinguish average behavior from individual behavior by decomposing the error term, $v_{it}$, into

$$v_{it} = \alpha_i + \lambda_t + u_{it} \tag{7.2.22}$$

where $\alpha_i$ and $\lambda_t$ denote the effects of omitted individual-specific and time-specific variables, respectively. Then Prob($y_i = 1 \mid \mathbf{x}, \alpha_i$) $\neq$ Prob($y_j = 1 \mid \mathbf{x}, \alpha_j$) if $\alpha_i \neq \alpha_j$. In this chapter, we demonstrate the misspecifications that can arise because of failure to control for unobserved characteristics of the individuals in panel data and discuss possible remedies.

## 7.3 PARAMETRIC APPROACH TO STATIC MODELS WITH HETEROGENEITY

Statistical models developed for analyzing cross-sectional data essentially ignore individual differences and treat the sum of the individual-specific effect and the time-varying omitted-variable effect as a pure chance event. However, as the example in Chapter 1 shows, a discovery of a group of married women having an average yearly labor participation rate of 50 percent could lead to

[2] The variable $y_{i0}$ is sometimes omitted from the specification because it is determined by $y_{i0} = 1 - \Sigma_{j=1}^{m} y_{ij}$. For instance, a dichotomous model is often simply characterized by a single binary variable $y_i, i = 1, \ldots, N$.

[3] It should be noted that in generalizing the results of the binary case to the multiresponse case, we should allow for the fact that although $y_{ij}$ and $y_{i'j}$ are independent for $i \neq i'$, $y_{ij}$ and $y_{ij'}$ are not, because $\text{Cov}(y_{ij}, y_{ij'}) = -F_{ij}F_{ij'}$.

diametrically opposite inferences. At one extreme, each woman in a homogeneous population could have a 50 percent chance of being in the labor force in any given year, whereas at the other extreme 50 percent of women in a heterogeneous population might always work and 50 percent never work. Either explanation is consistent with the finding relying on given cross-sectional data. To discriminate among the many possible explanations, we need information on individual labor-force histories in different subintervals of life cycle. Panel data, having information on intertemporal dynamics of individual entities, provide the possibility to separate a model of individual behavior from a model of average behavior of a group of individuals.

Suppose there are sample observations $(y_{it}, \mathbf{x}_{it})$ for $i = 1, \ldots, N$ and $t = 1, \ldots, T$, where $y_{it}$ is binary with $y_{it} = 1$ if $y_{it}^*$ given by (7.2.1) is greater than 0 and 0 otherwise. For simplicity, we shall assume that the heterogeneity across cross-sectional units is time invariant,[4] and these individual-specific effects are captured by decomposing the error term $v_{it}$ in (7.2.1) as $\alpha_i + u_{it}$. When $\alpha_i$ are treated as fixed, $\text{Var}(v_{it} \mid \alpha_i) = \text{Var}(u_{it}) = \sigma_u^2$. When $\alpha_i$ are treated as random, we assume that $E\alpha_i = E\alpha_i u_{it} = 0$, and $\text{Var}(v_{it}) = \sigma_u^2 + \sigma_\alpha^2$. However, as discussed earlier, when the dependent variables are binary, the scale factor is not identifiable. Thus, for ease of exposition, we normalize the variance of $u$, $\sigma_u^2$, to be equal to 1 for the parametric specifications discussed in Section 7.2.

The existence of such unobserved permanent components allows individuals who are homogeneous in terms of their observed characteristics to be heterogeneous in response probabilities, $F(\beta'\mathbf{x}_{it} + \alpha_i)$. For example, heterogeneity will imply that the sequential-participation behavior of a woman, $F(\beta'\mathbf{x} + \alpha_i)$, within a group of women with observationally identical $\mathbf{x}$ differ systematically from $F(\beta'\mathbf{x})$ or the average behavior of the group, $\int F(\beta'\mathbf{x} + \alpha) dH(\alpha \mid \mathbf{x})$, where $H(\alpha \mid \mathbf{x})$ gives the population probability (or empirical distribution) for $\alpha$ conditional on $\mathbf{x}$.[5] In this section, we discuss the statistical inference of the common parameters $\beta$ based on a parametric specification of $F(\cdot)$.

### 7.3.1 Fixed-Effects Models

#### *7.3.1.1 Maximum-Likelihood Estimator*

If the individual specific effect, $\alpha_i$, is assumed to be fixed,[6] then both $\alpha_i$ and $\beta$ are unknown parameters to be estimated for the model $\text{Prob}(y_{it} = 1 \mid \mathbf{x}_{it}, \alpha_i) = F(\beta'\mathbf{x}_{it} + \alpha_i)$. When $T$ is finite, there is only a limited number of observations to provide information on $\alpha_i$. Thus, we have the familiar incidental-parameter problem (Neyman and Scott 1948). Any estimation of the $\alpha_i$ is meaningless

[4] For a random-coefficient formulation of probit models, see Hausman and Wise (1978).

[5] Note that, in general, $\int F(\boldsymbol{\beta}'\mathbf{x} + \alpha) dH(\alpha \mid \mathbf{x}) \neq F[\boldsymbol{\beta}'\mathbf{x} + E(\alpha \mid \mathbf{x})]$.

[6] Note that for notational ease, we now use only $\alpha_i$ instead of both $\alpha_i$ and $\alpha_i^*$. Readers should bear in mind that whenever $\alpha_i$ are treated as fixed, they are not viewed as the deviation from the common mean $\mu$; rather, they are viewed as the sum of $\mu$ and the individual deviation. On the other hand, when $\alpha_i$ are treated as random, we assume that $E\alpha_i = 0$.

if we intend to judge the estimators by their large-sample properties. We shall therefore concentrate on estimation of the common parameters, $\beta$.

Unfortunately, contrary to the linear-regression case where the individual effects $\alpha_i$ can be eliminated by taking a linear transformation such as first difference, in general, it is hard to find simple transformation to eliminate the incidental parameters from a nonlinear model. The MLEs for $\alpha_i$ and $\beta$ are not independent of each other for the discrete-choice models. When $T$ is fixed, the inconsistency of $\hat{\alpha}_i$ is transmitted into the MLE for $\beta$. Hence, even if $\beta$ is the same for all $i$ and $t$ the MLE of $\beta$ remains inconsistent if $T$ is finite no matter how large $N$ is.

We demonstrate the inconsistency of the MLE for $\beta$ by considering a logit model. The log-likelihood function for this model is

$$\log L = -\sum_{i=1}^{N}\sum_{t=1}^{T} \log\left[1 + \exp\left(\beta'\mathbf{x}_{it} + \alpha_i\right)\right] + \sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}(\beta'\mathbf{x}_{it} + \alpha_i). \tag{7.3.1}$$

For ease of illustration, we consider a special case of $T = 2$, one explanatory variable, with $x_{i1} = 0$, and $x_{i2} = 1$. Then the first-derivative equations are

$$\begin{aligned}\frac{\partial \log L}{\partial \beta} &= \sum_{i=1}^{N}\sum_{t=1}^{2}\left[-\frac{e^{\beta\mathbf{x}_{it}+\alpha_i}}{1+e^{\beta\mathbf{x}_{it}+\alpha_i}} + y_{it}\right]\mathbf{x}_{it} \\ &= \sum_{i=1}^{N}\left[-\frac{e^{\beta+\alpha_i}}{1+e^{\beta+\alpha_i}} + y_{i2}\right] = 0,\end{aligned} \tag{7.3.2}$$

$$\frac{\partial \log L}{\partial \alpha_i} = \sum_{t=1}^{2}\left[-\frac{e^{\beta\mathbf{x}_{it}+\alpha_i}}{1+e^{\beta\mathbf{x}_{it}+\alpha_i}} + y_{it}\right] = 0 \tag{7.3.3}$$

Solving (7.3.3), we have

$$\begin{aligned}&\hat{\alpha}_i = \infty \text{ if } y_{i1} + y_{i2} = 2, \\ &\hat{\alpha}_i = -\infty \text{ if } y_{i1} + y_{i2} = 0, \\ &\hat{\alpha}_i = -\tfrac{\beta}{2} \text{ if } y_{i1} + y_{i2} = 1.\end{aligned} \tag{7.3.4}$$

Inserting (7.3.4) into (7.3.2) and letting $n_1$ denote the number of individuals with $y_{i1} + y_{i2} = 1$ and letting $n_2$ denote the number of individuals with $y_{i1} + y_{i2} = 2$, we have[7]

$$\sum_{i=1}^{N}\frac{e^{\beta+\alpha_i}}{1+e^{\beta+\alpha_i}} = n_1\frac{e^{\beta/2}}{1+e^{\beta/2}} + n_2 = \sum_{i=1}^{N} y_{i2}. \tag{7.3.5}$$

Therefore,

$$\hat{\beta} = 2\left\{\log\left(\sum_{i=1}^{N} y_{i2} - n_2\right) - \log\left(n_1 + n_2 - \sum_{i=1}^{N} y_{i2}\right)\right\}. \tag{7.3.6}$$

[7] The number of individuals with $y_{i1} + y_{i2} = 0$ is $N - n_1 + n_2$.

By a law of large numbers (Rao 1973, Chapter 2),

$$\begin{aligned}\text{plim}_{N\to\infty}\ & \frac{1}{N}\left(\sum_{i=1}^{N} y_{i2} - n_2\right) \\ &= \frac{1}{N}\sum_{i=1}^{N} \text{Prob}(y_{i1}=0, y_{i2}=1 \mid \beta, \alpha_i) \\ &= \frac{1}{N}\sum_{i=1}^{N} \frac{e^{\beta+\alpha_i}}{(1+e^{\alpha_i})(1+e^{\beta+\alpha_i})},\end{aligned} \tag{7.3.7}$$

$$\begin{aligned}\text{plim}_{N\to\infty}\ & \frac{1}{N}\left(n_1 + n_2 - \sum_{i=1}^{N} y_{i2}\right) \\ &= \frac{1}{N}\sum_{i=1}^{N} \text{Prob}(y_{i1}=1, y_{i2}=0 \mid \beta, \alpha_i) \\ &= \frac{1}{N}\sum_{i=1}^{N} \frac{e^{\alpha_i}}{(1+e^{\alpha_i})(1+e^{\beta+\alpha_i})}.\end{aligned} \tag{7.3.8}$$

Substituting $\hat{\alpha}_i = \frac{\beta}{2}$ into (7.3.7) and (7.3.8) yields

$$\text{plim}_{N\to\infty}\hat{\beta} = 2\beta, \tag{7.3.9}$$

which is not consistent.

### *7.3.1.2 Conditions for the Existence of a Consistent Estimator*

Neyman and Scott (1948) have suggested a general principle to find a consistent estimator for the (structural) parameter $\beta$ in the presence of the incidental parameters $\alpha_i$.[8] Suppose the dimension of $\beta$ is $K$, their idea is to find $K$ functions

$$\Psi_{Nj}(\mathbf{y}_1, \ldots, \mathbf{y}_N \mid \beta),\ j = 1, \ldots, K. \tag{7.3.10}$$

that are independent of the incidental parameters $\alpha_i$ and have the property that when $\beta$ are the true values $\Psi_{Nj}(\mathbf{y}_1, \ldots, \mathbf{y}_N \mid \beta)$ converges to some known constant, say 0, in probability as $N$ tends to infinity. Then an estimator $\hat{\beta}$ derived by solving $\Psi_{Nj}\left(\mathbf{y}_1, \ldots, \mathbf{y}_N \mid \hat{\beta}\right) = 0$ is consistent under suitable regularity conditions. For instance, $\hat{\beta}^* = (1/2)\hat{\beta}$ for the foregoing example of a fixed-effect logit model (7.3.1)–(7.3.3) is such an estimator.

In the case of a linear-probability model, either taking first difference over time or taking difference with respect to the individual mean eliminates the

[8] We call $\boldsymbol{\beta}$ the structural parameter because the value of $\boldsymbol{\beta}$ characterizes the structure of the complete sequence of random variables. It is the same for all $i$ and $t$. We call $\alpha_i$ an incidental parameter to emphasize that the value of $\alpha_i$ changes when $i$ changes.

individual-specific effect. The least-squares regression of the differenced equations yields a consistent estimator for $\beta$ when $N$ tends to infinity.

But in the general nonlinear models, simple forms of $\Psi(\cdot)$ are not always easy to find. For instance, in general, we do not know the probability limit of the MLE of a fixed-effects logit model. However, if a minimum sufficient statistic $\tau_i$ for the incidental parameter $\alpha_i$ exists and is not dependent on the structural parameter $\beta$, the conditional density,

$$f^*(\mathbf{y}_i \mid \beta, \tau_i) = \frac{f(\mathbf{y}_i \mid \beta, \alpha_i)}{g(\tau_i \mid \beta, \alpha_i)} \text{ for } g(\tau_i \mid \beta, \alpha_i) > 0, \tag{7.3.11}$$

no longer depends on $\alpha_i$.[9] Andersen (1970, 1973) has shown that maximizing the conditional density of $\mathbf{y}_1, \ldots, \mathbf{y}_N$, given $\tau_1, \ldots, \tau_N$,

$$\prod_{i=1}^{N} f^*(\mathbf{y}_i \mid \beta, \tau_i), \tag{7.3.12}$$

yields the first-order conditions $\Psi_{Nj}(\mathbf{y}_1, \ldots, \mathbf{y}_N \mid \hat{\beta}, \tau_1, \tau_2, \ldots, \tau_N) = 0$, for $j = 1, \ldots, K$. Solving these functions will give a consistent estimator of the common (structural) parameter $\beta$ under mild regularity conditions.[10]

To illustrate the conditional maximum-likelihood method, we use the logit model as an example. The joint probability of $\mathbf{y}_i$ is

$$\text{Prob } (\mathbf{y}_i) = \frac{\exp \{\alpha_i \sum_{t=1}^{T} y_{it} + \beta' \sum_{t=1}^{T} \mathbf{x}_{it} y_{it}\}}{\prod_{t=1}^{T} [1 + \exp (\beta' \mathbf{x}_{it} + \alpha_i)]}. \tag{7.3.13}$$

The logit form has the property that the denominator of Prob($y_{it}$) is always $[1 + \exp (\boldsymbol{\beta}' \mathbf{x}_{it} + \alpha_i)]$ independent of whether $y_{it} = 1$ or 0. On the other hand, for any sequence of dummy variable $d_{ijt}$, $D_{ij} = (d_{ij1}, d_{ij2}, \ldots, d_{ijT})$ where $d_{ijt} = 0$ or 1, the numerator of Prob($D_{ij}$) always has the form $\exp(\alpha_i \sum_{t=1}^{T} d_{ijt}) \cdot \exp [\boldsymbol{\beta}' \sum_{t=1}^{T} \mathbf{x}_{it} d_{ijt}]$. It is clear that $\sum_{t=1}^{T} y_{it}$ is a minimum sufficient statistic for $\alpha_i$. The conditional probability for $y_{it}$ given $\sum_{t=1}^{T} y_{it}$ is

$$\text{Prob } \left(\mathbf{y}_i \mid \sum_{t=1}^{T} y_{it}\right) = \frac{\exp \left[\beta' \sum_{t=1}^{T} \mathbf{x}_{it} y_{it}\right]}{\sum_{D_{ij} \in \bar{B}_i} \exp \{\beta' \sum_{t=1}^{T} \mathbf{x}_{it} d_{ijt}\}}, \tag{7.3.14}$$

where $\bar{B}_i = \{D_{ij} = (d_{ij1}, \ldots, d_{ijT}) \mid d_{ijt} = 0 \text{ or } 1, \text{ and } \sum_{t=1}^{T} d_{ijt} = \sum_{t=1}^{T} y_{it}, j = 1, 2, \ldots, \frac{T!}{s!(T-s)!}\}$, is the set of all possible distinct sequence

[9] Suppose that the observed random variables $\mathbf{y}$ have a certain joint distribution function that belongs to a specific family $\mathcal{J}$ of distribution functions. The statistic $S(\mathbf{y})$ (a function of the observed sample values $\mathbf{y}$) is called a sufficient statistic if the conditional expectation of any other statistic $H(\mathbf{y})$, given $S(\mathbf{y})$, is independent of $\mathcal{J}$. A statistic $S^*(\mathbf{y})$ is called a minimum sufficient statistic if it is a function of every sufficient statistic $S(\mathbf{y})$ for $\mathcal{J}$. For additional discussion, see Zacks (1971, Chapter 2).

[10] When $u_{it}$ are independently normally distributed, the LSDV estimator of $\boldsymbol{\beta}$ for the linear static model is the conditional MLE (Cornwell and Schmidt 1984).

$(d_{ij1}, d_{ij2}, \ldots, d_{ijT})$ satisfying $\sum_{t=1}^{T} d_{ijt} = \sum_{t=1}^{T} y_{it} = s$. There are $T+1$ distinct alternative sets corresponding to $\sum_{t=1}^{T} y_{it} = 0, 1, \ldots, T$. Groups for which $\sum_{t=1}^{T} y_{it} = 0$ or $T$ contribute 0 to the likelihood function, because the corresponding probability in this case is equal to 1 (with $\alpha_i = -\infty$ or $\infty$). So only $T-1$ alternative sets are relevant. The alternative sets for groups with $\sum_{t=1}^{T} y_{it} = s$ have $\binom{T}{S}$ elements, corresponding to the distinct sequences of $T$ trials with $s$ success.

Equation (7.3.14) is in a conditional logit form (McFadden 1974), with the alternative sets $(\bar{B}_i)$ varying across observations $i$. It does not depend on the incidental parameters, $\alpha_i$. Therefore, the conditional maximum-likelihood estimator of $\beta$ is consistent under mild conditions. For example, with $T = 2$, the only case of interest is $y_{i1} + y_{i2} = 1$. The two possibilities are $\omega_i = 1$, if $(y_{i1}, y_{i2}) = (0, 1)$, and $\omega_i = 0$, if $(y_{i1}, y_{i2}) = (1, 0)$.

The conditional probability of $\omega_i = 1$ given $y_{i1} + y_{i2} = 1$ is

$$\begin{aligned} \text{Prob}\,(\omega_i = 1 \mid y_{i1} + y_{i2} = 1) &= \frac{\text{Prob}\,(\omega_i = 1)}{\text{Prob}\,(\omega_i = 1) + \text{Prob}\,(\omega_i = 0)} \\ &= \frac{\exp\,[\beta'(\mathbf{x}_{i2} - \mathbf{x}_{i1})]}{1 + \exp[\beta'(\mathbf{x}_{i2} - \mathbf{x}_{i1})]} \\ &= F[\beta'(\mathbf{x}_{i2} - \mathbf{x}_{i1})]. \end{aligned} \tag{7.3.15}$$

Equation (7.3.15) is in the form of a binary logit function in which the two outcomes are (0,1) and (1,0), with explanatory variables $(\mathbf{x}_{i2} - \mathbf{x}_{i1})$. The conditional log-likelihood function is

$$\begin{aligned} \log L^* = \sum_{i \in \tilde{B}_1} \{&\omega_i \log F[\beta'(\mathbf{x}_{i2} - \mathbf{x}_{i1})] \\ &+ (1 - \omega_i) \log (1 - F[\beta'(\mathbf{x}_{i2} - \mathbf{x}_{i1})])\}, \end{aligned} \tag{7.3.16}$$

where $\tilde{B}_1 = \{i \mid y_{i1} + y_{i2} = 1\}$.

Although $\tilde{B}_1$ is a random set of indices, Chamberlain (1980) has shown that the inverse of the information matrix based on the conditional-likelihood function provides an asymptotic covariance matrix for the conditional MLE of $\beta$ when $N$ tends to infinity. This can be made more explicit by defining $d_i = 1$, if $y_{i1} + y_{i2} = 1$, and $d_i = 0$, otherwise, for the foregoing case in which $T = 2$. Then we have

$$\begin{aligned} J_{\tilde{B}_1} = \frac{\partial^2 \log L^*}{\partial\beta\partial\beta'} = -\sum_{i=1}^{N} d_i F[\beta'(\mathbf{x}_{i2} - \mathbf{x}_{i1})] \\ \{1 - F[\beta'(\mathbf{x}_{i2} - \mathbf{x}_{i1})]\}(\mathbf{x}_{i2} - \mathbf{x}_{i1}) \cdot (\mathbf{x}_{i2} - \mathbf{x}_{i1})'. \end{aligned} \tag{7.3.17}$$

The information matrix is

$$J = E(J_{\tilde{B}_1}) = -\sum_{i=1}^{N} P_i F[\beta'(\mathbf{x}_{i2} - \mathbf{x}_{i1})] \tag{7.3.18}$$
$$\{1 - F[\beta'(\mathbf{x}_{i2} - \mathbf{x}_{i1})]\}(\mathbf{x}_{i2} - \mathbf{x}_{i1}) \cdot (\mathbf{x}_{i2} - \mathbf{x}_{i1})',$$

where $P_i = E(d_i \mid \alpha_i) = F(\beta'\mathbf{x}_{i1} + \alpha_i)[1 - F(\beta'\mathbf{x}_{i2} + \alpha_i)] + [1 - F(\beta'\mathbf{x}_{i1} + \alpha_i)]\ F(\beta'\mathbf{x}_{i2} + \alpha_i)$. Because $d_i$ are independent, with $Ed_i = P_i$, and both $F$ and the variance of $d_i$ are uniformly bounded, by a strong law of large numbers,

$$\frac{1}{N} J_{\tilde{B}_1} - \frac{1}{N} J \text{ almost surely } \to 0 \text{ as } N \to \infty \tag{7.3.19}$$
$$\text{if } \sum_{i=1}^{N} \frac{1}{i^2} \mathbf{m}_i \mathbf{m}_i' < \infty,$$

where $\mathbf{m}_i$ replaces each element of $(\mathbf{x}_{i2} - \mathbf{x}_{i1})$ by its square. The condition for convergence clearly holds if $\mathbf{x}_{it}$ is uniformly bounded.

For the case of $T > 2$, there is no loss of generality in choosing the sequence $D_{i1} = (d_{i11}, \ldots, d_{i1T}, \sum_{t=1}^{T} d_{i1t} = \sum_{t=1}^{T} y_{it} = s, 1 \le s \le T - 1)$, as the normalizing factor. Hence we may rewrite the conditional probability (7.3.14) as

$$\text{Prob}\left(\mathbf{y}_i \mid \sum_{t=1}^{T} y_{it}\right) = \frac{\exp\{\beta' \sum_{t=1}^{T} \mathbf{x}_{it}(y_{it} - d_{i1t})\}}{1 + \sum_{D_{ij} \in (\tilde{B}_i - D_{i1})} \exp\{\beta' \sum_{t=1}^{T} \mathbf{x}_{it}(d_{ijt} - d_{i1t})\}} \tag{7.3.20}$$

Then the conditional log-likelihood function takes the form

$$\log L^* = \sum_{i \in C} \left\{ \beta' \sum_{t=1}^{T} \mathbf{x}_{it}(y_{it} - d_{i1t}) - \log \left[ 1 + \sum_{D_{ij} \in (\tilde{B}_i - D_{i1})} \exp \left\{ \beta' \sum_{t=1}^{T} \mathbf{x}_{it}(d_{ijt} - d_{i1t}) \right\} \right] \right\} \tag{7.3.21}$$

where $C = \{i \mid \sum_{t=1}^{T} y_{it} \neq T, \sum_{t=1}^{T} y_{it} \neq 0\}$.

Although we can find simple transformations of linear-probability and logit models that will satisfy the Neyman–Scott principle, we cannot find simple functions for the parameters of interest that are independent of the nuisance parameters $\alpha_i$ for probit models. That is, there does not appear to exist a consistent estimator of $\beta$ for the fixed-effects probit models.

*7.3.1.3 Some Monte Carlo Evidence*

Given that there exists a consistent estimator of $\beta$ for the fixed-effects logit model, but not for the fixed-effects probit model, and that in the binary case probit and logit models yield similar results, it appears that a case can be made for favoring the logit specification because of the existence of a consistent estimator for the structural parameter $\beta$. However, in the multivariate case, logit and probit models yield very different results. In this situation it will be useful to know the magnitude of the bias if the data actually call for a fixed-effects probit specification.

Heckman (1981b) conducted a limited set of Monte Carlo experiments to get some idea of the order of bias of the MLE for the fixed-effects probit models. His data were generated by the model

$$y_{it}^* = \beta x_{it} + \alpha_i + u_{it}, \quad i = 1, 2, \ldots, N, t = 1, \ldots, T, \tag{7.3.22}$$

and

$$y_{it} = \begin{cases} 1, & \text{if } y_{it}^* > 0, \\ 0, & \text{otherwise.} \end{cases}$$

The exogenous variable $x_{it}$ was generated by a Nerlove (1971a) process,

$$x_{it} = 0.1t + 0.5x_{i,t-1} + \epsilon_{it}, \tag{7.3.23}$$

where $\epsilon_{it}$ is a uniform random variable having mean 0 and range $-1/2$ to $1/2$. The variance $\sigma_u^2$ was set at 1. The scale of the variation of the fixed effect, $\sigma_\alpha^2$, is changed for different experiments. In each experiment, 25 samples of 100 individuals ($N = 100$) were selected for eight periods ($T = 8$).

The results of Heckman's experiment with the fixed-effects MLE of probit models are presented in Table 7.1. For $\beta = -0.1$, the fixed-effects estimator does well. The estimated value comes very close to the true value. For $\beta = -1$ or $\beta = 1$, the estimator does not perform as well, but the bias is never more than 10 percent and is always toward 0. Also, as the scale of the variation in the fixed-effects decreases, so does the bias.[11]

### 7.3.2 Random-Effects Models

When the individual specific effects $\alpha_i$ are treated as random, we may still use the fixed effects estimators to estimate the structural parameters $\beta$. The asymptotic properties of the fixed effects estimators of $\beta$ remain unchanged. However, if $\alpha_i$ are random, but are treated as fixed, the consequence, at its best, is a loss of efficiency in estimating $\beta$, but it could be worse, namely,

[11] Similar results also hold for the MLE of the fixed-effects logit model. Wright and Douglas (1976), who used Monte Carlo methods to investigate the performance of the MLE, found that when $T = 20$, the MLE is virtually unbiased, and its distribution is well described by a limiting normal distribution, with the variance–covariance matrix based on the inverse of the estimated-information matrix.

Table 7.1. *Average values of $\hat{\beta}$ for the fixed-effects probit model*

| | $\hat{\beta}$ | | |
|---|---|---|---|
| $\sigma_\alpha^2$ | $\beta = 1$ | $\beta = -0.1$ | $\beta = -1$ |
| 3 | 0.90 | −0.10 | −0.94 |
| 1 | 0.91 | −0.09 | −0.95 |
| 0.5 | 0.93 | −0.10 | −0.96 |

*Source:* Heckman (1981b, Table 4.1).

the resulting fixed effects estimators may be inconsistent as discussed in Section 7.3.1.

When $\alpha_i$ are independent of $\mathbf{x}_i$ and are a random sampling from a univariate distribution $G$, indexed by a finite number of parameters $\delta$, the log-likelihood function becomes

$$\log L = \sum_{i=1}^{N} \log \int \prod_{t=1}^{T} F(\beta' \mathbf{x}_{it} + \alpha)^{y_{it}} [1 - F(\beta' \mathbf{x}_{it} + \alpha)]^{1-y_{it}} dG(\alpha \mid \delta). \tag{7.3.24}$$

where $F(\cdot)$ is the distribution of the error term conditional on both $\mathbf{x}_i$ and $\alpha_i$. Equation (7.3.24) replaces the probability function for $y$ conditional on $\alpha$ by a probability function that is marginal on $\alpha$. It is a function of a finite number of parameters $(\beta', \delta')$. Thus, maximizing (7.3.24), under weak regularity conditions, will give consistent estimators for $\beta$ and $\delta$ as $N$ tends to infinity provided the distribution (or conditional distribution) of $\alpha$ is correctly specified. If $G(\alpha)$ is misspecified, maximizing (7.3.24) will yield inconsistent estimates when $T$ is fixed. However, when $T \to \infty$, the random effects estimator becomes consistent, irrespective of the form of the postulated distribution of individual effects. The reason is that:

$$\log f(\mathbf{y}_i \mid \mathbf{x}_i, \beta, \alpha_i) = \sum_{t=1}^{T} \log f(y_{it} \mid \mathbf{x}_{it}, \beta, \alpha_i)$$

is a sum of $T$ time series observation, so that the distribution of $\alpha$ becomes negligible compared to that of the likelihood as the number of time periods increases (Arellano and Bonhomme 2009).

If $\alpha_i$ is correlated with $\mathbf{x}_{it}$, maximizing (7.3.24) will not eliminate the omitted-variable bias. To allow for dependence between $\alpha$ and $\mathbf{x}$, we must specify a distribution for $\alpha$ conditional on $\mathbf{x}$, $G(\alpha \mid \mathbf{x})$ and consider the marginal

log-likelihood function

$$\log L = \sum_{i=1}^{N} \log \int \prod_{t=1}^{T} F(\beta' x_{it} + \alpha)^{y_{it}} [1 - F(\beta' x_{it} + \alpha)]^{1-y_{it}} dG(\alpha \mid \mathbf{x}) \tag{7.3.24'}$$

A convenient specification suggested by Chamberlain (1980, 1984) is to assume that $\alpha_i = \sum_{t=1}^{T} \mathbf{a}_t' \mathbf{x}_{it} + \eta_i = \mathbf{a}' \mathbf{x}_i + \eta_i$ where $\mathbf{a}' = (\mathbf{a}_1', \ldots, \mathbf{a}_T)$ and $\mathbf{x}_i' = (\mathbf{x}_{i1}', \ldots, \mathbf{x}_{iT}')$, and $\eta_i$ is the residual. However, there is a very important difference in this step compared with the linear case. In the linear case it was not restrictive to decompose $\alpha_i$ into its linear projection on $\mathbf{x}_i$ and an orthogonal residual. Now we are assuming that the regression function $E(\alpha_i \mid \mathbf{x}_i)$ is actually linear, that $\eta_i$ is independent of $\mathbf{x}_i$, and that $\eta_i$ has a specific probability distribution.

Given these assumptions, the log-likelihood function under our random-effects specification is

$$\begin{aligned} \log L = \sum_{i=1}^{N} \log \int \prod_{t=1}^{T} & F(\beta' \mathbf{x}_{it} + \mathbf{a}' \mathbf{x}_i + \eta)^{y_{it}} \\ & \cdot [1 - F(\beta' \mathbf{x}_{it} + \mathbf{a}' \mathbf{x}_i + \eta)]^{1-y_{it}} dG^*(\eta), \end{aligned} \tag{7.3.25}$$

where $G^*$ is a univariate distribution function for $\eta$. For example, if $F$ is a standard normal distribution function and we choose $G^*$ to be the distribution function of a normal random variable with mean 0 and variance $\sigma_\eta^2$, then our specification gives a multivariate probit model:

$$\begin{aligned} y_{it} &= 1 \text{ if } \beta' \mathbf{x}_{it} + \mathbf{a}' \mathbf{x}_i + \eta_i + u_{it} > 0, \quad t = 1, \ldots, T, \\ &= 0 \text{ if } \beta' \mathbf{x}_{it} + \mathbf{a}' \mathbf{x}_i + \eta_i + u_{it} \leq 0, \end{aligned} \tag{7.3.26}$$

where $\mathbf{u}_i + \mathbf{e}\eta_i$ is independent normal, with mean $\mathbf{0}$ and variance–covariance matrix $I_T + \sigma_\eta^2 \mathbf{e}\mathbf{e}'$.

The difference between (7.3.25) and (7.3.24) is only in the inclusion of the term $\mathbf{a}' \mathbf{x}_i$ to capture the dependence between the incidental parameters $\alpha_i$ and $\mathbf{x}_i$. Therefore, the essential characteristics with regard to estimation of (7.3.24) or (7.3.25) are the same. So we shall discuss only the procedure to estimate the model (7.3.24).

Maximizing (7.3.25) involves integration of $T$ dimensions, which can be computationally cumbersome. An alternative approach that simplifies the computation of the MLE to a univariate integration is to note that conditional on $\alpha_i$, the error terms, $v_{it} = \alpha_i + u_{it}$ are independently normally distributed with

mean $\alpha_i$ and variance 1, denoted by $\phi(v_{it} \mid \alpha_i)$ (Heckman 1981a). Then

$$Pr(y_{i1}, \ldots, y_{iT}) = \int_{c_{i1}}^{b_{i1}} \cdots \int_{c_{iT}}^{b_{iT}} \prod_{t=1}^{T} \phi(v_{it} \mid \alpha_i) G(\alpha_i \mid x_i) d\alpha_i dv_{i1}, \ldots, dv_{iT},$$

$$= \int_{-\infty}^{\infty} G(\alpha_i \mid \mathbf{x}_i) \prod_{t=1}^{T} [\Phi(b_{it} \mid \alpha_i) - \Phi(c_{it} \mid \alpha_i)] d\alpha_i, \tag{7.3.27}$$

where $\Phi(\cdot \mid \alpha_i)$ is the cumulative distribution function (cdf) of normal density with mean $\alpha_i$ and variance 1, $\phi(\cdot \mid \alpha_i)$, $c_{it} = -\beta' \mathbf{x}_{it}$, $b_{it} = \infty$ if $y_{it} = 1$ and $c_{it} = -\infty$, $b_{it} = -\beta' \mathbf{x}_{it}$ if $y_{it} = 0$, $G(\alpha_i \mid \mathbf{x}_i)$ is the probability density function of $\alpha_i$ given $\mathbf{x}_i$. If $G(\alpha_i \mid \mathbf{x}_i)$ is assumed to be normally distributed with variance $\sigma_\alpha^2$, and the expression (7.3.27) reduces a $T$-dimensional integration to a single integral whose integrand is a product of one normal density and $T$ differences of normal cumulative density functions for which highly accurate approximations are available. For instance, Butler and Moffitt (1982) suggests using Gaussian quadrature to achieve gains in computational efficiency. The Gaussian quadrature formula for evaluation of the necessary integral is the Hermite integration formula $\int_{-\infty}^{\infty} e^{-z^2} g(z) dz \approx \sum_{j=1}^{l} w_j g(z_j)$, where $l$ is the number of evaluation points, $w_j$ is the weight given to the $j$th point, and $g(z_j)$ is $g(z)$ evaluated at the $j$th point of $z$. The points and weights are available from Abramowitz and Stegun (1965) and Stroud and Secrest (1966).

A key question for computational feasibility of the Hermite formula is the number of points at which the integrand must be evaluated for accurate approximation. Several evaluations of the integral using four periods of arbitrary values of the data and coefficients on right-hand-side variables by Butler and Moffitt (1982) show that even two-point integration is highly accurate. Of course, in the context of a maximization algorithm, accuracy could be increased by raising the number of evaluation points as the likelihood function approaches its optimum.

Although maximizing (7.3.25) or (7.3.24) provides a consistent and efficient estimator for $\beta$, computationally it is much more involved. However, if both $u_{it}$ and $\eta_i$ (or $\alpha_i$) are normally distributed, a computationally simple approach that avoids numerical integration is to make use of the fact that the distribution for $y_{it}$ conditional on $\mathbf{x}_i$ but marginal on $\alpha_i$ also has a probit form:

$$\text{Prob}\,(y_{it} = 1) = \Phi\left[(1 + \sigma_\eta^2)^{-1/2} (\beta' \mathbf{x}_{it} + \mathbf{a}' \mathbf{x}_i)\right]. \tag{7.3.28}$$

Estimating each of $t$ cross-sectional univariate probit specifications by maximum-likelihood gives the estimated coefficients of $\mathbf{x}_{it}$ and $\mathbf{x}_i$ as $\hat{\boldsymbol{\pi}}_t$, $t = 1, 2, \ldots, T$, which will converge to[12]

$$\Pi = (1 + \sigma_\eta^2)^{-1/2} (I_T \otimes \beta' + \mathbf{e}\mathbf{a}') \tag{7.3.29}$$

[12] In the case in which $\alpha_i$ are uncorrelated with $\mathbf{x}_i$, we have $\mathbf{a} = \mathbf{0}$ and $\sigma_\eta^2 = \sigma_\alpha^2$.

as $N$ tends to infinity where $\Pi$ denotes the $T \times (T+1)K$ stacked $\boldsymbol{\pi}_t'$. Therefore, consistent estimators $(1+\sigma_\eta^2)^{-1/2}\beta'$ and $(1+\sigma_\eta^2)^{-1/2}\ \mathbf{a}'$ can be easily derived from (7.3.29). One can then follow Heckman's (1981a) suggestion by substituting these estimated values into (7.3.25) and optimizing the functions with respect to $\sigma_\eta^2$ conditional on $(1+\sigma_\eta^2)^{-1/2}\beta$ and $(1+\sigma_\eta^2)^{-1/2}\mathbf{a}$.

A more efficient estimator that avoids numerical integration is to impose the restriction (7.3.29) by $\boldsymbol{\pi} = \text{vec}\,(\Pi') = \mathbf{f}(\theta)$, where $\theta' = (\beta', \mathbf{a}', \sigma_\eta^2)$, and use a genralized method of moments (GMM) or minimum-distance estimator (see Chapter 3, Section 3.9), just as in the linear case. Chamberlain (1984) suggests that we choose $\hat{\theta}$ to minimize[13]

$$(\hat{\boldsymbol{\pi}} - \mathbf{f}(\theta))'\hat{\Omega}^{-1}(\hat{\boldsymbol{\pi}} - \mathbf{f}(\theta)) \tag{7.3.30}$$

where $\hat{\Omega}$ is a consistent estimator of

$$\Omega = J^{-1}\Delta J^{-1}, \tag{7.3.31}$$

where

$$J = \begin{bmatrix} J_1 & 0 & \dots & 0 \\ 0 & J_2 & & \\ \vdots & & \ddots & \\ 0 & & & J_T \end{bmatrix},$$

$$J_t = E\left\{\frac{\phi_{it}^2}{\Phi_{it}(1-\Phi_{it})}\mathbf{x}_i\mathbf{x}_i'\right\},$$

$$\Delta = E[\Phi_i \otimes \mathbf{x}_i\mathbf{x}_i'],$$

and where the $t, s$ element of the $T \times T$ matrix $\Phi_i$ is $\psi_{its} = c_{it}c_{is}$, with

$$c_{it} = \frac{y_{it} - \Phi_{it}}{\Phi_{it}(1-\Phi_{it})}\phi_{it}, \quad t = 1, \dots, T.$$

The standard normal distribution function $\Phi_{it}$ and the standard normal density function $\phi_{it}$ are evaluated at $\boldsymbol{\pi}'\mathbf{x}_i$. We can obtain a consistent estimator of $\Omega$ by replacing expectations by sample means and using $\hat{\boldsymbol{\pi}}$ in place of $\boldsymbol{\pi}$.

## 7.4 SEMIPARAMETRIC APPROACH TO STATIC MODELS

The parametric approach of estimating discrete choice model suffers from two drawbacks: (1) conditional on $\mathbf{x}$, the probability law of generating $(u_{it}, \alpha_i)$

[13] $\Omega$ is the asymptotic variance–covariance matrix of $\hat{\boldsymbol{\pi}}$ when no restrictions are imposed on the variance–covariance matrix of the $T \times 1$ normal random variable $\mathbf{u}_i + \mathbf{e}\eta_i$. We can relax the serial-independence assumption on $u_{it}$ and allow $E\mathbf{u}_i\mathbf{u}_i'$ to be an arbitrary positive definite matrix except for scale normalization. In this circumstance, $\Pi = \text{diag}\,\{(\sigma_{u1}^2 + \sigma_\eta^2)^{-1/2}, \dots, (\sigma_{uT}^2 + \sigma_\eta^2)^{-1/2}\}[I_T \otimes \beta' + \mathbf{e}\mathbf{a}']$.

is known a priori or conditional on $\mathbf{x}$ and $\alpha_i$, the probability law of $u_{it}$ is known a priori. (2) When $\alpha_i$ are fixed it appears that apart from logit and linear probability model, there does not exist a simple transformation that can get rid of the incidental parameters. The semiparametric approach not only avoids making specific distribution of $u_{it}$ but also allows consistent estimator of $\beta$ up to a scale whether $\alpha_i$ is treated as fixed or random.

### 7.4.1 Maximum Score Estimator

Manski (1975, 1985, 1987) suggests a maximum score estimator that maximizes the sample average function

$$H_N(\mathbf{b}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=2}^{T} sgn(\Delta\mathbf{x}_{it}'\mathbf{b})\Delta y_{it} \tag{7.4.1}$$

subject to the normalization condition $\mathbf{b}'\mathbf{b}$=1, where $\Delta\mathbf{x}_{it} = \mathbf{x}_{it} - \mathbf{x}_{i,t-1}$, $\Delta y_{it} = y_{it} - y_{i,t-1}$, $sgn(w) = 1$ if $w > 0$, 0 if $w = 0$, and $-1$ if $w < 0$. This is because under fairly general conditions (7.4.1) converges uniformly to

$$H(\mathbf{b}) = E[sgn(\Delta\mathbf{x}_{it}'\mathbf{b})\Delta y_{it}], \tag{7.4.2}$$

where $H(\mathbf{b})$ is maximized at $\mathbf{b} = \beta^*$, where $\beta^* = \frac{\beta}{\|\beta\|}$ and $\| \beta \|$ is the square root of the Euclidean norm $\sum_{k=1}^{K}\beta_k^2$.

To see this, we note that the binary choice model can be written in the form

$$y_{it} = \begin{cases} 1 & \text{if} \quad y_{it}^* > 0, \\ 0 & \text{if} \quad y_{it}^* \le 0, \end{cases} \tag{7.4.3}$$

where $y_{it}^*$ is given by (7.2.1) with $v_{it} = \alpha_i + u_{it}$. Under the assumption that $u_{it}$ is independently, identically distributed and is independent of $\mathbf{x}_i$ and $\alpha_i$ for given $i$ (i.e., $\mathbf{x}_{it}$ is strictly exogenous), we have

$$\begin{aligned} &\mathbf{x}_{it}'\beta > \mathbf{x}_{i,t-1}'\beta \Longleftrightarrow E(y_{it} \mid \mathbf{x}_{it}) > E(y_{i,t-1} \mid \mathbf{x}_{i,t-1}), \\ &\mathbf{x}_{it}'\beta = \mathbf{x}_{i,t-1}'\beta \Longleftrightarrow E(y_{it} \mid \mathbf{x}_{it}) = E(y_{i,t-1} \mid \mathbf{x}_{i,t-1}), \\ &\mathbf{x}_{it}'\beta < \mathbf{x}_{i,t-1}'\beta \Longleftrightarrow E(y_{it} \mid \mathbf{x}_{it}) < E(y_{i,t-1} \mid \mathbf{x}_{i,t-1}). \end{aligned} \tag{7.4.4}$$

Rewrite (7.4.4) in terms of first differences, we have the equivalent representation

$$\begin{aligned} &\Delta\mathbf{x}_{it}'\beta > 0 \Longleftrightarrow E[(y_{it} - y_{i,t-1}) > 0 \mid \Delta\mathbf{x}_{it}] \\ &\Delta\mathbf{x}_{it}'\beta = 0 \Longleftrightarrow E[(y_{it} - y_{i,t-1}) = 0 \mid \Delta\mathbf{x}_{it}], \\ &\Delta\mathbf{x}_{it}'\beta < 0 \Longleftrightarrow E[(y_{it} - y_{i,t-1}) < 0 \mid \Delta\mathbf{x}_{it}]. \end{aligned} \tag{7.4.5}$$

It is obvious that (7.4.5) continues to hold when $\tilde{\beta} = \beta c$ where $c > 0$. Therefore, we shall only consider the normalized vector $\beta^* = \frac{\beta}{\|\beta\|}$.

Then, for any $\mathbf{b}$ (satisfying $\mathbf{b}'\mathbf{b} = 1$) such that $\mathbf{b} \neq \beta^*$,

$$H(\beta^*) - H(\mathbf{b}) = E\{[sgn(\Delta\mathbf{x}_{it}'\beta^*) - sgn(\Delta\mathbf{x}_{it}'\mathbf{b})](y_{it} - y_{i,t-1})\} \tag{7.4.6}$$

$$= 2\int_{W_b} sgn(\Delta\mathbf{x}_{it}'\beta^*)E[y_t - y_{t-1} \mid \Delta\mathbf{x}]dF_{\Delta\mathbf{x}},$$

where $W_b = [\Delta\mathbf{x} : sgn(\Delta\mathbf{x}'\beta^*) \neq sgn(\Delta\mathbf{x}'\mathbf{b})]$, and $F_{\Delta\mathbf{x}}$ denotes the distribution of $\Delta\mathbf{x}$. Because of (7.4.5), the relation (7.4.6) implies that for all $\Delta\mathbf{x}$,

$$sgn\,(\Delta\mathbf{x}'\beta^*)E[y_t - y_{t-1} \mid \Delta\mathbf{x}] = \mid E[y_t - y_{t-1} \mid \Delta\mathbf{x}] \mid .$$

Therefore under the assumption that $\mathbf{x}$'s are unbounded,[14]

$$H(\beta^*) - H(\mathbf{b}) = 2\int_{W_b} \mid E[y_t - y_{t-1} \mid \Delta\mathbf{x}] \mid dF_{\Delta\mathbf{x}} \geq 0. \tag{7.4.7}$$

Manski (1985, 1987) has shown that under fairly general conditions, the estimator maximizing the criterion function (7.4.1) is a strongly consistent estimator for $\beta^*$.

As discussed in Chapter 3 and early sections of this chapter, when $T$ is small the MLE of the (structural) parameters $\beta$ is consistent as $N \to \infty$ for the linear model and inconsistent for the nonlinear model in the presence of incidental parameters $\alpha_i$ because in the former case we can eliminate $\alpha_i$ by differencing while in the latter case we cannot. Thus, the error of estimating $\alpha_i$ is transmitted into the estimator of $\beta$ in the nonlinear case. The Manski semiparametric approach makes use of the linear structure of the latent variable representation (7.2.1) or (7.4.4). The individual specific effects $\alpha_i$ can again be eliminated by differencing and hence the lack of knowledge of $\alpha_i$ no longer affects the estimation of $\beta$.

The Manski maximum score estimator is consistent as $N \to \infty$ for unknown conditional distribution of $u_{it}$ given $\alpha_i$ and $\mathbf{x}_{it}, \mathbf{x}_{i,t-1}$. However, it converges at the rate $N^{1/3}$ which is much slower than the usual speed of $N^{1/2}$ for the parametric approach. Moreover, Kim and Pollard (1990) have shown that $N^{1/3}$ times the centered maximum score estimator converges in distribution to the random variable that maximizes a certain Gaussian process. This result shows that the maximum score estimator is probably not very useful in application because the properties of the limiting distribution are largely unknown.

The objective function (7.4.1) is equivalent to

$$\max_b H_N^*(\mathbf{b}) = N^{-1}\sum_{i=1}^{N}\sum_{t=2}^{T}[2 \cdot \mathbf{1}(\Delta y_{it} = 1) - 1]\mathbf{1}(\Delta\mathbf{x}_{it}'\mathbf{b} > 0) \tag{7.4.8}$$

subject to $\mathbf{b}'\mathbf{b} = 1$, $\mathbf{1}(A)$ is the indicator of the event $A$ with $\mathbf{1}(A) = 1$ if $A$ occurs and 0 otherwise. The complexity of the maximum score estimator and its slow rate of convergence are due to the discontinuity of the function $H_N(\mathbf{b})$ or $H_N^*(\mathbf{b})$. Horowitz (1992) suggests avoiding these difficulties by replacing

[14] If $\mathbf{x}$ is bounded, then identification may fail if $u_{it}$ is not logistic (Chamberlain (2010)).

$H_N^*(\mathbf{b})$ with a sufficiently smooth function $\tilde{H}_N(\mathbf{b})$ whose almost sure limit as $N \to \infty$ is the same as that of $H_N^*(\mathbf{b})$. Let $K(\cdot)$ be a continuous function of the real line into itself such that

(i) $| K(v) | < M$ for some finite $M$ and all $v$ in $(-\infty, \infty)$,
(ii) $\lim_{v \to -\infty} K(v) = 0$ and $\lim_{v \to \infty} K(v) = 1$.

The $K(\cdot)$ here is analogous to a cumulative distribution function. Let $\{h_N : N = 1, 2, \ldots\}$ be a sequence of strictly positive real numbers satisfying $\lim_{N \to \infty} h_N = 0$. Define

$$\tilde{H}_N(\mathbf{b}) = N^{-1} \sum_{i=1}^{N} \sum_{t=2}^{T} [2 \cdot \mathbf{1}(\Delta y_{it} = 1) - 1] K(\mathbf{b}' \Delta \mathbf{x}_{it} / h_N). \quad (7.4.9)$$

Horowitz (1992) defines a smoothed maximum score estimator as any solution that maximizes (7.4.9). Like Manski's estimator, $\beta$ can be identified only up to scale. Instead of using the normalization $\| \beta^* \| = 1$, Horowitz (1992) finds it is more convenient to use the normalization that the coefficient of one component of $\Delta \mathbf{x}$, say $\Delta \mathbf{x}_1$, to be equal to 1 in absolute value if its coefficient $\beta_1 \neq 0$ and the probability distribution of $\Delta \mathbf{x}_1$ conditional on the remaining components is absolutely continuous (with respect to Lebesgue measure).

The smoothed maximum score estimator is strongly consistent under the assumption that the distribution of $\Delta u_{it} = u_{it} - u_{i,t-1}$ conditional on $\Delta \mathbf{x}_{it}$ is symmetrically distributed with mean equal to 0. The asymptotic behavior of the estimator can be analyzed by taking a Taylor expansion of the first-order conditions and applying a version of the central limit theorem and the law of large numbers. The smoothed estimator of $\beta$ is consistent and, after centering and suitable normalization, is asymptotically normally distributed. Its rate of convergence is at least as fast as $N^{-2/5}$ and, depending on how smooth the distribution of $u$ and $\beta' \Delta \mathbf{x}$ are, can be arbitrarily close to $N^{-1/2}$.

### 7.4.2 A Root-$N$ Consistent Semiparametric Estimator

The speed of convergence of the smoothed maximum score estimator depends on the speed of convergence of $h_N \to 0$. Lee (1999) suggests a root-$N$ consistent semiparametric estimator that does not depend on a smoothing parameter by maximizing the double sums

$$\{N(N-1)\}^{-1} \sum_{i \neq j} \sum_{t=2}^{T} \operatorname{sgn}(\Delta \mathbf{x}_{it}' \mathbf{b} - \Delta \mathbf{x}_{jt}' \mathbf{b})(\Delta y_{it} - \Delta y_{jt}) \Delta y_{it}^2 \, \Delta y_{jt}^2$$

$$= \{N(N-1)\}^{-1} \sum_{\substack{i \\ i<j,}} \sum_{\substack{j \\ \Delta y_{it} \neq \Delta y_{jt}}} \sum_{\substack{t=2 \\ \Delta y_{it} \neq 0 \\ \Delta y_{jt} \neq 0}}^{T} \operatorname{sgn}(\Delta \mathbf{x}_{it}' \mathbf{b} - \Delta \mathbf{x}_{jt}' \mathbf{b})(\Delta y_{it} - \Delta y_{jt}) \quad (7.4.10)$$

with respect to $\mathbf{b}$. The consistency of the Lee estimator, $\hat{\mathbf{b}}$, follows from the fact that although $\Delta y_{it} - \Delta y_{jt}$ can take five values $(0, \pm 1, \pm 2)$, the event that $(\Delta y_{it} - \Delta y_{jt})\Delta y_{jt}^2 \Delta y_{it}^2 \neq 0$ excludes $(0, \pm 1)$ to make $\Delta y_{it} - \Delta y_{jt}$ binary (2 or $-2$). Conditional on given $j$, the first average over $i$ and $t$ converges to

$$E\{\, sgn(\Delta \mathbf{x}'\mathbf{b} - \Delta \mathbf{x}_j'\mathbf{b})(\Delta y - \Delta y_j)\Delta y^2 \Delta y_j^2 \mid \Delta \mathbf{x}_j, \Delta y_j\} \qquad (7.4.11)$$

The $\sqrt{N}$ speed of convergence follows from the second average of the smooth function (7.4.10).

Normalizing $\beta_1 = 1$, the asymptotic covariance matrix of $\sqrt{N}\left(\tilde{\mathbf{b}} - \tilde{\beta}\right)$ is equal to

$$4(E \bigtriangledown_2 \tau)^{-1}(E \bigtriangledown_1 \tau \bigtriangledown_1 \tau')(E \bigtriangledown_2 \tau)^{-1}, \qquad (7.4.12)$$

where $\tilde{\beta} = (\beta_2, \ldots, \beta_K)'$, and $\tilde{\mathbf{b}}$, its estimator,

$$\tau\left(\Delta y_j, \Delta \mathbf{x}_j, \tilde{\mathbf{b}}\right) \equiv E_{i|j}\{sgn(\Delta \mathbf{x}_i'\mathbf{b} - \Delta \mathbf{x}_j'\mathbf{b})(\Delta y_i - \Delta y_j)\Delta y_i^2 \Delta y_j^2\}, \quad i \neq j,$$

with $E_{i|j}$ denoting the conditional expectation of $(\Delta y_i, \Delta \mathbf{x}_i')$ conditional on $(\Delta y_j, \Delta \mathbf{x}_j')$, $\bigtriangledown_1 \tau$ and $\bigtriangledown_2 \tau$ denote the first and second derivative matrices of $\tau\left(\Delta y_j, \Delta \mathbf{x}_j, \tilde{\mathbf{b}}\right)$ with respect to $\tilde{\mathbf{b}}$.

The parametric approach requires the specification of the distribution of $u$. If the distribution of $u$ is misspecified, the MLE of $\beta$ is inconsistent. The semiparametric approach does not require the specification of the distribution of $u$ and permits its distribution to depend on $\mathbf{x}$ in an unknown way (heteroskedasticity of unknown form). It is consistent up to a scale whether the unobserved individual effects are treated as fixed or correlated with $\mathbf{x}$. However, the step of differencing $\mathbf{x}_{it}$ eliminates time-invariant variables from the estimation. Lee's (1999) $\sqrt{N}$ consistent estimator takes the additional differencing across individuals, $\Delta \mathbf{x}_i - \Delta \mathbf{x}_j$, and further reduces the dimension of estimable parameters by eliminating "period individual-invariant" variables (e.g., time dummies and macroeconomic shocks common to all individuals) from the specification. Moreover, the requirement that $u_{it}$ and $u_{i,t-1}$ are identically distributed conditional on $(\alpha_i, \mathbf{x}_{it}, \mathbf{x}_{i,t-1})$ does not allow the presence of the lagged dependent variables in $\mathbf{x}_{it}$. Neither can a semiparametric approach be used to generate the predicted probability conditional on $\mathbf{x}$ as in the parametric approach. All it can estimate is the relative effects of the explanatory variables.

## 7.5 DYNAMIC MODELS

### 7.5.1 The General Model

The static models discussed in the previous sections assume that the probability of moving (or staying) in or out of a state is independent of the occurrence or nonoccurrence of the event in the past. However, in a variety of contexts, such as in the study of the incidence of accidents (Bates and Neyman 1951), brand loyalty (Chintagunta, Kyriazidou, and Perktold 2001), labor force participation

(Heckman and Willis 1977; Hyslop 1999), and unemployment (Layton 1978), it is often noted that individuals who have experienced an event in the past are more likely to experience the event in the future than individuals who have not. In other words, the conditional probability that an individual will experience the event in the future is a function of past experience.

To analyze the intertemporal relationship among discrete variables, Heckman (1978a, 1981b) proposed a general framework in terms of a latent-continuous-random-variable crossing the threshold. He let the continuous random variable $y_{it}^*$ be a function of $\mathbf{x}_{it}$ and past occurrence of the event,

$$y_{it}^* = \beta' \mathbf{x}_{it} + \sum_{l=1}^{t-1} \gamma_\ell y_{i,t-l} + \phi \sum_{s=1}^{t-1} \prod_{l=1}^{s} y_{i,t-l} + v_{it}, \quad i = 1, \ldots, N, \ t = 1, \ldots, T \tag{7.5.1}$$

and

$$y_{it} = \begin{cases} 1 & \text{if} \quad y_{it}^* > 0, \\ 0 & \text{if} \quad y_{it}^* \leq 0. \end{cases} \tag{7.5.2}$$

The error term $v_{it}$ is assumed to be independent of $\mathbf{x}_{it}$ and is independently distributed over $i$, with a general intertemporal variance–covariance matrix $E\mathbf{v}_i\mathbf{v}_i' = \Omega$. The coefficient $\gamma_l$ measures the effects of experience of the event $l$ periods ago on current values of $y_{it}^*$. The coefficient $\phi$ measures the effect of the cumulative recent spell of experience in the state for those still in the state on the current value of $y_{it}^*$.

Specifications (7.5.1) and (7.5.2) accommodate a wide variety of stochastic models that appear in the literature. For example, let $\mathbf{x}_{it} = 1$, and let $v_{it}$ be independently identically distributed. If $\gamma_l = 0, l = 2, \ldots, T-1$, and $\phi = 0$, equations (7.5.1) and (7.5.2) generate a time-homogenous first-order Markov process. If $\gamma_l = 0, 1, \ldots, T-1$, and $\phi \neq 0$, a renewal process is generated. If $\gamma_l = 0, l = 1, \ldots, T-1$ and $\phi = 0$, a simple Bernoulli model results. If one allows $v_{it}$ to follow an autoregressive moving-average scheme, but keeps the assumption that $\gamma_l = 0, l = 1, \ldots, T-1$, and $\phi = 0$, the Coleman (1964) latent Markov model emerges.

As said before, repeated observations of a given group of individuals over time permit us to construct a model in which individuals may differ in their propensity to experience the event with the same $\mathbf{x}$. Such heterogeneity is allowed by decomposing the error term $v_{it}$ as

$$v_{it} = \alpha_i + u_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T. \tag{7.5.3}$$

where $u_{it}$ is independently distributed over $i$, with arbitrary serial correlation, and $\alpha_i$ is individual-specific and can be treated as a fixed constant or as random. Thus, for example, if the previoius assumptions on the Markov process

$$\gamma_l = 0, l = 2, \ldots, T-1, \text{ and } \phi = 0$$

hold, but $v_{it}$ follows a "components-of-variance" scheme (7.5.3), a compound first-order Markov process, closely related to previous work on the mover-stayer model (Goodman 1961; Singer and Spilerman 1976), is generated.

Specifications (7.5.1)–(7.5.3) allow for three sources of persistence (after controlling for the observed explanatory variables, $\mathbf{x}$). Persistence can be the result of serial correlation in the error term, $u_{it}$, or the result of "unobserved heterogeneity," $\alpha_i$, or the result of true state dependence through the term $\gamma y_{i,t-1}$ or $\phi \prod_{l=1}^{*} y_{i,t-1}$. Distinguishing the sources of persistence is important because a policy that temporariliy increases the probability that $y = 1$ will have different implications about future probabilities of experiencing an event.

When the conditional probability of an individual staying in a state is a function of past experience, two new issues arise. One is how to treat the initial observations. The second is how to distinguish true state dependence from spurious state dependence in which the past $y_{it}$ appears in the specification merely as a proxy for the unobserved individual effects, $\alpha_i$. The first issue could play a role in deriving consistent estimators for a given model. The second issue is important because the time dependence among observed events could arise either from the fact that the actual experience of an event has modified individual behavior or from unobserved components that are correlated over time, or from a combination of both.

### 7.5.2 Initial Conditions

When dependence among time-ordered outcomes is considered, just as in the dynamic linear-regression model, the problem of initial conditions must be resolved for a likelihood approach before parameters generating the stochastic process can be estimated. To focus the discussion on the essential aspects of the problem of initial conditions and its solutions, we assume that there are no exogenous variables and that the observed data are generated by a first-order Markov process. Namely,

$$y_{it}^{*} = \beta_0 + \gamma y_{i,t-1} + v_{it}, \tag{7.5.4}$$

$$y_{it} = \begin{cases} 1, & \text{if } y_{it}^{*} > 0, \\ 0, & \text{if } y_{it}^{*} \leq 0. \end{cases}$$

For ease of exposition we shall also assume that $u_{it}$ is independently normally distributed with mean 0 and variance $\sigma_u^2$ normalized to be equal to 1. It should be noted that the general conclusions of the following discussion also hold for other types of distributions.

In much applied work in the social sciences, two assumptions for initial conditions are typically invoked: (1) the initial conditions or relevant presample history of the process are assumed to be truly exogenous, or (2) the process is assumed to be in equilibrium. Under the assumption that $y_{i0}$ is a fixed nonstochastic constant for individual $i$, the joint probability of $\mathbf{y}_i' = (y_{i1}, \ldots, y_{iT})$,

given $\alpha_i$, is

$$\prod_{t=1}^{T} F(y_{it} \mid y_{i,t-1}, \alpha_i) = \prod_{t=1}^{T} \Phi\{(\beta_0 + \gamma y_{i,t-1} + \alpha_i)(2y_{it} - 1)\}, \quad (7.5.5)$$

where $\Phi$ is the standard normal cumulative distribution function. Under the assumption that the process is in equilibrium, the limiting marginal probability for $y_{it} = 1$ for all $t$, given $\alpha_i$, is (Karlin and Taylor 1975)[15]

$$P_i = \frac{\Phi(\beta_0 + \alpha_i)}{1 - \Phi(\beta_0 + \gamma + \alpha_i) + \Phi(\beta_0 + \alpha_i)}, \quad (7.5.6)$$

and the limiting probability for $y_{it} = 0$ is $1 - P_i$. Thus the joint probability of $(y_{i0}, \ldots, y_{iT})$, given $\alpha_i$ is

$$\prod_{t=1}^{T} \Phi\{(\beta + \gamma y_{i,t-1} + \alpha_i)(2y_{it} - 1)\} P_i^{y_{i0}} (1 - P_i)^{1 - y_{i0}}. \quad (7.5.7)$$

If $\alpha_i$ is random, with distribution $G(\alpha)$, the likelihood function for the random-effects model under the first assumption is

$$L = \prod_{i=1}^{N} \int \prod_{t=1}^{T} \Phi\{(\beta_0 + \gamma y_{i,t-1} + \alpha)(2y_{it} - 1)\} dG(\alpha). \quad (7.5.8)$$

The likelihood function under the second assumption is

$$L = \prod_{i=1}^{N} \int \prod_{t=1}^{T} \Phi\{(\beta_0 + \gamma y_{i,t-1} + \alpha)(2y_{it} - 1)\} \cdot P_i^{y_{i0}} (1 - P_i)^{1 - y_{i0}} dG(\alpha). \quad (7.5.9)$$

The likelihood functions (7.5.8) and (7.5.9) under both sets of assumptions about initial conditions are of closed form. When $\alpha_i$ is treated as random, the MLEs for $\beta_0$, $\gamma$, and $\sigma_\alpha^2$ are consistent if $N$ tends to infinity or if both $N$ and

[15] The transition-probability matrix of our homogeneous two-state Markov chain is

$$\mathcal{P} = \begin{bmatrix} 1 - \Phi(\beta_0 + \alpha_i) & \Phi(\beta_0 + \alpha_i) \\ 1 - \Phi(\beta_0 + \gamma + a_i) & \Phi(\beta_0 + \gamma + \alpha_i) \end{bmatrix}.$$

By mathematical induction, the $n$-step transition matrix is

$$\begin{aligned} \mathcal{P}^n = {} & \frac{1}{1 - \Phi(\beta_0 + \gamma + \alpha_i) + \Phi(\beta_0 + \alpha_i)} \\ & \cdot \left\{ \begin{bmatrix} 1 - \Phi(\beta_0 + \gamma + \alpha_i) & \Phi(\beta_0 + \alpha_i) \\ 1 - \Phi(\beta_0 + \gamma + \alpha_i) & \Phi(\beta_0 + \alpha_i) \end{bmatrix} \right. \\ & + [\Phi(\beta_0 + \gamma + \alpha_i) - \Phi(\beta_0 + \alpha))]^n \\ & \left. \cdot \begin{bmatrix} \Phi(\beta_0 + \alpha_i) & -\Phi(\beta_0 + \alpha_i) \\ -[1 - \Phi(\beta_0 + \gamma + \alpha_i)] & 1 - \Phi(\beta_0 + \gamma + \alpha_i) \end{bmatrix} \right\}. \end{aligned}$$

$T$ tend to infinity. When $\alpha_i$ is treated as a fixed constant (7.5.5), the MLEs for $\beta_0$, $\gamma$, and $\alpha_i$ are consistent only when $T$ tends to infinity. If $T$ is finite, the MLE is biased. Moreover, the limited results from Monte Carlo experiments suggest that, contrary to the static case, the bias is significant (Heckman 1981b).

However, the assumption that initial conditions are fixed constants is justifiable only if the disturbances that generate the process are serially independent and if a genuinely new process is fortuitously observed at the beginning of the sample. If the process has been in operation prior to the time it is sampled, or if the disturbances of the model are serially dependent as in the presence of individual specific random effects, the initial conditions are not exogenous. The assumption that the process is in equilibrium also raises problems in many applications, especially when time-varying exogenous variables are driving the stochastic process.

Suppose that the analyst does not have access to the process from the beginning; then the initial state for individual $i$, $y_{i0}$, cannot be assumed fixed. The initial state is determined by the process generating the panel sample. The sample likelihood function for the fixed-effects model is

$$L = \prod_{i=1}^{N} \prod_{t=1}^{T} \Phi\{(\beta_0 + \gamma y_{i,t-1} + \alpha_i)(2y_{it} - 1)\}\ f\ (y_{i0} \mid \alpha_i), \tag{7.5.10}$$

and the sample likelihood function for the random-effects models is

$$L = \prod_{i=1}^{N} \int_{-\infty}^{\infty} \prod_{t=1}^{T} \Phi\{(\beta_0 + \gamma y_{i,t-1} + \alpha)(2y_{it} - 1)\}\ f\ (y_{i0} \mid \alpha) dG(\alpha), \tag{7.5.11}$$

where $f(y_{i0} \mid \alpha)$ denotes the marginal probability of $y_{i0}$ given $\alpha_i$. Thus, unless $T$ is very large, maximizing (7.5.5) or (7.5.10) yields inconsistent estimates.[16]

Because $y_{i0}$ is a function of unobserved past values, besides the fact that the marginal distribution of $f(y_{i0} \mid \alpha)$ is not easy to derive, maximizing (7.5.10) or (7.5.11) is also considerably involved. Heckman (1981b) therefore suggested that we approximate the initial conditions for a dynamic discrete model by the following procedure:

1. Approximate the probability of $y_{i0}$, the initial state in the sample, by a probit model, with index function

$$y_{i0}^* = Q(\mathbf{x}_{it}) + \epsilon_{i0}, \tag{7.5.12}$$

and

$$y_{i0} = \begin{cases} 1 & \text{if } y_{i0}^* > 0, \\ 0 & \text{if } y_{i0}^* \leq 0, \end{cases} \tag{7.5.13}$$

[16] This can be easily seen by noting that the expectation of the first-derivative vector of (7.5.5) or (7.5.8) with respect to the structural parameters does not vanish at the true parameter value when the expectations are evaluated under (7.5.10) or (7.5.11).

where $Q(\mathbf{x}_{it})$ is a general function of $\mathbf{x}_{it}$, $t = 0, \ldots, T$, usually specified as linear in $\mathbf{x}_{it}$, and $\epsilon_{i0}$ is assumed to be normally distributed, with mean 0 and variance 1.

2. Permit $\epsilon_{i0}$ to be freely correlated with $v_{it}$, $t = 1, \ldots, T$.
3. Estimate the model by maximum likelihood without imposing any restrictions between the parameters of the structural system and parameters of the approximate reduced-form probability for the initial state of the sample.

Heckman (1981b) conducted Monte Carlo studies comparing the performances of the MLEs when assumption on initial $y_{i0}$ and $\alpha_i$ conform with the true data generating process, an approximate reduced-form probability for $y_{i0}$, and false fixed $y_{i0}$ and $\alpha_i$ for a first-order Markov process. The data for his experiment were generated by the random-effects model

$$y_{it}^* = \beta x_{it} + \gamma y_{i,t-1} + \alpha_i + u_{it}, \tag{7.5.14}$$

$$y_{it} = \begin{cases} 1 & \text{if } y_{it}^* > 0, \\ 0 & \text{if } y_{it}^* \leq 0, \end{cases}$$

where the exogenous variable $x_{it}$ was generated by (7.3.23). He let the process operate for 25 periods before selecting samples of 8 ($= T$) periods for each of the 100 ($=N$) individuals used in the 25 samples for each parameter set. Heckman's Monte Carlo results are produced in Table 7.2.

These results show that contrary to the static model, the fixed-effects probit estimator performs poorly. The greater the variance of the individual effects ($\sigma_\alpha^2$), the greater the bias. The $t$ statistics based on the estimated information matrix also lead to a misleading inference by not rejecting the false null hypotheses of $\gamma = \beta = 0$ in the vast majority of samples.

By comparison, Heckman's approximate solution performs better. Although the estimates are still biased from the true values, their biases are not significant, particularly when they are compared with the ideal estimates. The $t$ statistics based on the approximate solutions are also much more reliable than in the fixed-effects probit model, because they lead to a correct inference in a greater proportion of the samples.

Heckman's Monte Carlo results also point to a disquieting feature. Namely, the MLE produces a biased estimator even under the ideal conditions with a correctly specified likelihood function. Because a panel with 100 observations of three periods is not uncommon, this finding deserves further study.

### 7.5.3 A Conditional Approach

The likelihood approach cannot yield a consistent estimator when $T$ is fixed and $N$ tends to infinity if the individual effects are fixed. If the individual effects are random and independent of $\mathbf{x}$, the consistency of the MLE depends on the correct formulation of the probability distributions of the effects and initial observations. A semiparametric approach cannot be implemented for

Table 7.2. *Monte Carlo results for first-order Markov process*

| | | $\sigma_\alpha^2 = 3$ | | | $\sigma_\alpha^2 = 1$ | | |
|---|---|---|---|---|---|---|---|
| $\gamma$ | | $\beta = -0.1$ | $\beta = 1$ | $\beta = 0$ | $\beta = -0.1$ | $\beta = 1$ | $\beta = 0$ |
| Values of $\hat{\gamma}$ and $\hat{\beta}$ for the random-effects estimator with known initial conditions[a] | | | | | | | |
| 0.5 | $\hat{\gamma}$ | n.a.[c] | 0.57 | n.a.[c] | | | |
| | $\hat{\beta}$ | n.a.[c] | 0.94 | —[d] | | | |
| 0.1 | $\hat{\gamma}$ | 0.13 | 0.12 | 0.14 | | | |
| | $\hat{\beta}$ | −0.11 | 1.10 | — | | | |
| Values of $\hat{\gamma}$ and $\hat{\beta}$ for the approximate random-effects estimation[a] | | | | | | | |
| 0.5 | $\hat{\gamma}$ | 0.63 | 0.60 | 0.70 | n.a.[c] | 0.54 | 0.62 |
| | $\hat{\beta}$ | −0.131 | 0.91 | — | n.a.[c] | 0.93 | — |
| 0.1 | $\hat{\gamma}$ | 0.14 | 0.13 | 0.17 | 0.11 | 0.11 | 0.13 |
| | $\hat{\beta}$ | −0.12 | 0.92 | — | −0.12 | 0.95 | — |
| Values of $\hat{\gamma}$ and $\hat{\beta}$ for the fixed-effects estimator[b] | | | | | | | |
| 0.5 | $\hat{\gamma}$ | 0.14 | 0.19 | 0.03 | n.a.[c] | 0.27 | 0.17 |
| | $\hat{\beta}$ | −0.07 | 1.21 | — | n.a.[c] | 1.17 | — |
| 0.1 | $\hat{\gamma}$ | −0.34 | −0.21 | −0.04 | −0.28 | −0.15 | −0.01 |
| | $\hat{\beta}$ | −0.06 | 1.14 | — | −0.08 | 1.12 | — |

[a] $N = 100; T = 3$.
[b] $N = 100; T = 8$.
[c] Data not available because the model was not estimated.
[d] Not estimated.
*Source:* Heckman (1981b, Table 4.2).

a dynamic model because the strict exogeneity condition of explanatory variables is violated with the presence of lagged dependent variables as explanatory variables. When strict exogenity condition of the explanatory variables is violated, $E(\Delta u_{it} \mid \mathbf{x}_{it}, \mathbf{x}_{i,t-1}, y_{i,t-1}, y_{i,t-2}) \neq 0$. In other words, the one-to-one correspondence relation of the form (7.4.4) is violated. Hence, the Manski (1985) type maximum score estimator cannot be implemented. Neither can the (unrestricted) conditional approach be implemented. Consider the case of $T = 2$. The basic idea of conditional approach is to consider the probability of $y_{i2} = 1$ or 0 conditional on explanatory variables in both periods and conditional on $y_{i1} \neq y_{i2}$. If the explanatory variables of Prob $(y_{i2} = 1)$ include $y_{i1}$, then the conditional probability is either 1 or 0 according as $y_{i1} = 0$ or 1, hence provides no information about $\gamma$ and $\beta$.

However, in the case that $T \geq 3$ and $\mathbf{x}_{it}$ follows certain special pattern. Honoré and Kyriazidou (2000a) show that it is possible to generalize the conditional probability approach to consistently estimate the unknown parameters for the logit model or to generalize the maximum score approach without the

need of formulating the distribution of $\alpha_i$ or the probability distribution of the initial observations for certain types of discrete choice models. However, the estimators converge to the true values at the speed considerably slower than the usual square root $N$ rate.

Consider the model (7.5.4) with the assumption that $u_{it}$ is logistically distributed, then the model of $(y_{i0}, \ldots, y_{iT})$ is of the form

$$P(y_{i0} = 1 \mid \alpha_i) = P_0(\alpha_i) \tag{7.5.15}$$

$$P(y_{it} = 1 \mid \alpha_i, y_{i0}, \ldots, y_{i,t-1}) = \frac{\exp(\gamma y_{i,t-1} + \alpha_i)}{1 + \exp(\gamma y_{i,t-1} + \alpha_i)}, \tag{7.5.16}$$

for $t = 1, 2, \ldots, T$.

When $T \geq 3$, Chamberlain (1993) has shown that inference on $\gamma$ can be made independent of $\alpha_i$ by using a conditional approach.

For ease of exposition, we shall assume that $T = 3$ (i.e., there are four time series observations for each $i$). Consider the events

$$\begin{aligned} A &= \{y_{i0}, y_{i1} = 0, y_{i2} = 1, y_{i3}\}, \\ B &= \{y_{i0}, y_{i1} = 1, y_{i2} = 0, y_{i3}\}. \end{aligned}$$

where $y_{i0}$ and $y_{i3}$ can be either 1 or 0. Then

$$\begin{aligned} P(A) = P_0(\alpha_i)^{y_{i0}}[1 - P_0(\alpha_i)]^{1-y_{i0}} \cdot \frac{1}{1 + \exp(\gamma y_{i0} + \alpha_i)} \\ \cdot \frac{\exp(\alpha_i)}{1 + \exp(\alpha_i)} \cdot \frac{\exp(y_{i3}(\gamma + \alpha_i))}{1 + \exp(\gamma + \alpha_i)} \end{aligned} \tag{7.5.17}$$

and

$$\begin{aligned} P(B) = P_0(\alpha_i)^{y_{i0}}[1 - P_0(\alpha_i)]^{1-y_{i0}} \cdot \frac{\exp(\gamma y_{i0} + \alpha_i)}{1 + \exp(\gamma y_{i0} + \alpha_i)} \\ \cdot \frac{1}{1 + \exp(\gamma + \alpha_i)} \cdot \frac{\exp(\alpha_i y_{i3})}{1 + \exp(\alpha_i)}. \end{aligned} \tag{7.5.18}$$

Hence

$$\begin{aligned} P(A \mid A \cup B) &= P(A \mid y_{i0}, y_{i1} + y_{i2} = 1, y_{i3}) \\ &= \frac{\exp(\gamma y_{i3})}{\exp(\gamma y_{i3}) + \exp(\gamma y_{i0})} \\ &= \frac{1}{1 + \exp[\gamma(y_{i0} - y_{i3})]}, \end{aligned} \tag{7.5.19}$$

and

$$\begin{aligned} P(B \mid A \cup B) &= P(B \mid y_{i0}, y_{i1} + y_{i2} = 1, y_{i3}) \\ &= 1 - P(A \mid A \cup B) \\ &= \frac{\exp[\gamma(y_{i0} - y_{i3})]}{1 + \exp[\gamma(y_{i0} - y_{i3})]}. \end{aligned} \tag{7.5.20}$$

Equation (7.5.19) and (7.5.20) are in the binary logit form and does not depend on $\alpha_i$. The conditional log-likelihood

$$\log \tilde{L} = \sum_{i=1}^{N} 1(y_{i1} + y_{i2} = 1)\{y_{i1}[\gamma(y_{i0} - y_{i3})] - \log [1 + \exp \gamma(y_{i0} - y_{i3})]\} \tag{7.5.21}$$

is in the conditional logit form. Maximizing (7.5.21) yields $\sqrt{N}$ consistent estimator of $\gamma$, where $1(A) = 1$ if $A$ occurs and 0 otherwise.

When exogenous variables $\mathbf{x}_{it}$ also appear as explanatory variables in the latent response function

$$y_{it}^* = \beta' \mathbf{x}_{it} + \gamma y_{i,t-1} + \alpha_i + u_{it}, \tag{7.5.22}$$

we may write

$$P(y_{i0} = 1 \mid \mathbf{x}_i, \alpha_i) = P_0(\mathbf{x}_i, \alpha_i), \tag{7.5.23}$$

$$\begin{aligned} P(y_{it} &= 1 \mid \mathbf{x}_i, \alpha_i, y_{i0}, \ldots, y_{i,t-1}) \\ &= \frac{\exp (\mathbf{x}_{it}'\beta + \gamma y_{i,t-1} + \alpha_i)}{1 + \exp (\mathbf{x}_{it}'\beta + \gamma y_{i,t-1} + \alpha_i)}, \quad t = 1, \ldots, T. \end{aligned} \tag{7.5.24}$$

Let $P(y_{i0}) = P_0(\mathbf{x}_i, \alpha_i)^{y_{i0}}[1 - P_0(\mathbf{x}_i, \alpha_i)]^{1-y_{i0}}$. Suppose $T = 3$. Under (7.5.24),

$$\begin{aligned} P(A) = P(y_{i0}) \cdot &\frac{1}{1 + \exp (\mathbf{x}_{i1}'\boldsymbol{\beta} + \gamma y_{i0} + \alpha_i)} \cdot \frac{\exp (\mathbf{x}_{i2}'\boldsymbol{\beta} + \alpha_i)}{1 + \exp (\mathbf{x}_{i2}'\beta + \alpha_i)} \\ &\cdot \frac{\exp [(\mathbf{x}_{i3}'\boldsymbol{\beta} + \gamma + \alpha_i) y_{i3}]}{1 + \exp (\mathbf{x}_{i3}'\boldsymbol{\beta} + \gamma + \alpha_i)}. \end{aligned} \tag{7.5.25}$$

$$\begin{aligned} P(B) = P(y_{i0}) \cdot &\frac{\exp (\mathbf{x}_{i1}'\boldsymbol{\beta} + \gamma y_{i0} + \alpha_i)}{1 + \exp (\mathbf{x}_{i1}'\boldsymbol{\beta} + \gamma y_{i0} + \alpha_i)} \\ &\cdot \frac{1}{1 + \exp (\mathbf{x}_{i2}'\boldsymbol{\beta} + \gamma + \alpha_i)} \cdot \frac{[\exp (\mathbf{x}_{i3}'\boldsymbol{\beta} + \alpha_i) y_{i3}]}{1 + \exp (\mathbf{x}_{i3}'\boldsymbol{\beta} + \alpha_i)}. \end{aligned} \tag{7.5.26}$$

The denominator of $P(A)$ and $P(B)$ are different depending the sequence is of $(y_{i1} = 0, y_{i2} = 1)$ or $(y_{i1} = 1, y_{i2} = 0)$. Therefore, in general, $P(A \mid \mathbf{x}_i, \alpha_i, A \cup B)$ will depend on $\alpha_i$. However, if $\mathbf{x}_{i2} = \mathbf{x}_{i3}$, then the denominator of $P(A)$ and $P(B)$ are identical. Using the same conditioning method, Honoré and Kyriazidou (2000a) show that

$$\begin{aligned} &P(A \mid \mathbf{x}_i, \alpha_i, A \cup B, \mathbf{x}_{i2} = \mathbf{x}_{i3}) \\ &\quad = \frac{1}{1 + \exp [(\mathbf{x}_{i1} - \mathbf{x}_{i2})'\beta + \gamma(y_{i0} - y_{i3})]}, \end{aligned} \tag{7.5.27}$$

which does not depend on $\alpha_i$. If $\mathbf{x}_{it}$ is continuous, it may be rare that $\mathbf{x}_{i2} = \mathbf{x}_{i3}$. Honoré and Kyriazidou (2000a) propose estimating $\beta$ and $\gamma$ by maximizing

$$\sum_{i=1}^{N} \mathbf{1}(y_{i1} + y_{i2} = 1) K\left(\frac{\mathbf{x}_{i2} - \mathbf{x}_{i3}}{h_N}\right) \ln\left\{\frac{\exp\left[(\mathbf{x}_{i1} - \mathbf{x}_{i2})'\boldsymbol{\beta} + \gamma(y_{i0} - y_{i3})\right]^{y_{i1}}}{1 + \exp\left[(\mathbf{x}_{i1} - \mathbf{x}_{i2})'\boldsymbol{\beta} + \gamma(y_{i0} - y_{i3})\right]}\right\} \tag{7.5.28}$$

with respect to $\boldsymbol{\beta}$ and $\gamma$ (over some compact set) if $P(\mathbf{x}_{i2} = \mathbf{x}_{i3}) > 0$. Here $K(\cdot)$ is a kernel density function that gives appropriate weight to observation $i$, while $h_N$ is a bandwidth which shrinks to 0 as $N$ tends to infinity. The asymptotic theory will require that $K(\cdot)$ be chosen so that a number of regularity conditions are satisfied such as $\mid K(\cdot) \mid < M$ for some constant $M$, and $K(v) \to 0$ as $\mid v \mid \to \infty$ and $\int K(v)dv = 1$. For instance, $K(v)$ is often taken to be the standard normal density function and $h_N = cN^{-1/5}$ for some constant $c$. The effect of the term $K(\frac{\mathbf{x}_{i2} - \mathbf{x}_{i3}}{h_N})$ is to give more weight to observations for which $\mathbf{x}_{i2}$ is close to $\mathbf{x}_{i3}$. Their estimator is consistent and asymptotically normal although their speed of convergence is only $\sqrt{Nh_N^k}$, which is considerably slower than $\sqrt{N}$ where $k$ is the dimension of $\mathbf{x}_{it}$.

The conditional approach works for the logit model but it does not seem applicable for general nonlinear models. However, if the nonlinearity can be put in the single index form $F(a)$ with the transformation function $F$ being a strictly increasing distribution function, then Manski (1987) maximum score estimator for the static case can be generalized to the case where the lagged dependent variable is included in the explanatory variable set by considering

$$\begin{aligned} &P(A \mid \mathbf{x}_i, \alpha_i, \mathbf{x}_{i2} = \mathbf{x}_{i3}) \\ &= P_0(\mathbf{x}_i, \alpha_i)^{y_{i0}}[1 - P_0(\mathbf{x}_i, \alpha_i)]^{1-y_{i0}} \\ &\quad \cdot [1 - F(\mathbf{x}_{i1}'\beta + \gamma y_{i0} + \alpha_i)] \times F(\mathbf{x}_{i2}'\beta + \alpha_i) \\ &\quad \cdot [1 - F(\mathbf{x}_{i2}'\beta + \gamma + \alpha_i)]^{1-y_{i3}} \times F(\mathbf{x}_{i2}'\beta + \gamma + \alpha_i)^{y_{i3}} \end{aligned} \tag{7.5.29}$$

and

$$\begin{aligned} &P(B \mid \mathbf{x}_i, \alpha_i, \mathbf{x}_{i2} = \mathbf{x}_{i3}) \\ &= P_0(\mathbf{x}_i, \alpha_i)^{y_{i0}}[1 - P_0(\mathbf{x}_i, \alpha_i)]^{1-y_{i0}} \\ &\quad \cdot F(\mathbf{x}_{i1}'\beta + \gamma y_{i0} + \alpha_i) \times [1 - F(\mathbf{x}_{i2}'\beta + \gamma + \alpha_i)] \\ &\quad \cdot [1 - F(\mathbf{x}_{i2}'\beta + \alpha_i)]^{1-y_{i3}} \times F(\mathbf{x}_{i2}'\beta + \alpha_i)^{y_{i3}}. \end{aligned} \tag{7.5.30}$$

If $y_{i3} = 0$, then

$$\begin{aligned}
&\frac{P(A \mid \mathbf{x}_i, \alpha_i, \mathbf{x}_{i2} = \mathbf{x}_{i3})}{P(B \mid \mathbf{x}_i, \alpha_i, \mathbf{x}_{i2} = \mathbf{x}_{i3})} \\
&\quad = \frac{[1 - F(\mathbf{x}_{i1}'\beta + \gamma y_{i0} + \alpha_i)]}{[1 - F(\mathbf{x}_{i2}'\beta + \gamma + \alpha_i)]} \cdot \frac{F(\mathbf{x}_{i2}'\beta + \alpha_i)}{F(\mathbf{x}_{i1}'\beta + \gamma y_{i0} + \alpha_i)} \\
&\quad = \frac{[1 - F(\mathbf{x}_{i1}'\beta + \gamma y_{i0} + \alpha_i)]}{[1 - F(\mathbf{x}_{i2}'\beta + \gamma y_{i3} + \alpha_i)]} \cdot \frac{F(\mathbf{x}_{i2}'\beta + \gamma y_{i3} + \alpha_i)}{F(\mathbf{x}_{i1}'\beta + \gamma y_{i0} + \alpha_i)},
\end{aligned} \tag{7.5.31}$$

where the second equality follows from the fact that $y_{i3} = 0$. If $y_{i3} = 1$, then

$$\begin{aligned}
&\frac{P(A \mid \mathbf{x}_i, \alpha_i, \mathbf{x}_{i2} = \mathbf{x}_{i3})}{P(B \mid \mathbf{x}_i, \alpha_i, \mathbf{x}_{i2} = \mathbf{x}_{i3})} \\
&\quad = \frac{[1 - F(\mathbf{x}_{i1}'\beta + \gamma y_{i0} + \alpha_i)]}{[1 - F(\mathbf{x}_{i2}'\beta + \gamma + \alpha_i)]} \cdot \frac{F(\mathbf{x}_{i2}'\beta + \gamma + \alpha_i)}{F(\mathbf{x}_{i1}'\beta + \gamma y_{i0} + \alpha_i)} \\
&\quad = \frac{[1 - F(\mathbf{x}_{i1}'\beta + \gamma y_{i0} + \alpha_i)]}{[1 - F(\mathbf{x}_{i2}'\beta + \gamma y_{i3} + \alpha_i))]} \cdot \frac{F(\mathbf{x}_{i2}'\beta + \gamma y_{i3} + \alpha_i)}{F(\mathbf{x}_{i1}'\beta + \gamma y_{i0} + \alpha_i)},
\end{aligned} \tag{7.5.32}$$

where the second equality follows from the fact that $y_{i3} = 1$, so that $\gamma y_{i3} = \gamma$. In either case, the monotoncity of $F$ implies that

$$\frac{P(A)}{P(B)} \begin{cases} > 1 \text{ if } \mathbf{x}_{i2}'\beta + \gamma y_{i3} > \mathbf{x}_{i1}'\beta + \gamma y_{i0}, \\ < 1 \text{ if } \mathbf{x}_{i2}'\beta + \gamma y_{i3} < \mathbf{x}_{i1}\beta + \gamma y_{i0}. \end{cases}$$

Therefore,

$$\begin{aligned}
&\operatorname{sgn}[P(A \mid \mathbf{x}_i, \alpha_i, \mathbf{x}_{i2} = \mathbf{x}_{i3}) - P(B \mid \mathbf{x}_i, \alpha_i, \mathbf{x}_{i2} = \mathbf{x}_{i3})] \\
&\quad = \operatorname{sgn}[(\mathbf{x}_{i2} - \mathbf{x}_{i1})'\beta + \gamma(y_{i3} - y_{i0})].
\end{aligned} \tag{7.5.33}$$

Hence, Honoré and Kyriazidou (2000a) propose a maximum score estimator that maximizes the score function

$$\sum_{i=1}^{N} K\left(\frac{\mathbf{x}_{i2} - \mathbf{x}_{i3}}{h_N}\right)(y_{i2} - y_{i1})\operatorname{sgn}[(\mathbf{x}_{i2} - \mathbf{x}_{i1})'\beta + \gamma(y_{i3} - y_{i0})] \tag{7.5.34}$$

with respect to $\beta$ and $\gamma$. The Honoré and Kyriazidou estimator is consistent (up to a scale) if the density of $\mathbf{x}_{i2} - \mathbf{x}_{i3}$, $f(\mathbf{x}_{i2} - \mathbf{x}_{i3})$, is strictly positive at 0, $f(0) > 0$. (This assumption is required for consistency.)

We have discussed the estimation of panel data dynamic discrete choice model assuming that $T = 3$. It can be easily generalized to the case of $T > 3$ by maximizing the objective function that is based on sequences where an individual switches between alternatives in any two of the middle $T - 1$

periods:

$$\sum_{i=1}^{N} \sum_{1 \le s < t \le T-1} \mathbf{1}\{y_{is} + y_{it} = 1\} K\left(\frac{\mathbf{x}_{i,t+1} - \mathbf{x}_{i,s+1}}{h_N}\right) \cdot \ln\left(\frac{\exp\left[(\mathbf{x}_{is} - \mathbf{x}_{it})'\beta + \gamma(y_{i,s-1} - y_{i,t+1}) + \gamma(y_{i,s+1} - y_{i,t-1})\mathbf{1}(t-s>\mathbf{1})\right]^{y_{is}}}{\mathbf{1} + \exp\left[(\mathbf{x}_{is} - \mathbf{x}_{it})'\beta + \gamma(y_{i,s-1} - y_{i,t+1}) + \gamma(y_{i,s+1} - y_{i,t-1})\mathbf{1}(t-s>\mathbf{1})\right]}\right) \tag{7.5.35}$$

The conditional approach does not require modeling of the initial observations of the sample. Neither does it make any assumptions about the statistical relationship of the individual effects with the observed explanatory variables or with the initial conditions. However, it also suffers from the limitation that $\mathbf{x}_{is} - \mathbf{x}_{it}$ has support in a neighborhood of 0 for any $t \neq s$, which rules out time dummies as explanatory variables.[17] The fact that individual effects cannot be estimated also means that it is not possible to carry out predictions or compute elasticities for individual agents at specified values of the explanatory variables.

### 7.5.4 State Dependence versus Heterogeneity

There are two diametrically opposite explanations for the often observed empirical regularity with which individuals who have experienced an event in the past are more likely to experience that event in the future. One explanation is that as a consequence of experiencing an event, preferences, prices, or constraints relevant to future choices are altered. A second explanation is that individuals may differ in certain unmeasured variables that influence their probability of experiencing the event but are not influenced by the experience of the event. If these variables are correlated over time and are not properly controlled, previous experience may appear to be a determinant of future experience solely because it is a proxy for such temporally persistent unobservables. Heckman (1978a, 1981a,c) has termed the former case "true state dependence" and the latter case "spurious state dependence," because in the former case, past experience has a genuine behavioral effect in the sense that an otherwise identical individual who has not experienced the event will behave differently in the future than an individual who has experienced the event. In the latter case, previous experience appears to be a determinant of future experience solely because it is a proxy for temporally persistent unobservables that determine choices.

The problem of distinguishing between true and spurious state dependencies is of considerable substantive interest. To demonstrate this, let us consider some work in the theory of unemployment. Phelps (1972) argued that current

[17] See Arellano and Carrasco (2003) for a GMM approach to estimate the dynamic random-effects probit model.

unemployment has a real and lasting effect on the probability of future unemployment. Hence, short-term economic policies that alleviate unemployment tend to lower aggregate unemployment rates in the long run by preventing the loss of work-enhancing market experience. On the other hand, Cripps and Tarling (1974) maintained the opposite view in their analysis of the incidence and duration of unemployment. They assumed that individuals differ in their propensity to experience unemployment and in their unemployment duration times and those differences cannot be fully accounted for by measured variables. They further assumed that the actual experience of having been unemployed or the duration of past unemployment does not affect future incident or duration. Hence, in their model, short-term economic policies have no effect on long-term unemployment.

Because the unobserved individual effects, $\alpha_i$, persist over time, ignoring these effects of unmeasured variables (heterogeneity) creates serially correlated residuals. This suggests that we cannot use the conditional probability, given past occurrence not equal to the marginal probability alone, Prob $(y_{it} \mid y_{i,t-s}, \mathbf{x}_{it}) \neq$ Prob $(y_{it} \mid \mathbf{x}_{it})$, to test for true state dependence against spurious state dependence, because this inequality may be a result of past information on $y$ yielding information on the unobserved specific effects. A proper test for dependence should control for the unobserved individual-specific effects.

When conditional on the individual effects, $\alpha_i$, the error term $u_{it}$ is serially uncorrelated, a test for state dependence can be implemented by controlling the individual effects and testing for the conditional probability equal to the marginal probability,

$$\text{Prob}\,(y_{it} \mid y_{i,t-s}, \mathbf{x}_{it}, \alpha_i) = \text{Prob}(y_{it} \mid \mathbf{x}_{it}, \alpha_i). \tag{7.5.36}$$

When $N$ is fixed and $T \to \infty$, likelihood ratio tests can be implemented to test (7.5.36).[18] However, if $T$ is finite, controlling $\alpha_i$ to obtain consistent estimator for the coefficient of lagged dependent variable imposes very restrictive conditions on the data which severly limits the power of the test, as shown in Section 7.4.

If $\alpha_i$ are treated as random and the conditional distribution of $\alpha_i$ given $\mathbf{x}_i$ is known, a more powerful test is to use an unconditional approach. Thus, one may test true state dependence versus spurious state dependence by testing the

[18] Let $P_{it} = \text{Prob}\,(y_{it} \mid \mathbf{x}_{it}, \alpha_i)$ and $P_{it}^* = \text{Prob}\,(y_{it} \mid y_{i,t-\ell}, \mathbf{x}_{it}, \alpha_i)$. Let $\hat{P}_{it}$ and $\hat{P}_{it}^*$ be the MLEs obtained by maximizing $\mathcal{L} = \prod_i \prod_t P_{it}^{y_{it}}(1 - P_{it})^{1-y_{it}}$ and $\mathcal{L}^* = \prod_i \prod_t P_{it}^{*y_{it}}(1 - P_{it}^*)^{1-y_{it}}$ with respect to unknown parameters, respectively. A likelihood-ratio test statistic for the null hypothesis (7.5.36) is $-2 \log [\mathcal{L}(\hat{P}_{it})/\mathcal{L}(\hat{P}_{it}^*)]$. When conditional on $\mathbf{x}_{it}$ and $\alpha_i$, there are repeated observations; we can also use the Pesaran chi-square goodness-of-fit statistic to test (7.5.36). For details, see Bishop, Fienberg, and Holland (1975, Chapter 7). However, in the finite-$T$ case, the testing procedure cannot be implemented, as the $\alpha_i's$ are unknown and cannot be consistently estimated.

significance of the MLE of $\gamma$ of the log-likelihood

$$\sum_{i=1}^{N} \log \int \prod_{t=1}^{T} \left\{ F(\mathbf{x}_{it}'\beta + \gamma y_{i,t-1} + \alpha_i)^{y_{it}} \left[ 1 - F(\mathbf{x}_{it}'\beta + \gamma y_{i,t-1} + \alpha_i) \right]^{1-y_{it}} \right. \tag{7.5.37}$$

$$\left. \cdot P(\mathbf{x}_i, \alpha)^{y_{i0}} \left[ 1 - P(\mathbf{x}_i, \alpha) \right]^{1-y_{i0}} \right\} G(\alpha_i \mid \mathbf{x}_i) d\alpha_i .$$

When conditional on the individual effects, $\alpha_i$, the error term $u_{it}$ remains serially correlated, the problem becomes more complicated. The conditional probability, Prob $(y_{it} \mid y_{i,t-1}, \alpha_i)$, not being equal to the marginal probability, Prob $(y_{it} \mid \alpha_i)$, could be because of past $y_{it}$ containing information on $u_{it}$. A test for state dependence cannot simply rely on the multinominal distribution of the $(y_{i1}, \ldots, y_{iT})$ sequence. The general framework (7.5.1) and (7.5.2) proposed by Heckman (1978a, 1981a,b) accommodates very general sorts of heterogeneity and structural dependence. It permits an analyst to combine models and test among competing specifications within a unified framework. However, the computations of maximum-likelihood methods for the general models could be quite involved. It would be useful to rely on simple methods to explore data before implementing the computationally cumbersome maximum-likelihood method for a specific model.

Chamberlain (1978b) suggested a simple method to distinguish true state dependence from spurious state dependence. He noted that just as in the continuous models, a key distinction between state dependence and serial correlation is whether or not there is a dynamic response to an intervention. This distinction can be made clear by examining (7.5.1). If $\gamma = 0$, a change in $\mathbf{x}$ has its full effect immediately, whereas if $\gamma \neq 0$, this implies a distributed-lag response to a change in $\mathbf{x}$. The lag structure relating $y$ to $\mathbf{x}$ is not related to the serial correlation in $u$. If $\mathbf{x}$ is increased in period $t$ and then returned to its former level, the probability of $y_{i,t+1}$ is not affected if $\gamma = 0$, because by assumption the distribution of $u_{it}$ was not affected. If $\gamma \neq 0$, then the one-period shift in $\mathbf{x}$ will have lasting effects. An intervention that affects the probability of $y$ in period $t$ will continue to affect the probability of $y$ in period $t + 1$, even though the intervention was presented only in period $t$. In contrast, an interpretation of serial correlation is that the shocks $(u)$ tend to persist for more than one period and that $y_{i,t-s}$ is informative only in helping to infer $u_{it}$ and hence to predict $u_{it}$. Therefore, a test that is not very sensitive to functional form is to simply include lagged $\mathbf{x}'s$ without lagged $y$. After conditioning on the individual-specific effect $\alpha$, there may be two possible outcomes. If there is no state dependence, then

$$\text{Prob}(y_{it} = 1 \mid \mathbf{x}_{it}, \mathbf{x}_{i,t-1}, \ldots, \alpha_i) = \text{ Prob } (y_{it} = 1 \mid \mathbf{x}_{it}, \alpha_i). \tag{7.5.38}$$

If there is state dependence, then

$$\text{Prob}(y_{it} = 1 \mid \mathbf{x}_{it}, \mathbf{x}_{i,t-1}, \ldots, \alpha_i) \neq \text{ Prob } (y_{it} = 1 \mid \mathbf{x}_{it}, \alpha_i). \tag{7.5.39}$$

While the combination of (7.5.38) and (7.5.39) provides a simple form to distinguish pure heterogeneity, state dependence, and serial correlation, we cannot make further distinctions with regard to different forms of state dependence and heterogeneity, and serial correlation should (7.5.38) be rejected. Models of the form (7.5.1) and (7.5.2) may still have to be used to further narrow down possible specifications.

### 7.5.5 Two Examples

The control of heterogeneity plays a crucial role in distinguishing true state dependence from spurious state dependence. Neglecting heterogeneity and the issue of initial observations can also seriously bias the coefficient estimates. It is important in estimating dynamic models that the heterogeneity in the sample be treated correctly. To demonstrate this, we use the female-employment models estimated by Heckman (1981c) and household brand choices estimated by Chintagunta, Kyriazidou, and Perktold (2001) as examples.

#### *7.5.5.1 Female Employment*

Heckman (1981c) used the first three-year sample of women aged 45–59 in 1968 from the Michigan Panel Survey of Income dynamics to study married women's employment decisions. A woman is defined to be a market participant if she worked for money any time in the sample year. The set of explanatory variables is as follows: the woman's education; family income, excluding the wife's earnings; number of children younger than six; number of children at home; unemployment rate in the county in which the woman resided; the wage of unskilled labor in the county (a measure of the availability of substitutes for a woman's time in the home); the national unemployment rate for prime-age men (a measure of aggregate labor-market tightness); two types of prior work experience: within-sample work experience and presample work experience. The effect of previous work experience is broken into two components, because it is likely that presample experience exerts a weaker measured effect on current participation decisions than more recent experience. Furthermore, because the data on presample work experience are based on a retrospective question and therefore are likely to be measured with error, Heckman replaces them by predicted values based on a set of regressors.

Heckman fitted the data to various multivariate probit models of the form (7.5.1) and (7.5.2) to investigate whether or not work experience raises the probability that a woman will work in the future (by raising her wage rates) and to investigate the importance of controlling for heterogeneity in utilizing panel data. Maximum-likelihood-coefficient estimates for the state-dependent models under the assumptions of stationary intertemporal covariance matrix

$$\Omega = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ & 1 & \rho_{23} \\ & & 1 \end{bmatrix},$$

first-order Markov process ($v_{it} = \rho v_{i,t-1} + u_{it}$), and no heterogeneity ($v_{it} = u_{it}$) are presented in columns 1, 2, and 3, respectively, of Table 7.3.[19] Coefficient estimates for no state dependence with general stationary intertemporal correlation, first-order Markov process, conventional random effects error-component formulation $v_{it} = \alpha_i + u_{it}$, equivalent to imposing the restriction that $\rho_{12} = \rho_{13} = \rho_{23} = \sigma_\alpha^2/(\sigma_u^2 + \sigma_\alpha^2)$, and no heterogeneity are presented in columns 4, 5, 6, and 7, respectively. A Heckman–Willis (1977) model with time-invariant exogenous variables and conventional error-component formulation was also estimated and is presented in column 8.

Likelihood ratio test statistics (twice the difference of the log-likelihood value) against the most general model (column 1, Table 7.3) indicate the acceptance of recent labor-market experience as an important determinant of current employment decision, with unobservables determining employment choices following a first-order Markov process (column 2, Table 7.3) as a maintained hypothesis, and the statistics clearly reject all other formulations. In other words, Heckman's study found that work experience, as a form of general and specific human capital investment, raises the probability that a woman will work in the future, even after accounting for serial correlation of a very general type. It also maintained that there exist unobserved variables that affect labor participations. However, initial differences in unobserved variables tend to be eliminated with the passage of time. But this homogenizing effect is offset in part by the impact of prior work experience that tends to accentuate initial differences in the propensity to work.

Comparison of the estimates of the maintained hypothesis with estimates of other models indicates that the effect of recent market experience on employment is dramatically overstated in a model that neglects heterogeneity. The estimated effect of recent market experience on current employment status recorded in column 3, Table 7.3, overstates the impact by a factor of 10 (1.46 vs. 0.143)! Too much credit will be attributed to past experience as a determinant of employment if intertemporal correlation in the unobservables is ignored. Likewise for the estimated impact of national unemployment on employment. On the other hand, the effect of children on employment is understated in models that ignore heterogeneity.

Comparisons of various models' predictive performance on sample-run patterns (temporal employment status) are presented in Table 7.4. It shows that dynamic models ignoring heterogeneity under-predict the number of individuals who work all of the time and over-predict the number who do not work at all. It also overstates the estimated frequency of turnover in the labor force. In fact, comparing the performances of the predicted run patterns for the dynamic and static models without heterogeneity (column 3 and 7 or Table 7.3 and columns 3 and 4 of Table 7.4) suggests that introducing "lagged employment status" into a model as a substitute for a more careful treatment of heterogeneity is an imperfect procedure. In this case, it is worse than using no proxy at all. Nor

[19] A nonstationary model was also estimated by Heckman (1981c), but because the data did not reject stationarity, we shall treat the model as having stationary covariance.

Table 7.3. *Estimates of employment models for women aged 45–59 in 1968*[a]

| Variable | (1) | (2) | (3) |
|---|---|---|---|
| Intercept | −2.576 (4.6) | 1.653 (2.5) | 0.227 (0.4) |
| No. of children aged<6 | −0.816 (2.7) | −0.840 (2.3) | −0.814 (2.1) |
| County unemployment rate (%) | −0.035 (1.5) | −0.027 (1.0) | −0.018 (0.57) |
| County wage rate ($/h) | 0.104 (0.91) | 0.104 (0.91) | 0.004 (0.02) |
| Total no. of children | −0.146 (4.3) | −0.117 (2.2) | −0.090 (2.4) |
| Wife's education (years) | 0.162 (6.5) | 0.105 (2.8) | 0.104 (3.7) |
| Family income, excluding wife's earnings | $-0.363 \times 10^{-4}$ (4.8) | $-0.267 \times 10^{-4}$ (2.7) | $-0.32 \times 10^{-4}$ (3.6) |
| National unemployment rate | −0.106 (0.51) | −0.254 (1.4) | −1.30 (6) |
| Recent experience | 0.143 (0.95) | 0.273 (1.5) | 1.46 (12.2) |
| Predicted presample experience | 0.072 (5.8) | 0.059 (3.4) | 0.045 (3.4) |
| Serial-correlation coefficient: | | | |
| $\rho_{12}$ | 0.913 | — | — |
| $\rho_{13}$ | 0.845 | | |
| $\rho_{23}$ | 0.910 | | |
| $\rho$ | — | 0.873 (14.0) | — |
| $\sigma_\alpha^2/(\sigma_u^2+\sigma_\alpha^2)$ | — | — | — |
| Log likelihood | −237.74 | −240.32 | −263.65 |

[a] Asymptotic normal test statistics in parentheses; these statistics were obtained from the estimating information matrix.

does a simple static model with a "components-of-variance" scheme (column 8 of Table 7.3, column 5 of Table 7.4) perform any better. Dynamic models that neglect heterogeneity (column 3 of Table 7.4) overestimate labor-market turnover, whereas the static model with a conventional variance components formulation (column 5 of Table 7.4) overstates the extent of heterogeneity and the degree of intertemporal correlation. It over-predicts the number who never work during these three years and underpredicts the number who always work.

This example suggests that considerable care should be exercised in utilizing panel data to discriminate among state dependence, heterogeneity, and serial correlations. Improper control for heterogeneity can lead to erroneous parameter estimates and dramatically overstate the impact of past experience on current choices.

### *7.5.5.2 Household Brand Choices*

Chintagunta, Kyriazidou, and Perktold (2001) use the A.C. Nielson data on yogurt purchases in Sioux Falls, South Dakota between September 17, 1986 and August 1, 1988 to study yogurt brand loyalty. They focus on the 6 oz.

Table 7.3. *(cont.)*

| (4) | (5) | (6) | (7) | (8) |
|---|---|---|---|---|
| −2.367 (6.4) | −2.011 (3.4) | −2.37 (5.5) | −3.53 (4.6) | −1.5 (0) |
| −0.742 (2.6) | −0.793 (2.1) | −0.70 (2.0) | −1.42 (2.3) | −0.69 (1.2) |
| −0.030 (1.5) | −0.027 (1.2) | −0.03 (1.6) | −0.059 (1.3) | 0.046 (11) |
| 0.090 (0.93) | 0.139 (1.5) | 0.13 (1.4) | 0.27 (1.1) | 0.105 (0.68) |
| −0.124 (4.9) | −0.116 (2.2) | −0.161 (4.9) | −0.203 (3.9) | −0.160 (6.1) |
| 0.152 (7.3) | 0.095 (2.5) | 0.077 (3) | 0.196 (4.8) | 0.105 (3.3) |
| $-0.312 \times 10^{-4}$ (5.2) | $-0.207 \times 10^{-4}$ (2.3) | $-0.2 \times 10^{-4}$ (2.6) | $-0.65 \times 10^{-4}$ (5.1) | $-0.385 \times 10^{-4}$ (20) |
| −0.003 (0.38) | −0.021 (0.26) | 0.02 (3) | 1.03 (0.14) | −0.71 (0) |
| —[b] | — | — | — | — |
| 0.062 (0.38) | 0.062 (3.5) | 0.091 (7.0) | 0.101 (5.4) | 0.095 (11.0) |
| 0.917 | — | — | — | — |
| 0.873 | — | — | — | — |
| 0.946 | — | — | — | — |
| — | −0.942 (50) | — | — | — |
| — | — | 0.92 (4.5) | — | 0.941 (4.1) |
| −239.81 | −243.11 | −244.7 | −367.3 | −242.37 |

[b] Not estimated.

*Source:* Heckman (1981c, Table 3.2).

packages of the two dominant yogurt brands, Yoplait and Nordica, for the analysis. These brands account for 18.4 and 19.5 percent of yogurt purchases by weight. Only data for households that have at least two consecutive purchases of any one of the two brands are considered. This leaves 737 households and 5618 purchase occasions, out of which 2718 are for Yoplait and the remaining 2900 for Nordica. The panel is unbalanced.[20] The minimum number of purchase occasions per household is 2 and the maximum is 305. The mean number of purchase is 9.5 and the median is 5.

The model they estimate is given by

$$\text{Prob}(y_{it} = 1 \mid \mathbf{x}_{it}, y_{i0}, \ldots, y_{i,t-1}, \alpha_i) = \frac{\exp(\mathbf{x}'_{it}\beta + \gamma y_{i,t-1} + \alpha_i)}{1 + \exp(\mathbf{x}'_{it}\beta + \gamma y_{i,t-1} + \alpha_i)}, \tag{7.5.40}$$

where $y_{it} = 1$ if household $i$ chooses Yoplait in period $t$ and $y_{it} = 0$ if household $i$ chooses Nordica. The exogenous variables in $\mathbf{x}_{it}$ are the difference in the

[20] One can modify the estimator (7.5.33) by replacing $T$ with $T_i$.

Table 7.4. *Comparisons of employment models using run data: Women aged 45–59 in 1968*

| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| Run pattern | Actual number | Number predicted from state-dependent model with heterogeneity (column 2 of Table 7.3) | Probit model that ignores heterogeneity (column 3 of Table 7.3) | Probit model that ignores heterogeneity and recent-sample state dependence (column 7 of Table 7.3) | Number predicted from Heckman–Willis model (column 8 of Table 7.3) |
| 0,0,0 | 96 | 94.2 | 145.3 | 36.1 | 139.5 |
| 0,0,1 | 5 | 17.6 | 38.5 | 20.5 | 4.1 |
| 0,1,0 | 4 | 1.8 | 1.9 | 20.2 | 4.1 |
| 1,0,0 | 8 | 2.6 | 0.35 | 20.6 | 4.1 |
| 1,1,0 | 5 | 1.4 | 0.02 | 21.2 | 3.6 |
| 1,0,1 | 2 | 2.4 | 1.38 | 21.1 | 3.6 |
| 0,1,1 | 2 | 16.4 | 8.51 | 21.7 | 3.6 |
| 1,1,1 | 76 | 61.5 | 2.05 | 36.6 | 34.9 |
| $\chi^{2\,c}$ | — | 48.5 | 4,419 | 221.8 | 66.3 |

[a] Data for 1971, 1972, and 1973, three years following the sample data, were used to estimate the model.

[b] 0 corresponds to not working; 1 corresponds to working; thus, 1,1,0 corresponds to a woman who worked the first two years of the sample and did not work in the final year.

[c] This is the standard chi-square statistic for goodness of fit. The higher the value of the statistic, the worse the fit.

*Source:* Heckman (1981c).

natural logarithm of the price (coefficient denoted by $\beta_P$) and the difference in the dummy variables for the two brands that describe whether the brand was displayed in the store and featured in an advertisement that week (coefficients denoted by $\beta_D$ and $\beta_F$ respectively). Among the many models they estimated, Table 7.5 presents the results of

1. The pooled logit model, with the lagged choice treated as exogenous assuming there are no individual specific effects (PLL)
2. The Chamberlain (1982) conditional logit approach with the lagged choice treated as exogenous (CLL)
3. The pooled logit approach with normally distributed random effects with mean $\mu$ and variance $\sigma_\alpha^2$, with the initial choice treated as exogenous (PLL-HET)
4. The pooled logit approach with normally distributed random effects and the initial probability of choosing 1 given $(\mathbf{x}_i, \alpha_i)$ assuming at the

Table 7.5. *Estimates of brand choices using various approaches (standard errors in parentheses)*

| Model | $\beta_p$ | $\beta_d$ | $\beta_f$ | $\gamma$ | $\mu_\alpha$ | $\sigma_\alpha$ |
|---|---|---|---|---|---|---|
| CLL | −3.347 | 0.828 | 0.924 | −0.068 | | |
| | (0.399) | (0.278) | (0.141) | (0.140) | | |
| PLL | −3.049 | 0.853 | 1.392 | 3.458 | −0.333 | |
| | (0.249) | (0.174) | (0.091) | (0.084) | (0.102) | |
| PLLHET | −3.821 | 1.031 | 1.456 | 2.126 | 0.198 | 1.677 |
| | (0.313) | (0.217) | (0.113) | (0.114) | (0.150) | (0.086) |
| PLLHETE | −4.053 | 0.803 | 1.401 | 1.598 | 0.046 | 1.770 |
| | (0.274) | (0.178) | (0.115) | (0.115) | (0.133) | (0.102) |
| HK05 | −3.477 | 0.261 | 0.782 | 1.223 | | |
| | (0.679) | (0.470) | (0.267) | (0.352) | | |
| HK10 | −3.128 | 0.248 | 0.759 | 1.198 | | |
| | (0.658) | (0.365) | (0.228) | (0.317) | | |
| HK30 | −2.644 | 0.289 | 0.724 | 1.192 | | |
| | (0.782) | (0.315) | (0.195) | (0.291) | | |
| PLLHET-S[a] | −3.419 | 1.095 | 1.291 | 1.550 | 0.681 | 1.161 |
| | (0.326) | (0.239) | (0.119) | (0.117) | (0.156) | (0.081) |

[a] The PLLHET estimates after excluding those households that are completely loyal to one brand.

*Source*: Chintagunta, Kyriazidou, and Perktold (2001, Table 3).

steady state, which is approximated by

$$\frac{F(\bar{\mathbf{x}}_i'\beta + \alpha_i)}{1 - F(\bar{\mathbf{x}}_i'\beta + \gamma + \alpha_i) + F(\bar{\mathbf{x}}_i'\beta + \alpha_i)}, \tag{7.5.41}$$

where $F(a) = \exp(a)/(1 + \exp(a))$ and $\bar{\mathbf{x}}_i$ denotes the individual time series mean of $\mathbf{x}_{it}$ (PLLHETE)

5. The Honoré and Kyriazidou (2000a) approach, where $h_N = c \cdot N^{-1/5}$ with $c = 0.5$ (HK05), 1.0 (HK10), and 3.0 (HK30)

Table 7.5 reveals that almost all procedures yield statistically significant coefficients with the expected signs. An increase in the price of a brand reduces the probability of choosing the brand, and the presence of a store display or of a feature advertisement for a brand makes purchase of that brand more likely. Also, apart from CLL, all methods produce positive and statistically significant estimates for $\gamma$, that is, a previous purchase of a brand increases the probability of purchasing the same brand in the next period. The lagged choice is found to have a large positive effect in brand choice for pooled methods assuming no heterogeneity: PLL estimates of $\gamma$ is 3.5. However, introducing heterogeneity lowers it substantially to 2.1 (PLLHET). The estimate of $\gamma$ further drops to 1.598 (PLL-HETE) when the initial observations are treated as endogenous, and drops to about 1.2 using the Honoré–Kyriazidou estimator. Nevertheless, they do indicate that after controlling for the effects of $\alpha_i$, a previous purchase of a

brand increases the probability of purchasing the same brand in the next period, although their impact is substantially reduced compared to the case of assuming no heterogeneity. There is also an indication of substantial heterogeneity in the sample. All methods that estimate random effects give high values for the standard deviation of the household effects, $\sigma_\alpha$ about 1.7, bearing in mind that $\sigma_u$ is normalized to 1 only.

In general, the size of the estimated parameters varies considerably across estimation methods. There is also some sensitivity in the HK point estimates of all coefficients with respect to the bandwidth choice. To investigate this issue further and identify situations where the different methods are most reliable in producing point estimates, Chintagunta, Kyriazidou, and Perktold (2001) further conduct Monte Carlo studies. Their results indicate that the conditional likelihood procedures are the most robust in estimating the coefficients on the exogenous variables. However, the coefficient on the lagged dependent variable is significantly underestimated. The pooled procedures are quite sensitive to model misspecification, often yielding large biases for key economic parameters. The estimator proposed by Honoré and Kyriazidou (2000a) performs quite satisfactory despite a loss of precision because their method de facto only uses substantially smaller number of observations than other methods due to the use of the weighting scheme $K\left(\frac{\mathbf{x}_{it}-\mathbf{x}_{is}}{h_N}\right)$.

## 7.6 ALTERNATIVE APPROACHES FOR IDENTIFYING STATE DEPENDENCE

Section 7.5 focuses on getting consistent estimators for dynamic panel discrete choice models with individual-specific effects. If individual-specific effects are treated as random, the consistency of dynamic models requires the knowledge of the conditional distribution of individual-specific effects $\alpha_i$ given the $T$ time series observations of the $K \times 1$ exogenous variables, $\mathbf{x}_{it}, \mathbf{x}_i' = (\mathbf{x}_{i1}', \ldots, \mathbf{x}_{iT}')$, $G(\alpha_i \mid \mathbf{x}_i)$, and the initial value distribution given $\mathbf{x}_i$, $P(y_{i0} \mid \mathbf{x}_i)$. If $\alpha_i$ is treated as a fixed constant, the consistency of the MLE requires $T \to \infty$. If $T$ is finite, the conditions for obtaining consistent estimator of the coefficients of exogenous variables and lagged dependent variables impose severe restrictions on the observed data that only very small proportion of the sample may be utilized, if they satisfy the conditions at all. In this section, we consider alternative approaches to identify the dynamic dependence–bias reduced estimator for fixed-effects models; bounding parameters without the knowledge of $G(\alpha \mid \mathbf{x})$ and $P(y_0 \mid \mathbf{x})$ for random effects models; and approximate model.

### 7.6.1 Bias-Adjusted Estimator

Controlling the impact of unobserved heterogeneity in linear models are relatively straightforward (e.g., see Chapters 3 and 4). Controlling the impact of

unobserved heterogeneity that is correlated with explanatory variables in nonlinear models is much more difficult. When $T$ is finite, the estimators of the parameters of interest (structural parameters) are inconsistent no matter how large $N$ is. This inconsistency occurs because only a finite number of observations are available to estimate each individual effect $\alpha_i$ while the estimation of structural parameters depends on $\alpha_i$. Increasing $T$ does not necessarily fully solve this problem if $N$ also grows with $T$ (e.g., see Alvarez and Arellano 2003; Hahn and Newey 2004). In this section we consider methods that reduce the bias of the estimator to the order of $\frac{1}{T^2}$.

Let $\boldsymbol{\theta}$ denote the parameters of interest (structural parameters) and $\alpha_i$ denote the unobserved individual-specific effects (incidental parameters). Let $\hat{\boldsymbol{\theta}}_T$ denote the estimator of $\boldsymbol{\theta}$ based on $NT$ panel data $(y_{it}, \mathbf{x}_{it})$ and $\hat{\alpha}_i$, the estimated $\alpha_i$, say the fixed effects MLE (7.3.2) for static logit model ((7.3.2) and 7.3.3)) or dynamic logit model (7.5.16). In general, because of the error in the estimation of $\alpha_i$ when $T$ is fixed, as $N \to \infty$, $\hat{\boldsymbol{\theta}}_T \to \boldsymbol{\theta}_T$, where

$$\boldsymbol{\theta}_T = \boldsymbol{\theta} + \frac{B}{T} + \frac{D}{T^2} + O\left(\frac{1}{T^3}\right) \tag{7.6.1}$$

for some $B$ and $D$. This bias should be small for large $T$. However, if $N$ grows at the same rate as $T$ when $T \to \infty$, the fixed-effects estimator $\hat{\boldsymbol{\theta}}$ is asymptotically biased. For $\frac{N}{T} \to c \neq 0$,

$$\sqrt{NT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \sqrt{NT}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + \sqrt{NT} \cdot \frac{B}{T} + O\left(\sqrt{\frac{N}{T^3}}\right) \tag{7.6.2}$$

will have asymptotic normal distribution centered at $\sqrt{c}B$. (e.g., the fixed-effects estimator for the dynamic panel data model (4.2.3)).

Hahn and Newey (2004) suggest a jackknife estimator to reduce the bias,

$$\tilde{\theta} \equiv T\hat{\boldsymbol{\theta}} - \frac{T-1}{T}\sum_{t=1}^{T} \hat{\boldsymbol{\theta}}(t), \tag{7.6.3}$$

where $\hat{\boldsymbol{\theta}}(t)$ be the fixed effects estimator based on the subsample excluding the observations of the $t$th period. If $\boldsymbol{\theta}_T$ has the form (7.6.1), then the estimator $\tilde{\boldsymbol{\theta}}$ will converge in probability to

$$\begin{aligned} &(T\boldsymbol{\theta}_T - (T-1)\boldsymbol{\theta}_{T-1}) \\ &\quad = \boldsymbol{\theta} + \left(\frac{1}{T} - \frac{1}{T-1}\right) D + O\left(\frac{1}{T^2}\right) \\ &\quad = \boldsymbol{\theta} + O\left(\frac{1}{T^2}\right). \end{aligned} \tag{7.6.4}$$

Thus, the jackknife estimator reduces the bias to the order of $\frac{1}{T^2}$. However, in addition to the fact that the jackknife estimator (7.6.3) requires the estimation of

$(T+1)$ fixed effects estimators, the asymptotic covariance matrix of $\tilde{\boldsymbol{\theta}}$ is complicated to derive unless $(y_{it}, \mathbf{x}_{it})$ are contemporaneously and intertemporally independently distributed (over $i$ and $t$).

An alternative approach is to obtain an estimated $B$, $\hat{B}$, then forming a bias corrected estimator

$$\hat{\boldsymbol{\theta}}^* = \hat{\boldsymbol{\theta}} - \frac{\hat{B}}{T}, \tag{7.6.5}$$

(e.g. (4.7.28)). The advantage of (7.6.5) is that it reduces the bias while the formula for the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}^*$ remains the same as that of $\hat{\boldsymbol{\theta}}$. However, the derivation of $\hat{B}$ can be complicated.

For instance, consider the panel dynamic binary choice model of the form,

$$\begin{aligned} y_{it} &= 1(\mathbf{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma + \alpha_i + u_{it} > 0), \\ i &= 1, \ldots, N, \\ t &= 1, \ldots, T, \\ y_{i0} &\text{ observable,} \end{aligned} \tag{7.6.6}$$

where $1(A) = 1$ if event A occurs and 0 otherwise. We suppose that $u_{it}$ is independently, identically distributed with mean 0. Then

$$\begin{aligned} E(y_{it} \mid y_{i,t-1}, \mathbf{x}_{it}, \alpha_i) &= \text{Prob}\,(y_{it} = 1 \mid y_{i,t-1}, \mathbf{x}_{it}, \alpha_i) \\ &= F(\mathbf{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma + \alpha_i) \\ &= F_{it}, \end{aligned} \tag{7.6.7}$$

where $F$ is the integral of the probability distribution function of $u_{it}$ from $-(\mathbf{x}_{it}\boldsymbol{\beta} + y_{i,t-1}\gamma + \alpha_i)$ to $\infty$. When $\alpha_i$ is considered fixed, the log-likelihood function conditional on $y_{i0}$ takes the form

$$\log L = \sum_{i=1}^{N}\sum_{t=1}^{T}[y_{it} \log F_{it} + (1 - y_{it}) \log (1 - F_{it})] \tag{7.6.8}$$

The MLE of $\boldsymbol{\theta}' = (\boldsymbol{\beta}', \gamma)$ and $\alpha_i$ are obtained by solving the following first-order conditions simultaneously:

$$\frac{\partial \log L}{\partial \alpha_i} \mid_{\hat{\alpha}_i} = 0, \quad i = 1, \ldots, N, \tag{7.6.9}$$

$$\frac{\partial \log L}{\partial \boldsymbol{\theta}} \mid_{\hat{\boldsymbol{\theta}}} = \mathbf{0}. \tag{7.6.10}$$

Substituting the solutions of (7.6.9) as function of $\boldsymbol{\theta}$ to (7.6.8) yields the concentrated log-likelihood function

$$\log L^* = \sum_{i=1}^{N} \ell_i(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta})), \tag{7.6.11}$$

where

$$\ell_i(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta})) = \sum_{t=1}^{T} \Big\{ y_{it} \log F(\mathbf{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma + \hat{\alpha}_i(\boldsymbol{\theta})) + (1 - y_{it}) \log [1 - F(\mathbf{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma + \hat{\alpha}_i(\boldsymbol{\theta}))] \Big\}.$$

Then the MLE of $\boldsymbol{\theta}$ is the solution of the following first-order conditions:

$$\frac{1}{NT} \sum_{i=1}^{N} \left[ \frac{\partial \ell_i(\boldsymbol{\theta}, \hat{\alpha}_i(\theta))}{\partial \boldsymbol{\theta}} + \frac{\partial \ell_i(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta}))}{\partial \hat{\alpha}_i(\boldsymbol{\theta})} \times \frac{\partial \hat{\alpha}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right]_{\hat{\boldsymbol{\theta}}} = \mathbf{0}. \quad (7.6.12)$$

The estimating equation (7.6.12) depends on $\hat{\alpha}_i$. When $T \to \infty$, $\hat{\alpha}_i \to \alpha_i$, the MLE of $\boldsymbol{\theta}$ is consistent. When $T$ is finite, $\hat{\alpha}_i \neq \alpha_i$, then (7.6.12) evaluated at $\boldsymbol{\theta}$ does not converge to 0. Hence the MLE of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$, is not consistent. The bias of the MLE is of order $\frac{1}{T}$. The analytical solution for $\hat{B}_T$ can be derived by taking a Taylor series expansion of (7.6.12) (e.g., see Hahn and Kuersteiner 2011).

Instead of obtaining $\hat{B}_T$ directly, Carro (2007) proposes to derive the bias corrected MLE directly by taking the Taylor series expansion of the score function (7.6.12) around $\alpha_i$ and evaluate it at the true value $\boldsymbol{\theta}$ yields

$$d_{\theta i}(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta})) = d_{\theta i}(\boldsymbol{\theta}, \alpha_i) + d_{\theta\alpha_i i}(\hat{\alpha}_i(\boldsymbol{\theta}) - \alpha_i) + \frac{1}{2} d_{\theta\alpha_i\alpha_i i}(\hat{\alpha}_i(\boldsymbol{\theta}) - \alpha_i)^2 + O_p(T^{-\frac{1}{2}}), \; i = 1, \ldots, N. \quad (7.6.13)$$

where $d_{\theta i}(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta})) = \frac{\partial \ell_i(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} = \frac{\partial \ell_i(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} + \frac{\partial \ell_i(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta}))}{\partial \hat{\alpha}_i(\boldsymbol{\theta}))} \cdot \frac{\partial \hat{\alpha}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$, $d_{\boldsymbol{\theta}\alpha_i i} = \frac{\partial^2 \ell_i(\boldsymbol{\theta}, \hat{\alpha}(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta} \partial \alpha_i}$, $d_{\theta\alpha_i\alpha_i i} = \frac{\partial^3 \ell_i(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta} \partial \alpha_i \partial \alpha_i}$. Making use of McCullah (1987) asymptotic expansion for $(\hat{\alpha}_i(\boldsymbol{\theta}) - \alpha_i)$ and $(\hat{\alpha}_i(\boldsymbol{\theta}) - \alpha_i)^2$, Carro (2007) derives the bias-corrected estimator from the modified score function of $d_{\theta_i} = \frac{\partial \ell_i(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}}$,

$$\sum_{i=1}^{N} d_{\theta_i}^{*} = \sum_{i=1}^{N} \Big\{ d_{\theta_i}(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta})) - \frac{1}{2} \frac{1}{d_{\alpha_i\alpha_i i}(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta}))} \Big( d_{\theta\alpha_i\alpha_i i}(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta})) + d_{\alpha_i\alpha_i\alpha_i}(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta})) \frac{\partial \hat{\alpha}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \Big) + \frac{\partial}{\partial \alpha_i} \Big( \frac{1}{E[d_{\alpha_i\alpha_i i}(\boldsymbol{\theta}, \alpha_i)]} E[d_{\theta\alpha_i i}(\boldsymbol{\theta}, \alpha_i)] \Big) \mid_{\hat{\alpha}_i(\theta)} \Big\}_{\hat{\boldsymbol{\theta}}^{*}} = \mathbf{0}, \quad (7.6.14)$$

where $d_{\alpha_i\alpha_i i} = \frac{\partial^2 \ell_i(\boldsymbol{\theta}, \alpha_i(\boldsymbol{\theta}))}{\partial \alpha_i^2}$ and $d_{\alpha_i\alpha_i\alpha_i} = \frac{\partial^3 \ell_i(\boldsymbol{\theta}, \hat{\alpha}_i(\boldsymbol{\theta}))}{\partial \hat{\alpha}_i^3(\boldsymbol{\theta})}$. Carro (2007) shows that the bias of the modified MLE, $\hat{\boldsymbol{\theta}}^{*}$, is also of order $\left(\frac{1}{T^2}\right)$ and has the same asymptotic variance as the MLE. His Monte Carlo studies show that the bias of the modified MLE is small with $T = 8$.

### 7.6.2 Bounding Parameters

When $y_{i0}$ and $\alpha_i$ are treated as random, the joint likelihood of $f(\mathbf{y}_i, y_{i0} \mid \mathbf{x}_i)$ can be written in the form of conditional density of $f(\mathbf{y}_i \mid y_{i0}, \mathbf{x}_i)$ times the marginal density $f(y_{i0} \mid \mathbf{x}_i)$,

$$\begin{aligned} & f(\mathbf{y}_i, y_{i0} \mid \mathbf{x}_i) \\ & \quad = \int f(\mathbf{y}_i \mid y_{i0}, \mathbf{x}_i, \alpha_i) f(\mathbf{y}_{i0} \mid \mathbf{x}_i, \alpha_i) G(\alpha_i \mid \mathbf{x}_i) d\alpha_i, \\ & i = 1, \ldots, N, \end{aligned} \tag{7.6.15}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$, $\mathbf{x}_i = (\mathbf{x}_{i1}', \ldots, \mathbf{x}_{iT}')$, and $G(\alpha_i \mid \mathbf{x}_i)$ denotes the conditional density of $\alpha_i$ given $\mathbf{x}_i$. For model (7.6.6) with $u_{it}$ following a standard normal distribution, $N(0, 1)$,

$$\begin{aligned} f(\mathbf{y}_i \mid y_{i0}, \mathbf{x}_i, \alpha_i) = \prod_{t=1}^{T} & [\Phi(\mathbf{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma + \alpha_i)]^{y_{it}} \\ & \cdot [1 - \Phi(\mathbf{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma + \alpha_i)]^{1-y_{it}} \\ & i = 1, \ldots, N. \end{aligned} \tag{7.6.16}$$

If $u_{it}$ follows a logistic distribution

$$\begin{aligned} f(\mathbf{y}_i \mid y_{i0}, \mathbf{x}_i, \alpha_i) = \prod_{t=1}^{T} & \frac{\exp\left[(\mathbf{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma + \alpha_i)\right]^{y_{it}}}{1 + \exp\left(\mathbf{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma + \alpha_i\right)}, \\ & i = 1, \ldots, N. \end{aligned} \tag{7.6.17}$$

When $G(\alpha \mid \mathbf{x})$ and the initial distribution $P(y_0 \mid \mathbf{x})$ are known, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \gamma)$ can be estimated by the MLE. However, $G(\alpha \mid \mathbf{x})$ and $f(y_0 \mid \mathbf{x})$ are usually unknown. Although in principle, one can still maximize (7.6.15), the usual regularity conditions for the consistency of the MLE (e.g., Kiefer and Wolfowitz 1956) is violated because $G(\alpha \mid \mathbf{x})$ is infinite dimensional (Cosslett 1981).

If $\mathbf{x}$ and $\alpha$ are discrete, Honoré and Tamer (2006) suggest a linear program approach to provide the bound of $\boldsymbol{\theta}$. Let $A_j = (d_{j1}, \ldots, d_{jT})$ be the $1 \times T$ sequence of binary variables $d_{jt}$. Let $\mathcal{A}$ denote the set of all $2^T$ possible sequence of 0's and 1's, $A_j$. Let $P(y_{i0} \mid \mathbf{x}_i, \alpha_i)$ denote the probability of $y_{i0} = 1$ given $\mathbf{x}_i$ and $\alpha_i$ and $f_0(\alpha, \mathbf{x})$ denote the distribution of $y_{i0}$ given $\mathbf{x}$ and $\alpha$. Then, conditional on $P(y_{i0} \mid \mathbf{x}_i, \alpha_i)$,

$$\begin{aligned} f(\mathbf{y}_i \mid \mathbf{x}_i, f_0(y_{i0} \mid \mathbf{x}_i, \boldsymbol{\alpha}_i), \alpha_i) = {} & P(y_{i0} \mid \mathbf{x}_i, \alpha_i) f(\mathbf{y}_i \mid y_{i0} = 1, \mathbf{x}_i, \alpha_i) \\ & + (1 - P(y_{i0} \mid \mathbf{x}_i, \alpha_i)) f(\mathbf{y}_i \mid y_{i0} = 0, \mathbf{x}_i, \alpha_i), \end{aligned} \tag{7.6.18}$$

and

$$f(\mathbf{y}_i \mid \mathbf{x}_i, f_0(\cdot,\cdot), \boldsymbol{\theta}) = \int f(\mathbf{y}_i \mid \mathbf{x}_i, \alpha, f_0(\cdot,\cdot), \boldsymbol{\theta}) dG(\alpha \mid \mathbf{x}_i). \tag{7.6.19}$$

Let $\pi(A \mid \mathbf{x}, \alpha; f_0(\cdot,\cdot), \boldsymbol{\theta})$ and $P(A \mid \mathbf{x})$ be the probability of an event $A$ in $\mathcal{A}$ given $(\mathbf{x}, \alpha)$ predicted by the model and the probability of an event $A$ occurs given $\mathbf{x}$, respectively. Then $\pi(A \mid \mathbf{x}, f_0(\cdot,\cdot), \boldsymbol{\theta}) = \int \pi(A \mid \mathbf{x}, \alpha; f_0(\cdot,\cdot), \boldsymbol{\theta}) dG(\alpha \mid \mathbf{x})$. Define the set of $(f_0(\cdot,\cdot), \boldsymbol{\theta})$ that is consistent with a particular data-generating process with probabilities $\mathcal{P}(\mathcal{A} \mid \mathbf{x})$ as

$$\Psi = \Big\{ (f_0(\cdot,\cdot), \boldsymbol{\theta}) : P[\pi(\mathcal{A} \mid \mathbf{x}; f_0(\cdot,\cdot), \boldsymbol{\theta}) = P(\mathcal{A} \mid \mathbf{x})] = 1 \Big\}. \tag{7.6.20}$$

Then the bound of $\boldsymbol{\theta}$ is given by

$$\ominus = \Big\{ \boldsymbol{\theta} : \exists f_0(\cdot,\cdot) \text{ such that } P[\pi(\mathcal{A} \mid \mathbf{x}; f_0(\cdot,\cdot), \boldsymbol{\theta}) = P(\mathcal{A} \mid \mathbf{x})] = 1) \Big\}. \tag{7.6.21}$$

Suppose $\alpha$ has a discrete distribution with a known maximum number of points of support, $M$. The points of support are denoted by $a_m$ and the probability of $\alpha_i = a_m$ given $\mathbf{x}$ denoted by $\rho_{mx}$. Then

$$\begin{aligned} \pi(\mathcal{A} \mid f_0(\cdot,\cdot), \mathbf{x}, \boldsymbol{\theta}) &= \sum_{m=1}^{M} \rho_{mx} \Big[ f_0(a_m, \mathbf{x}) \pi(\mathcal{A} \mid y_0 = 1, \boldsymbol{\theta}, \mathbf{x}; a_m) \\ &\quad + (1 - f_0(a_m, \mathbf{x})) \pi(\mathcal{A} \mid y_0 = 0, \boldsymbol{\theta}, \mathbf{x}; a_m) \Big] \\ &= \sum_{m=1}^{M} z_{mx} \pi(\mathcal{A} \mid y_0 = 1, \mathbf{x}, \boldsymbol{\theta}; a_m) + \sum_{m=1}^{M} z_{M+m,x} \pi(\mathcal{A} \mid y_0 = 0, \mathbf{x}, \boldsymbol{\theta}; a_m), \end{aligned} \tag{7.6.22}$$

where $z_{mx} = \rho_{mx} f_0(a_m, x)$ and $z_{M+m,x} = \rho_{mx}[1 - f_0(a_m, \mathbf{x})]$ for $m = 1, \ldots, M$. The identified set $\ominus$, consists of the value of $\boldsymbol{\theta}$ for which the following equations have a solution for $\{z_{mx}\}_{m=1}^{2M}$:

$$\sum_{m=1}^{M} z_{mx} \pi(A \mid y_0 = 1, \mathbf{x}, \boldsymbol{\theta}; a_m) + \sum_{m=1}^{M} z_{M+m,x} \pi(A \mid y_0 = 0, \mathbf{x}, \boldsymbol{\theta}; a_m) = P(A \mid \mathbf{x}), \tag{7.6.23}$$

and for all $A \in \mathcal{A}$,

$$\sum_{m=1}^{2M} z_{mx} = 1, z_{mx} \geq 0. \tag{7.6.24}$$

Equation's (7.6.23) and (7.6.24) have exactly the same structure as the constraints in a linear programming problem, so checking whether a particular $\boldsymbol{\theta}$ belongs to $\ominus$ can be done in the same way that checks for a feasible solution in a linear programming problem provided $P(A \mid \mathbf{x})$ can be consistently estimated. Therefore, Honoré and Tamer (2006) suggest bounding $\boldsymbol{\theta}$ by considering the linear programming problem:

$$\begin{aligned} &\text{maximize} \quad \sum_j -v_{jx} \\ &\{z_{mx}, \{v_{jx}\}\} \end{aligned} \tag{7.6.25}$$

where

$$v_{jx} = P(A_j \mid \mathbf{x}) - \sum_{m=1}^{M} z_{mx} \pi(A_j \mid y_{i0} = 1, \mathbf{x}, \boldsymbol{\theta}; a_m)$$

$$- \sum_{m=1}^{M} z_{M+m,x} \pi(A_j \mid y_{i0} = 0, \mathbf{x}, \boldsymbol{\theta}; a_m)$$

$$\text{for all } A_j \in \mathcal{A}, \ j = 1, \ldots, 2^T, \tag{7.6.26}$$

$$1 - \sum_{m=1}^{2M} z_{mx} = v_{0x}, \tag{7.6.27}$$

$$z_{mx} \geq 0, \tag{7.6.28}$$

$$v_{jx} \geq 0. \tag{7.6.29}$$

The optimal function value for (7.6.25) is 0 if and only if all $v_{jx} = 0$, that is, if a solution exists to (7.6.23) and (7.6.24). If (7.6.23) and (7.6.24) do not have a solution, the maximum function value in (7.6.25) is negative. Following Manski and Tamer (2002) it can be shown that a consistent estimator of the identified region can be constructed by checking whether, for a given $\boldsymbol{\theta}$, the sample objective function is within $\epsilon$ of the maximum value of 0 where $P(A)$ is substituted by its consistent estimator. Because $\mathbf{x}$ is discrete, one can mimic this argument for each value in the support of $\mathbf{x}_i$ which will then contribute a set of constraints to the linear programming problem.

### 7.6.3 Approximate Model

The dynamic logit model (7.5.24) implies that the conditional distribution of a sequence of response variables, $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$ given $\alpha_i$, $\mathbf{x}_i$, and $y_{i0}$, can

be expressed as

$$P(\mathbf{y}_i \mid \mathbf{x}_i, \alpha_i, y_{i0}) = \frac{\exp\left(\alpha_i \sum_{t=1}^{T} y_{it} + \sum_{t=1}^{T} y_{it}(\mathbf{x}_{it}'\boldsymbol{\beta}) + y_{i*}\gamma\right)}{\sum_{t=1}^{T}[1 + \exp\left(\mathbf{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma + \alpha_i\right)]}, i = 1, \ldots, N, \tag{7.6.30}$$

where $y_{i*} = \sum_{t=1}^{T} y_{i,t-1}y_{it}$.

The Honoré and Kyriazidou (2000a) conditional approach discussed in Section 7.5 requires a specification of a suitable kernel function and the bandwidth parameters to weigh the response configuration of each subject in the sample on the basis of exogenous explanatory variables in which only exogenous variables are close to each other receiving large weights, implying a substantial reduction of the rate of convergence of the estimator to the true parameter value. Moreover, conditional on certain configurations leads to a response function in terms of time changes of the covariates, implying the exclusion of time-invariant variables. Noting that the dynamic logit model (7.6.30) implies that the conditional log-odds ratio between $(y_{it}, y_{i,t-1})$ equals to

$$\log \frac{P(y_{it} = 0 \mid \alpha_i, \mathbf{x}_i, y_{i,t-1} = 0) \cdot P(y_{it} = 1 \mid \alpha_i, \mathbf{x}_i, y_{i,t-1} = 1)}{P(y_{it} = 0 \mid \alpha_i, \mathbf{x}_i, y_{i,t-1} = 1) \cdot P(y_{it} = 1 \mid \alpha_i, \mathbf{x}_i, y_{i,t-1} = 0)} = \gamma, \tag{7.6.31}$$

Bartolucci and Nigro (2010) suggest using the Cox (1972) quadratic exponential model to approximate (7.6.30),[21]

$$P^*(\mathbf{y}_i \mid \mathbf{x}_i, y_{i0}, \delta_i) = \frac{\exp\left[\delta_i(\sum_{t=1}^{T} y_{it}) + \sum_{t=1}^{T} y_{it}(\mathbf{x}_{it}'\boldsymbol{\phi}_1) + y_{iT}(\psi + \mathbf{x}_{iT}'\boldsymbol{\phi}_2) + y_{i*}\tau\right]}{\sum_{d_i} \exp[\delta_i(\sum_{t=1}^{T} d_{ijt}) + \sum_{t=1}^{T} d_{ijt}(\mathbf{x}_{it}'\boldsymbol{\phi}_1) + d_{ijT}(\psi + \mathbf{x}_{iT}'\boldsymbol{\phi}_2) + d_{ij}^*\tau]} \tag{7.6.32}$$

where $\mathbf{d}_{ij} = (d_{ij1}, \ldots, d_{ijT})$ denote the $j$th possible binary response sequence, $\sum_{d_i}$ denotes the sum over all possible response sequence of $\mathbf{d}_{ij}$, such that $\sum_{t=1}^{T} d_{ijt} = \sum_{t=1}^{T} y_{it}$, and $d_{ij}^* = d_{ij1}y_{io} + \sum_{t=1}^{T} d_{ijt}d_{ij,t-1}$. Model (7.6.32) implies that

$$P^*(y_{it} \mid \mathbf{x}_i, \delta_i, y_{i0}, \ldots, y_{i,t-1}) = \frac{\exp\{y_{it}[\delta_i + \mathbf{x}_{it}'\boldsymbol{\phi}_1 + y_{i,t-1}\tau + e_t^*(\delta_i, \mathbf{x}_i)]\}}{1 + \exp\left[\delta_i + \mathbf{x}_{it}'\boldsymbol{\phi}_1 + y_{i,t-1}\tau + e_t^*(\delta_i, \mathbf{x}_i)\right]}, \tag{7.6.33}$$

[21] To differentiate the approximate model from the dynamic logit model (7.6.30), we use $\delta_i$ to represent the individual-specific effects and $\boldsymbol{\phi}_1$ to represent the coefficients of $\mathbf{x}_{it}$ in the approximate model (7.6.32).

where, for $t < T$,

$$\begin{aligned} e_t^*(\delta_i, \mathbf{x}_i) &= \log \frac{1 + \exp\left[\delta_i + \mathbf{x}'_{i,t+1}\boldsymbol{\phi}_1 + e_{t+1}^*(\delta_i, \mathbf{x}_i) + \tau\right]}{1 + \exp\left[\delta_i + \mathbf{x}'_{i,t+1}\boldsymbol{\phi}_1 + e_{t+1}^*(\delta_i, \mathbf{x}_i)\right]} \\ &= \log \frac{P(y_{i,t+1} = 0 \mid \delta_i, \mathbf{x}_i, y_{it} = 0)}{P(y_{i,t+1} = 0 \mid \delta_i, \mathbf{x}_i, y_{it} = 1)}. \end{aligned} \tag{7.6.34}$$

The corrections term (7.6.34) depends on future covariates. For the last period, it is approximated by

$$e_T^*(\delta_i, \mathbf{x}_i) = \psi + \mathbf{x}'_{iT}\boldsymbol{\phi}_2. \tag{7.6.35}$$

Model (7.6.33) may be viewed as a latent response model of the form

$$y_{it}^* = \mathbf{x}'_{it}\boldsymbol{\phi}_1 + \delta_i + y_{i,t-1}\tau + e_t^*(\delta_i, \mathbf{x}_i) + \eta_{it}, \tag{7.6.36}$$

with logistically distributed stochastic term $\eta_{it}$. The correction term $e_t^*(\delta_i, \mathbf{x}_i)$ may be interpreted as a measure of the effect of the present choice $y_{it}$ on the expected utility (or propensity) at period $(t + 1)$. The parameter $\tau$ for the state dependence is the log-odds ratio between any pairs of variables $(y_{i,t-1}, y_{it})$, conditional on all the other response variables or marginal with respect to these variables.

The difference between the approximate model (7.6.32) and the dynamic logit model (7.6.30) is in the denominator. The former does not depend on the actual sequence $\mathbf{y}_i$, while the latter does. The advantage of model (7.6.32) or (7.6.33) is that the parameters for the unobserved heterogeneity, $\delta_i$, can be eliminated by conditioning on the sum of response variables over time just like the static logit model (7.3.14). When $y_{i0}$ are observable, the structural parameters can be estimated by the conditional maximum likelihood estimator as discussed in (7.3.21) when $T \geq 2$.

The relations between (7.6.32) and (7.6.30) can be seen through a Taylor series expansion of the nonlinear term of the logarithm of the dynamic logit model (7.6.30) at $\alpha_i = \tilde{\alpha}_i$, $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$ and $\gamma = 0$,

$$\begin{aligned} &\sum_{t=1}^{T} \log\left[1 + \exp\left(\mathbf{x}'_{it}\boldsymbol{\beta} + y_{i,t-1}\gamma + \alpha_i\right)\right] \\ &\quad \simeq \sum_{t=1}^{T} \{ \log\left[1 + \exp\left(\mathbf{x}'_{it}\tilde{\boldsymbol{\beta}} + \tilde{\alpha}_i\right)\right] + \tilde{q}_{it} \\ &\qquad \cdot \left[\mathbf{x}'_{it}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) + (\alpha_i - \tilde{\alpha}_i)\right]\} + \tilde{q}_{i1} y_{i0}\gamma + \sum_{t=1}^{T} \tilde{q}_{it} y_{i,t-1}\gamma, \end{aligned} \tag{7.6.37}$$

where

$$\tilde{q}_{it} = \frac{\exp(\tilde{\alpha}_i + \mathbf{x}_{it}'\tilde{\boldsymbol{\beta}})}{1 + \exp(\tilde{\alpha}_i + \mathbf{x}_{it}'\tilde{\boldsymbol{\beta}})}. \tag{7.6.38}$$

Substituting (7.6.37) into the logarithm of (7.6.30) and renormalizing the exponential of the resulting expression yields the approximate model for (7.6.30) as

$$\begin{aligned} &P^*(\mathbf{y}_i \mid \mathbf{x}_i, \alpha_i, y_{i0}) \\ &\quad = \frac{\exp(\alpha_i(\sum_{t=1}^T y_{it}) + \sum_{t=1}^T y_{it}(\mathbf{x}_{it}'\boldsymbol{\beta}) - \sum_{t=2}^T \tilde{q}_{it} y_{i,t-1}\gamma + y_{i*}\gamma)}{\sum_{d_i} \exp[\alpha_i(\sum_{t=1}^T d_{ijt}) + \sum_{t=1}^T d_{ijt}(\mathbf{x}_{it}'\boldsymbol{\beta}) - (\sum_{t=2}^T \tilde{q}_{it} d_{i,t-1})\gamma + d_{ij*}\gamma)]}, \\ &\quad i = 1, \ldots, N. \end{aligned} \tag{7.6.39}$$

When $\gamma$ is indeed equal to 0, the true model and the approximating model coincide. Both become the static logit model (7.3.13). The approximating model (7.6.32) or (7.6.39) implies that the conditional logit of $y_{it}$ given $\mathbf{x}_i'$, $\alpha_i$ and $y_{i0}, \ldots, y_{i,t-1}$,is equal to

$$\begin{aligned} &\log \frac{P^*(y_{it} = 1 \mid \mathbf{x}_i, \alpha_i, y_{i0}, \ldots, y_{i,t-1})}{P^*(y_{it} = 0 \mid \mathbf{x}_i, \alpha_i, y_{i0}, \ldots, y_{i,t-1})} \\ &\quad = \begin{cases} \mathbf{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma + \alpha_i + e_t(\alpha_i, \mathbf{x}_i) - \tilde{q}_{i,t+1}\gamma, & \text{if } t < T, \\ \mathbf{x}_{it}'\boldsymbol{\beta} + y_{i,t-1}\gamma + \alpha_i, & \text{if } t = T, \end{cases} \end{aligned} \tag{7.6.40}$$

where

$$\begin{aligned} e_t(\alpha_i, \mathbf{x}_i) &= \log \frac{P^*(y_{i,t+1} = 0 \mid \mathbf{x}_i, \alpha_i, y_{it} = 0)}{P^*(y_{i,t+1} = 0 \mid \mathbf{x}_i, \alpha_i, y_{it} = 1)} \\ &= \bar{q}_{i,t+1}\gamma. \end{aligned} \tag{7.6.41}$$

Equation (7.6.40) implies that

$$\log \frac{P^*(y_{it} = 1 \mid \mathbf{x}_i, \alpha_i, y_{i,t-1} = 1)}{P^*(y_{it} = 0 \mid \mathbf{x}_i, \alpha_i, y_{i,t-1} = 1)} - \log \frac{P^*(y_{it} = 1 \mid \mathbf{x}_i\, a_i, y_{i,t-1} = 0)}{P^*(y_{it} = 0 \mid \mathbf{x}_i, \alpha_i, y_{i,t-1} = 0)} = \gamma. \tag{7.6.42}$$

Just like the dynamic logit model, the approximating model implies $y_{it}$ is conditionally independent of $y_{i0}, \ldots, y_{i,t-2}$ given $\mathbf{x}_i, \alpha_i$ and $y_{i,t-1}$ for $t = 2, \ldots, T$ and is conditionally independent of $y_{i0}, \ldots, y_{i,t-2}, y_{i,t+2}, \ldots, y_{iT}$ given $\mathbf{x}_i, \alpha_i, y_{i,t-1}$ and $y_{i,t+1}$ for $t = 2, \ldots, T-1$. However, it has the advantage that the minimum sufficient statistics for $\alpha_i$ is now $\sum_{t=1}^T y_{it}$. Hence the

conditional distribution of $\mathbf{y}_i$ given $\sum_{t=1}^{T} y_{it}$, where $0 < \sum_{t=1}^{T} y_{it} < T$,

$$P^*(\mathbf{y}_i \mid \mathbf{x}_i, \alpha_i, y_{i0}, \sum_{t=1}^{T} y_{it})$$

$$= \frac{\exp\left\{\sum_{t=2}^{T} y_{it}(\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta} - \sum_{t=2}^{T} \tilde{q}_{it} y_{i,t-1}\gamma + y_{i*}\gamma)\right\}}{\sum_{d_i} \exp\{\sum_{t=2}^{T} d_{ijt}(\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta} - \sum_{t=2}^{T} \tilde{q}_{it} d_{ij,t-1}\gamma + d_{ij*}\gamma)},$$

$$i = 1, \ldots, N. \tag{7.6.43}$$

To obtain the pseudo-conditional MLE of the pseudo-likelihood function (7.6.43). Bartolucci and Nigro (2010) suggest first assuming there was no state dependence ($\gamma = 0$) and maximizing the conditional log-likelihood of the static logit model (7.3.21) for those $i$ where $0 < \sum_{t=1}^{T} y_{it} < T$ to obtain a preliminary estimate $\bar{\boldsymbol{\beta}}$. Then substituting $\bar{\boldsymbol{\beta}}$ into (7.6.42) to obtain the revised pseudo-conditional MLE of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ through the Newton–Raphson iterative procedure. Their Monte Carlo studies show that the pseudo-conditional MLE has a very low bias for data generated by a dynamic logit model.

CHAPTER 8

# Sample Truncation and Sample Selection

## 8.1 INTRODUCTION

In economics, the ranges of dependent variables are often constrained in some way. For instance, in his pioneering work on household expenditure on durable goods, Tobin (1958) used a regression model that specifically took account of the fact that the expenditure (the dependent variable of his regression model) cannot be negative. Tobin called this type of model the model of limited dependent variables. It and its various generalizations are known as Tobit models because of their similarities to probit models.[1] In statistics they are known as truncated or censored regression models. The model is called truncated if the observations outside a specific range are totally lost, while it is called censored if we can at least observe the proportion of samples having realized values falling outside the observed range and some of the explanatory variables.

It is more convenient to relate an observed sample $y$ that is subject to truncation or selection with a latent response function,

$$y^* = \mathbf{x}'\boldsymbol{\beta} + u, \tag{8.1.1}$$

where $\mathbf{x}$ is a $K \times 1$ vector of exogenous variables and $u$ is the error term that is independently, identically distributed (i.i.d) with mean 0 and variance $\sigma_u^2$. Without loss of generality, suppose that the observed $y$ are related to the latent variable $y^*$ by

$$y = \begin{cases} y^*, & \text{if } y^* > 0, \\ 0, & \text{if } y^* \leq 0. \end{cases} \tag{8.1.2}$$

Models of the form (8.1.1) and (8.1.2) are called censored regression models because the data consist of those points of $(y_i^*, \mathbf{x}_i)$ if $y_i^* > 0$ and $(0, \mathbf{x}_i)$ if $y_i^* \leq 0$ for $i = 1, \ldots, N$. The truncated data consist only of points of $(y_i^*, \mathbf{x}_i)$ where $y_i^* > 0$.

[1] See Amemiya (1985) and Maddala (1983) for extensive discussions of various types of Tobit models.

The conditional expectation of $y$ given $\mathbf{x}$ for truncated data is equal to

$$E(y \mid \mathbf{x}, y > 0) = E(y^* \mid \mathbf{x}, y^* > 0) = \mathbf{x}'\boldsymbol{\beta} + E(u \mid u > -\mathbf{x}'\boldsymbol{\beta}). \tag{8.1.3}$$

The conditional expectation of $y$ given $\mathbf{x}$ for censored data is equal to

$$\begin{aligned} E(y \mid \mathbf{x}) &= \text{Prob}\,(y = 0) \cdot 0 \; + \; \text{Prob}\,(y > 0 \mid \mathbf{x}) \cdot E(y \mid y > 0, \mathbf{x}) \\ &= \text{Prob}\,(u \leq -\mathbf{x}'\boldsymbol{\beta}) \cdot 0 \\ &\quad + \text{Prob}\,(u > -\mathbf{x}'\boldsymbol{\beta}) E(y^* \mid \mathbf{x}; u > -\mathbf{x}'\boldsymbol{\beta}) \\ &= \text{Prob}\,(u > -\mathbf{x}'\boldsymbol{\beta})[\mathbf{x}'\boldsymbol{\beta} + E(u \mid u > -\mathbf{x}'\boldsymbol{\beta})]. \end{aligned} \tag{8.1.4}$$

If $u$ is independently normally distributed with mean 0 and variance $\sigma_u^2$, then

$$\text{Prob}\,(u > -\mathbf{x}'\boldsymbol{\beta}) = 1 - \Phi\left(\frac{-\mathbf{x}'\boldsymbol{\beta}}{\sigma_u}\right) = \Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma_u}\right), \tag{8.1.5}$$

and

$$E(u \mid u > -\mathbf{x}'\boldsymbol{\beta}) = \sigma_u \cdot \frac{\phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma_u}\right)}{\Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma_u}\right)}, \tag{8.1.6}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are standard normal density and cumulative (or integrated) normal, respectively. Equations (8.1.3) and (8.1.5) show that truncation or censoring of the dependent variables introduces dependence between the error term and the regressors for the model

$$y = \mathbf{x}'\boldsymbol{\beta} + \epsilon, \tag{8.1.7}$$

where the error

$$\epsilon = \nu + E(y \mid \mathbf{x}) - \mathbf{x}'\boldsymbol{\beta}. \tag{8.1.8}$$

Although $\nu = y - E(y \mid \mathbf{x})$ has $E(\nu \mid \mathbf{x}) = 0$, but $E(\epsilon \mid \mathbf{x}) \neq 0$. Therefore, the least squares estimator of (8.1.7) is biased and inconsistent.

The likelihood function of the truncated data is equal to

$$L_1 = \prod_1 [\text{Prob}\,(y_i > 0 \mid \mathbf{x}_i)]^{-1} f(y_i) \tag{8.1.9}$$

where $f(\cdot)$ denotes the density of $y_i^*$ (or $u_i$) and $\prod_1$ means the product over those $i$ for which $y_i > 0$. The likelihood function of the censored data is

equal to

$$
\begin{aligned}
L_2 &= \left\{ \prod_0 \text{Prob}\,(y_i = 0 \mid \mathbf{x}_i) \cdot \prod_1 \text{Prob}\,(y_i > 0 \mid \mathbf{x}_i) \right\} \\
&\quad \cdot \left\{ \prod_1 [\text{Prob}\,(y_i > 0 \mid \mathbf{x}_i)]^{-1} f(y_i) \right\} \\
&= \prod_0 \text{Prob}\,(y_i = 0 \mid \mathbf{x}_i) \prod_1 f(y_i),
\end{aligned} \tag{8.1.10}
$$

where $\prod_0$ means the product over those $i$ for which $y_i^* \le 0$. In the case that $u_i$ is independently normally distributed with mean 0 and variance $\sigma_u^2$, $f(y_i) = (2\pi)^{-\frac{1}{2}} \sigma_u^{-1} \exp\{-\frac{1}{2\sigma_u^2}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2\}$ and $\text{Prob}\,(y_i = 0 \mid \mathbf{x}_i) = \Phi\big(\frac{-\mathbf{x}_i'\boldsymbol{\beta}}{\sigma_u}\big) = 1 - \Phi\big(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma_u}\big)$.

Maximizing (8.1.9) or (8.1.10) with respect to $\boldsymbol{\theta}' = (\boldsymbol{\beta}', \sigma_u^2)$ yields the maximum likelihood estimator (MLE). The MLE, $\hat{\boldsymbol{\theta}}$, is consistent and is asymptotically normally distributed. The asymptotic covariance matrix of the MLE, asy cov $[\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})]$, is equal to the inverse of the information matrix $\big[ - E \frac{1}{N} \frac{\partial^2 \log L_j}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \big]^{-1}$, which may be approximated by $\big[ - \frac{1}{N} \frac{\partial^2 \log L_j}{\partial \theta \partial \theta'} \big|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} \big]^{-1}$, $j = 1, 2$. However, the MLE is highly nonlinear. A Newton–Raphson type iterative scheme may have to be used to obtain the MLE. Alternatively, if $u$ is normally distributed, Heckman (1976a) suggests the following two-step estimator:

1. Maximize the first curly part of the likelihood function (8.1.10) by probit MLE with respect to $\boldsymbol{\delta} = \frac{1}{\sigma_u}\boldsymbol{\beta}$, yielding $\hat{\boldsymbol{\delta}}$.
2. Substitute $\hat{\boldsymbol{\delta}}$ for $\boldsymbol{\delta}$ into the truncated model

$$
\begin{aligned}
y_i &= E(y_i \mid \mathbf{x}_i; y_i > 0) + \eta_i \\
&= \mathbf{x}_i'\boldsymbol{\beta} + \sigma_u \frac{\phi(\mathbf{x}_i'\boldsymbol{\delta})}{\Phi(\mathbf{x}_i'\boldsymbol{\delta})} + \eta_i, \quad \text{for those } i \text{ such that } y_i > 0,
\end{aligned} \tag{8.1.11}
$$

where $E(\eta_i \mid \mathbf{x}_i) = 0$ and Var $(\eta_i \mid \mathbf{x}_i) = \sigma_u^2[1 - (\mathbf{x}_i'\boldsymbol{\delta})\lambda_i - \lambda_i^2]$ and $\lambda_i = \frac{\phi(\mathbf{x}_i'\boldsymbol{\delta})}{\Phi(\mathbf{x}_i'\boldsymbol{\delta})}$. Regress $y_i$ on $\mathbf{x}_i$ and $\frac{\phi(\mathbf{x}_i'\hat{\boldsymbol{\delta}})}{\Phi(\mathbf{x}_i'\hat{\boldsymbol{\delta}})}$ by least squares, using only the positive observations of $y_i$.

The Heckman two-step estimator is consistent. The formula for computing the asymptotic variance–covariance matrix of Heckman's estimator is given by Amemiya (1978b). But the Heckman two-step estimator is not as efficient as the MLE.

Both the MLE of (8.1.10) and the Heckman two-step estimator (8.1.11) are consistent only if $u$ is independently normally distributed with constant

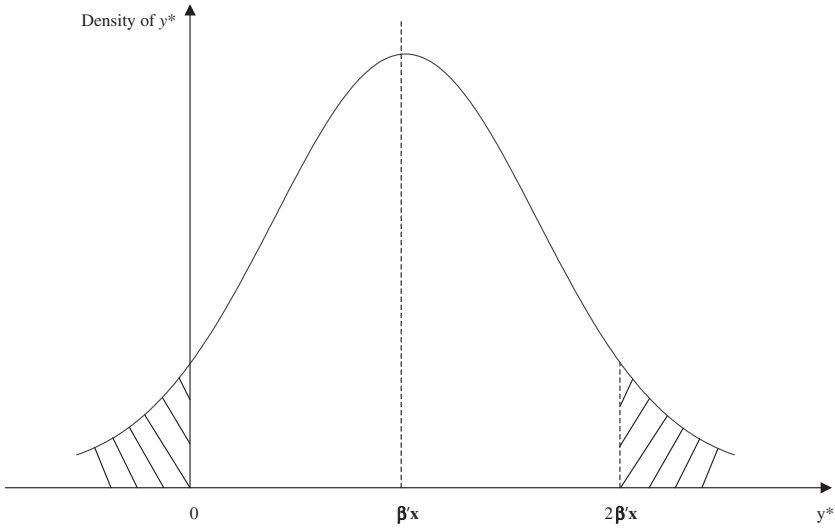


Figure 8.1. Density of $y^*$ censored or truncated at 0.

variance. Of course, the idea of the MLE and the Heckman two-step estimator can still be implemented with proper modification if the identically distributed density function of $u$ is correctly specified. A lot of times an investigator does not have the knowledge of the density function of $u$ or $u$ is not identically distributed. Under the assumption that it is symmetrically distributed around 0, Powell (1986) proves that by applying the least-squares method to the symmetrically censored or truncated data yields a consistent estimator that is robust to the assumption of the probability density function of $u$ and heteroscedasticity of the unknown form.

The problem of censoring or truncation is that conditional on $\mathbf{x}$, $y$ is no longer symmetrically distributed around $\mathbf{x}'\boldsymbol{\beta}$ even though $u$ is symmetrically distributed around 0. Consider the case where $\mathbf{x}_i'\boldsymbol{\beta} > 0$ and $y_i = y_i^*$ if $y_i^* > 0$. Data points for which $u_i \leq -\mathbf{x}_i'\boldsymbol{\beta}$ are either censored or omitted. However, we can restore symmetry by censoring or throwing away observations with $u_i \geq \mathbf{x}_i'\boldsymbol{\beta}$ or $y_i \geq 2\mathbf{x}_i'\boldsymbol{\beta}$ as shown in Figure 8.1 so that the remaining observations fall between $(0, 2\mathbf{x}'\boldsymbol{\beta})$. Because of the symmetry of $u$, the corresponding dependent variables are again symmetrically distributed about $\mathbf{x}'\boldsymbol{\beta}$ (Hsiao 1976). However, any observations correspond to $\mathbf{x}'\boldsymbol{\beta} < 0$ are all lying on one side of $\mathbf{x}'\boldsymbol{\beta}$. There are no corresponding observations lying on the other side of $\mathbf{x}'\boldsymbol{\beta}$, they have to be thrown away.

To make this approach more explicit, consider first the case in which the dependent variable is truncated at 0. In such a truncated sample, data points for which $u_i \leq -\mathbf{x}_i'\boldsymbol{\beta}$ when $\mathbf{x}_i'\boldsymbol{\beta} > 0$ are omitted. But if data points with $u_i \geq \mathbf{x}_i'\boldsymbol{\beta}$ are also excluded from the sample, then any remaining observations would have error terms lying within the interval $(-\mathbf{x}_i'\boldsymbol{\beta}, \mathbf{x}_i'\boldsymbol{\beta})$ (any observations for

which $\mathbf{x}_i'\boldsymbol{\beta} \leq 0$ are automatically deleted). Because of the symmetry of the distribution of $u$, the residuals for the "symmetrically truncated" sample will also be symmetrically distributed about 0. The corresponding dependent variable would take values between 0 and $2\mathbf{x}_i'\boldsymbol{\beta}$ as shown in the region AOB of Figure 8.2. In other words, points $b$ and $c$ in Figure 8.2 are thrown away (point $a$ is not observed). Therefore, the moment conditions

$$E[1(y < 2\mathbf{x}'\boldsymbol{\beta})(y - \mathbf{x}'\boldsymbol{\beta}) \mid \mathbf{x}] = 0, \tag{8.1.12}$$

and

$$E[1(y < 2\mathbf{x}'\boldsymbol{\beta})(y - \mathbf{x}'\boldsymbol{\beta})\mathbf{x}] = \mathbf{0}, \tag{8.1.13}$$

hold, where $1(A)$ denotes the indicator function that takes the value 1 if $A$ occurs and 0 otherwise.

The sample analog of (8.1.13) is

$$\frac{1}{N}\sum_{i=1}^{N} 1(y_i < 2\mathbf{x}_i'\hat{\boldsymbol{\beta}})(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})\mathbf{x}_i = 0 \tag{8.1.14}$$

which is the first-order condition of applying the least-squares principle to symmetrically trimmed truncated data falling in the region AOB.

Definition of the symmetrically trimmed estimator for a censored sample is similarly motivated. The error terms of the censored regression model are of the form $u_i^* = \max\{u_i, -\mathbf{x}_i'\boldsymbol{\beta}\}$, (i.e., point $a$ in Figure 8.2 is moved to the corresponding circled point $a'$). "Symmetric censoring" would replace $u_i^*$ with $\min\{u_i^*, \mathbf{x}_i'\boldsymbol{\beta}\}$ whenever $\mathbf{x}_i'\boldsymbol{\beta} > 0$, and would delete the observation otherwise. In other words, the dependent variable $y_i = \max\{0, y_i^*\}$ is replaced with $\min\{y_i, 2\mathbf{x}_i'\boldsymbol{\beta}\}$ as the points $a, b, c$ in Figure 8.2 have been moved to the corresponding circled points $(a', b', c')$. Therefore,

$$E\{1(\mathbf{x}'\boldsymbol{\beta} > 0)[\min(y, 2\mathbf{x}'\boldsymbol{\beta}) - \mathbf{x}'\boldsymbol{\beta}] \mid \mathbf{x}\} = 0, \tag{8.1.15}$$

and

$$E\{1(\mathbf{x}'\boldsymbol{\beta} > 0)[\min(y, 2\mathbf{x}'\boldsymbol{\beta}) - \mathbf{x}'\boldsymbol{\beta}]\mathbf{x}\} = \mathbf{0}. \tag{8.1.16}$$

The sample analog of (8.1.16) is

$$\frac{1}{N}\sum_{i=1}^{N} 1(\mathbf{x}_i'\hat{\boldsymbol{\beta}} > 0)[\min\{y_i, 2\mathbf{x}_i'\hat{\boldsymbol{\beta}}\} - \mathbf{x}_i'\hat{\boldsymbol{\beta}}]\mathbf{x}_i = \mathbf{0}. \tag{8.1.17}$$

Equation (8.1.17) is the first-order condition of applying the least-squares principle to the symmetrically censored data for observations in the region AOB and the boundary OA and OB (the circled points in Fig. 8.2).

However, there could be multiple roots that satisfy (8.1.14) or (8.1.17) because of the requirement $1(\mathbf{x}_i'\boldsymbol{\beta} > 0)$. For instance, $\hat{\boldsymbol{\beta}} = \mathbf{0}$ is one such root. To ensure the uniqueness of $\hat{\boldsymbol{\beta}}$ to satisfy these conditions, Powell (1986) proposes

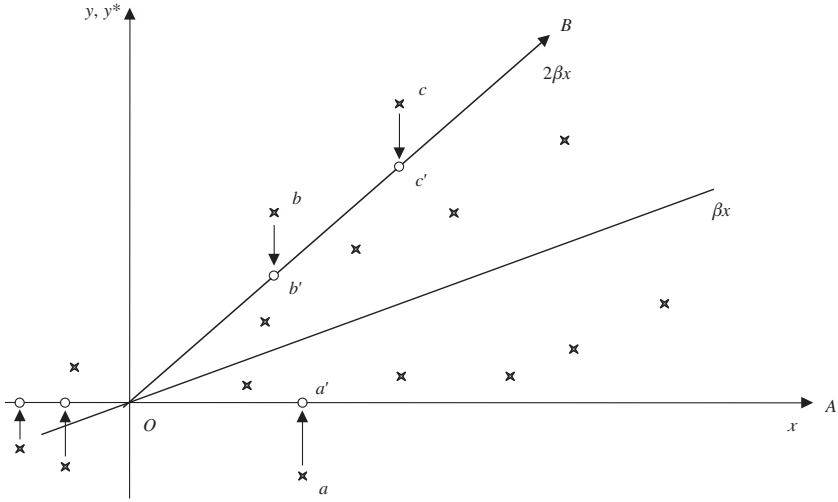


Figure 8.2. Distribution of $y$ and $y^*$ under symmetric trimming.

the symmetrically trimmed least squares estimator as the $\hat{\boldsymbol{\beta}}$ that minimizes

$$R_N(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left\{ y_i - \max \left( \frac{1}{2} y_i, \mathbf{x}_i' \boldsymbol{\beta} \right) \right\}^2, \tag{8.1.18}$$

for the truncated data, and

$$\begin{aligned} S_N(\boldsymbol{\beta}) = & \sum_{i=1}^{N} \left\{ y_i - \max \left( \frac{1}{2} y_i, \boldsymbol{\beta}' \mathbf{x}_i \right) \right\}^2 \\ & + \sum_{i=1}^{N} 1(y_i > 2\mathbf{x}'\boldsymbol{\beta}) \left\{ \left( \frac{1}{2} y_i \right)^2 - [\max (0, \mathbf{x}_i' \boldsymbol{\beta})]^2 \right\} \end{aligned} \tag{8.1.19}$$

for the censored data. When $u_i$ are mutually independently unimodally symmetrically distributed, the objective function (8.1.18) is convex in $\boldsymbol{\beta}$. The motivation for $R_N(\boldsymbol{\beta})$ is that not only will they yield first-order conditions of the form (8.1.14), it also serves to eliminate inconsistent roots that satisfy (8.1.14) with the additional "global" restrictions that for observations correspond to $\mathbf{x}'\boldsymbol{\beta} \leq 0$, $Ey_i^2$ will be smaller than those correspond to $\mathbf{x}'\boldsymbol{\beta} > 0$. Therefore, if $\mathbf{x}_i'\hat{\boldsymbol{\beta}} < 0$ while $\mathbf{x}_i'\boldsymbol{\beta} > 0$, it introduces a penalty of $(\frac{1}{2} y_i)^2$ in $R_N(\boldsymbol{\beta})$.

The motivation for $S_N(\boldsymbol{\beta})$ (8.1.19) is that for observations greater than $2\mathbf{x}'\boldsymbol{\beta}$, $S_N(\boldsymbol{\beta})$ will have partial derivatives equal to $-2(\mathbf{x}'\boldsymbol{\beta})\mathbf{x}$ if $\mathbf{x}'\boldsymbol{\beta} > 0$ and for observations correspond to $\mathbf{x}'\boldsymbol{\beta} < 0$ it will have 0 weight in the first-order condition (8.1.17), while in the meantime it imposes a penalty factor $\frac{1}{2} y_i^2$ in $S_N(\hat{\boldsymbol{\beta}})$ for observations corresponding to $\mathbf{x}_i'\hat{\boldsymbol{\beta}} < 0$ while $\mathbf{x}_i'\boldsymbol{\beta} > 0$. However, we no longer need unimodality of $u$ for censored data to ensure that the objective

function $S_N(\boldsymbol{\beta})$ is convex in $\boldsymbol{\beta}$. All we need is $u$ being independently symmetrically distributed. Powell (1986) shows that minimizing (8.1.18) or (8.1.19) yields $\sqrt{N}$ consistent and asymptotically normally distributed estimator.

The least-squares method yields the mean. The least absolute deviation method yields the median (e.g., Amemiya 1984). When $E(y^* \mid \mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$, censoring affects the mean, $E(y \mid \mathbf{x})$, but does not affect the median; therefore Powell (1984) suggests a least absolute deviation estimator of $\boldsymbol{\beta}$ by minimizing

$$\tilde{S} = \frac{1}{N}\sum_{i=1}^{N} \mid y_i - \max\,(0,\ \mathbf{x}_i'\boldsymbol{\beta}) \mid . \tag{8.1.20}$$

When data are truncated at 0, negatively realized $y^*(u < -\mathbf{x}'\boldsymbol{\beta})$ are unobserved. To restore the symmetry, Powell (1984) suggests minimizing

$$\tilde{R} = \frac{1}{N}\sum_{i=1}^{N} \left| y_i - \max\left(\frac{1}{2}\, y_i, \mathbf{x}_i'\beta\right)\right| . \tag{8.1.21}$$

The exogenously determined limited dependent variable models can be generalized to consider a variety of endogenously determined sample selection issues. For instance, in the Gronau (1976) and Heckman's (1976a) female labor supply model the hours worked are observed only for those women who decide to participate in the labor force. In other words, instead of an exogenously given truncating or censoring value, they are endogenously and stochastically determined by a selection equation

$$d_i^* = \mathbf{w}_i'\mathbf{a} + v_i, \quad i = 1, \ldots, N, \tag{8.1.22}$$

where $\mathbf{w}_i$ is a vector of exogenous variables, $\mathbf{a}$ is the parameter vector and $v_i$ is the random error term assumed to be i.i.d. with mean 0 and variance normalized to be 1. The sample $(y_i, d_i), i = 1, \ldots, N$ are related to $y_i^*$ and $d_i^*$ by the rule

$$d = \begin{cases} 1, & \text{if} \quad d^* > 0, \\ 0, & \text{if} \quad d^* \leq 0, \end{cases} \tag{8.1.23}$$

$$y = \begin{cases} y^*, & \text{if} \quad d = 1, \\ 0, & \text{if} \quad d = 0. \end{cases} \tag{8.1.24}$$

Model of (8.1.1), (8.1.22)–(8.1.24) is called the type II Tobit model by Amemiya (1985). Then

$$E(y_i \mid d_i = 1) = \mathbf{x}_i'\boldsymbol{\beta} + E(u_i \mid v_i > -\mathbf{w}_i'\mathbf{a}). \tag{8.1.25}$$

The likelihood function of $(y_i, d_i)$ is

$$\begin{aligned} L &= \prod_{c} \text{Prob}\,(d_i = 0) \prod_{\bar{c}} f(y_i^* \mid d_i = 1)\,\text{Prob}\,(d_i = 1), \\ &= \prod_{c} \text{Prob}\,(d_i = 0) \cdot \prod_{\bar{c}} \text{Prob}\,(d_i^* > 0 \mid y_i) f(y_i), \end{aligned} \tag{8.1.26}$$

where $c = \{i \mid d_i = 0\}$ and $\bar{c}$ denotes its complement. If the joint distribution of $(u, v)$ is specified, one can estimate this model by the MLE. For instance, if $(u, v)$ is jointly normally distributed with mean $(0, 0)$ and covariance matrix $\begin{pmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{vu} & 1 \end{pmatrix}$, then

$$E(u \mid v > -\mathbf{w}'\mathbf{a}) = \sigma_{uv} \frac{\phi(\mathbf{w}'a)}{\Phi(\mathbf{w}'\mathbf{a})}, \tag{8.1.27}$$

$$\text{Prob}\,(d = 0) = [1 - \Phi(\mathbf{w}'\mathbf{a})] = \Phi(-\mathbf{w}'\mathbf{a}), \tag{8.1.28}$$

$$\text{Prob}\,(d = 1 \mid y) = \Phi\left\{\mathbf{w}'\mathbf{a} + \frac{\sigma_{uv}}{\sigma_u}(y - \mathbf{x}'\boldsymbol{\beta})\right\}. \tag{8.1.29}$$

Alternatively, Heckman's (1979) two-stage method can be applied: first, estimate $\mathbf{a}$ by a probit MLE of $d_i, i = 1, \ldots, N$. Evaluate $\phi(\mathbf{a}'\mathbf{w}_i)/\Phi(\mathbf{a}'\mathbf{w}_i)$ using the estimated $\mathbf{a}$. Second, regress $y_i$ on $\mathbf{x}_i$ and $\phi(\hat{\mathbf{a}}'\mathbf{w}_i)/\Phi(\hat{\mathbf{a}}'\mathbf{w}_i)$ using data corresponding to $d_i = 1$ only.

Just like the standard Tobit model, the consistency and asymptotic normality of the MLE and Heckman two-stage estimator for the endogenously determined selection depend critically on the correct assumption of the joint probability distribution of $(u, v)$. When the distribution of $(u, v)$ is unknown, the coefficients of $\mathbf{x}$ that are not overlapping with $\mathbf{w}$ can be estimated by a semiparametric method.

For ease of exposition, suppose that there are no variables appearing in both $\mathbf{x}$ and $\mathbf{w}$; then as noted by Robinson (1988b), the model of (8.1.1), (8.1.23), and (8.1.24) conditional on $d_i = 1$ becomes a partially linear model of the form:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \lambda(\mathbf{w}_i) + \epsilon_i, \tag{8.1.30}$$

where $\lambda(\mathbf{w}_i)$ denotes the unknown selection factor. The expectation of $y_i$ conditional on $\mathbf{w}_i$ and $d_i = 1$ is equal to

$$E(y_i \mid \mathbf{w}_i, d_i = 1) = \boldsymbol{\beta}' E(\mathbf{x}_i \mid \mathbf{w}_i, d_i = 1) + \lambda(\mathbf{w}_i). \tag{8.1.31}$$

Subtracting (8.1.31) from (8.1.30) yields

$$y_i - E(y_i \mid \mathbf{w}_i, d_i = 1) = \boldsymbol{\beta}'(\mathbf{x}_i - E(\mathbf{x}_i \mid \mathbf{w}_i, d_i = 1)) + \epsilon_i, \tag{8.1.32}$$

where $E(\epsilon_i \mid \mathbf{w}_i, \mathbf{x}_i, d_i = 1) = 0$. Thus, Robinson (1988b) suggests estimating $\boldsymbol{\beta}$ by

$$\boldsymbol{\beta} = \left\{E(\mathbf{x} - E(\mathbf{x} \mid \mathbf{w}))[\mathbf{x} - E(\mathbf{x} \mid \mathbf{w})]'\right\}^{-1} E[(\mathbf{x} - E(\mathbf{x} \mid \mathbf{w}))][y - E(y \mid \mathbf{w})], \tag{8.1.33}$$

using the truncated sample.

The first-stage conditional expectation for the estimator (8.1.31) can be estimated by the nonparametric method. For instance, one may use the kernel

method to estimate the density of $y$ at $y_a$ (e.g., Härdle 1990; Robinson 1989)

$$\hat{f}(y_a) = \frac{1}{Nh_N} \sum_{i=1}^{N} k\left(\frac{y_i - y_a}{h_N}\right), \tag{8.1.34}$$

where $h_N$ is a positive number called the "bandwidth" or "smoothing" parameter that tends to 0 as $N \to \infty$, $k(u)$ is a kernel function that is a bounded symmetric probability density function (pdf) that integrates to 1. Similarly, one can construct a kernel estimator of a multivariate pdf at $\mathbf{w}_a$, $f(\mathbf{w}_a)$ by

$$\hat{f}(\mathbf{w}_a) = \frac{1}{N \mid H_m \mid} \sum_{i=1}^{N} k_m\big(H_m^{-1}(\mathbf{w}_i - \mathbf{w}_a)\big), \tag{8.1.35}$$

where $\mathbf{w}$ is a $m \times 1$ vector of random variables, $k_m$ is a kernel function on $m$ dimensional space, and $H_m$ is a positive definite matrix. For instance, $k_m(\mathbf{u})$ can be the multivariate normal density function or $k_m(\mathbf{u}) = \prod_{j=1}^{m} k(u_j)$, $\mathbf{u}' = (u_1, \ldots, u_m)$, $H_m = \text{ diag } (h_{1N}, \ldots, h_{mN})$.

Kernel estimates of a conditional pdf, $f(y_a \mid \mathbf{w}_a)$ or conditional expectations $Eg(y \mid \mathbf{w}_a)$ may be derived from the kernel estimates of the joint pdf and marginal pdf. Thus, the conditional pdf may be estimated by

$$\hat{f}(y_a \mid \mathbf{w}_a) = \frac{\hat{f}(y_a, \mathbf{w}_a)}{\hat{f}(\mathbf{w}_a)} \tag{8.1.36}$$

and the conditional expectation by

$$\hat{E}g(y \mid \mathbf{w}_a) = \frac{1}{N \mid H_m \mid} \sum_{i=1}^{N} g(y_i) k_m(H_m^{-1}(\mathbf{w}_i - \mathbf{w}_a)) / \hat{f}(\mathbf{w}_a). \tag{8.1.37}$$

The Robinson (1988b) approach does not allow the identification of the parameters of variables that appear both in the regression equation, $\mathbf{x}$, and the selection equation, $\mathbf{w}$. When there are variables appearing in both $\mathbf{x}$ and $\mathbf{w}$, Newey (2009) suggests a two-step series method of estimating $\boldsymbol{\beta}$ provided that the selection correction term of (8.1.30), $\lambda(w_i, d_i = 1)$, is a function of the single index, $\mathbf{w}_i'\mathbf{a}$,

$$\lambda(\mathbf{w}, d = 1) = E[u \mid v(\mathbf{w}'\mathbf{a}), d = 1]. \tag{8.1.38}$$

The first step of Newey's method uses the distribution-free method discussed in Chapter 7 or Klein and Spady (1993) to estimate $\mathbf{a}$. The second step consists of a linear regression of $d_i y_i$ on $d_i \mathbf{x}_i$ and the approximations of $\lambda(w_i)$. Newey suggests approximating $\lambda(\mathbf{w}_i)$ by either a polynomial function of $(\mathbf{w}_i'\hat{\mathbf{a}})$ or a spline function, $\mathbf{P}_N^K(\mathbf{w}'\mathbf{a}) = (P_{1K}(\mathbf{w}'\mathbf{a}), P_{2K}(\mathbf{w}'\mathbf{a}), \ldots, P_{KK}(\mathbf{w}'\mathbf{a}))'$ with the property that for large $K$, a linear combination of $\mathbf{P}_N^K(\mathbf{w}'\mathbf{a})$ can approximate an unknown function of $\lambda(\mathbf{w}'\mathbf{a})$ well. Newey (2009) shows that the two-step series estimation of $\boldsymbol{\beta}$ is consistent and asymptotically normally distributed when $N \to \infty$, $K \to \infty$, and $\sqrt{N}K^{-s-t+1} \to 0$ where $s \geq 5$ and $K^7/N \to 0$

if $P_N^K(\mathbf{w}'\mathbf{a})$ is a power series or $m \geq t-1$, $s \geq 3$, and $K^4/N \to 0$ if $P_N^K(\mathbf{w}'\mathbf{a})$ is a spline of degree $m$ in $(\mathbf{w}'\mathbf{a})$.[2]

If the selection factor $\lambda(\mathbf{w}_i)$ is a function of a "single index," $\mathbf{w}_i'\mathbf{a}$, and the components of $\mathbf{w}_i$ are a subvector of $\mathbf{x}_i$, instead of subtracting (8.1.32) from (8.1.31) to eliminate the unknown selection factor $\lambda(\mathbf{w}_i)$, Ahn and Powell (1993) note that for those individuals with $\mathbf{w}_i'\mathbf{a} = \mathbf{w}_j'\mathbf{a}$, $\lambda(\mathbf{w}_i'\mathbf{a}) = \lambda(\mathbf{w}_j'\mathbf{a})$. Thus, conditional on $(\mathbf{w}_i'\mathbf{a} = \mathbf{w}_j'\mathbf{a}, d_i = 1, d_j = 1)$,

$$(y_i - y_j) = (\mathbf{x}_i - \mathbf{x}_j)'\boldsymbol{\beta} + (\epsilon_i - \epsilon_j), \tag{8.1.39}$$

where the error term $(\epsilon_i - \epsilon_j)$ is symmetrically distributed around 0. They show that if $\lambda$ is a sufficiently "smooth" function, and $\hat{\mathbf{a}}$ is a consistent estimator of $\mathbf{a}$, observations for which the difference $(\mathbf{w}_i - \mathbf{w}_j)'\hat{\mathbf{a}}$ is close to 0 should have $\lambda(\mathbf{x}_i'\hat{\mathbf{a}}) - \lambda(\mathbf{w}_j'\hat{\mathbf{a}}) \simeq 0$. Therefore, Ahn and Powell (1993) proposes a two-step procedure. In the first step, consistent semiparametric estimates of the coefficients of the "selection" equation are obtained. The result is used to obtain estimates of the "single index, $\mathbf{x}_i'\mathbf{a}$," variables characterizing the selectivity bias in the equation of index. The second step of the approach estimates the parameters of interest by a weighted least-squares (or instrumental) variables regression of pairwise differences in dependent variables in the sample on the corresponding differences in explanatory variables:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{AP} = &\left[\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} K\left(\frac{(\mathbf{w}_i - \mathbf{w}_j)'\hat{\mathbf{a}}}{h_N}\right)\cdot(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)'d_i d_j\right]^{-1} \\ &\cdot\left[\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} K\left(\frac{(\mathbf{w}_i - \mathbf{w}_j)'\hat{\mathbf{a}}}{h_N}\right)\cdot(\mathbf{x}_i - \mathbf{x}_j)(y_i - y_j)d_i d_j\right],\end{aligned} \tag{8.1.40}$$

where $K(\cdot)$ is a kernel density weighting function that is bounded, symmetric, and tends to 0 as the absolute value of its argument increases, and $h_N$ is a positive constant (or bandwidth) that decreases to 0, $N(h_N)^\delta \longrightarrow 0$ as $N \to \infty$, where $\delta \in (6, 8)$. Often, standard normal density is used as a kernel function. The effect of multiplying the $K(\cdot)$ is to give more weights to observations with $\frac{1}{h_N}(\mathbf{w}_i - \mathbf{w}_j)'\hat{\mathbf{a}} \simeq 0$ and less weight to those observations that $\mathbf{w}_i'\hat{\mathbf{a}}$ is different from $\mathbf{w}_j'\hat{\mathbf{a}}$ so that in the limit only observations with $\mathbf{w}_i'\mathbf{a} = \mathbf{w}_j'\mathbf{a}$ are used in (8.1.39) and (8.1.40) converges to a weighted least-squares estimator for the

[2] For instance, a spline of degree $m$ in $(\mathbf{w}'\hat{\mathbf{a}})$ with $L$ evenly spaced knots on $[-1, 1]$ can be based on

$$\begin{aligned}P_{kK} &= (\mathbf{w}'\mathbf{a})^{k-1}, 1 \leq k \leq m+1, \\ &= \left\{[(\mathbf{w}'\mathbf{a}) + 1 - 2(k-m-1)/(L+1)]_+\right\}^m, m+2 \leq k \leq m+1+L \equiv K,\end{aligned}$$

where $b_+ \equiv 1(b > 0)\cdot b$.

truncated data,

$$\hat{\boldsymbol{\beta}}_{AP} \longrightarrow \left\{E\{f(\mathbf{w}'\mathbf{a})[\mathbf{x} - E(\mathbf{x} \mid \mathbf{w}'\mathbf{a})][x - E(\mathbf{x} \mid \mathbf{w}'\mathbf{a})]'\right\}^{-1} \cdot \left\{E\{f(\mathbf{w}'\mathbf{a})[\mathbf{x} - E(\mathbf{x} \mid \mathbf{w}'\mathbf{a})][y - E(y \mid \mathbf{w}'\mathbf{a})]\right\}, \tag{8.1.41}$$

where $f(\mathbf{w}'\mathbf{a})$ denotes the density function of $\mathbf{w}'\mathbf{a}$, which is assumed to be continuous and bounded above.

Both the Robinson (1988b) semiparametric estimator and the Powell type pairwise differencing estimator converge to the true value at the speed of $N^{-1/2}$. However, neither method can provide an estimate of the intercept term because differencing the observation conditional on $\mathbf{w}$ or $\mathbf{w}'\mathbf{a}$, although it eliminates the selection factor $\lambda(\mathbf{w})$, it also eliminates the constant term, nor can $\mathbf{x}$ and $\mathbf{w}$ be identical. Chen (1999) notes that if $(u, v)$ are jointly symmetrical and $\mathbf{w}$ includes a constant term,

$$\begin{aligned} &E(u \mid v > -\mathbf{w}'\mathbf{a}) \operatorname{Prob}(v > -\mathbf{w}'a) - E(u \mid v > \mathbf{w}'a) \operatorname{Prob}(v > \mathbf{w}'\mathbf{a}) \\ &= \int_{-\infty}^{\infty}\int_{-\mathbf{w}'\mathbf{a}}^{\infty} uf(u, v)du\,dv - \int_{-\infty}^{\infty}\int_{\mathbf{w}'\mathbf{a}}^{\infty} uf(u, v)du\,dv \\ &= \int_{-\infty}^{\infty}\int_{-\mathbf{w}'\mathbf{a}}^{\mathbf{w}'\mathbf{a}} uf(u, v)du\,dv = 0, \end{aligned} \tag{8.1.42}$$

where, without loss of generality, we let $\mathbf{w}'\mathbf{a} > 0$. It follows that

$$\begin{aligned} &E[d_i y_i - d_j y_j - (d_i \mathbf{x}_i - d_j \mathbf{x}_j)'\boldsymbol{\beta} \mid \mathbf{w}_i'\mathbf{a} = -\mathbf{w}_j'\mathbf{a}, \mathbf{w}_i, \mathbf{w}_j] \\ &= E[d_i u_i - d_j u_j \mid \mathbf{w}_i'\mathbf{a} = -\mathbf{w}_j'\mathbf{a}, \mathbf{w}_i, \mathbf{w}_j] = 0. \end{aligned} \tag{8.1.43}$$

Because $E[d_i - d_j \mid \mathbf{w}_i'\mathbf{a} = -\mathbf{w}_j'\mathbf{a}, \mathbf{w}_i, \mathbf{w}_j] = 2 \operatorname{Prob}(d_i = 1 \mid \mathbf{w}_i'\mathbf{a}) - 1 \neq 0$ and the conditioning is on $\mathbf{w}_i'\mathbf{a} = -\mathbf{w}_j'\mathbf{a}$, not on $\mathbf{w}_i'\mathbf{a} = \mathbf{w}_j'\mathbf{a}$, the moment condition (8.1.43) allows the identification of the intercept and the slope parameters without the need to impose the exclusion restriction that at least one component of $\mathbf{x}$ is excluded from $\mathbf{w}$. Therefore, Chen (1999) suggests a $\sqrt{N}$ consistent instrumental variable estimator for the intercept and the slope parameters as

$$\begin{aligned} \hat{\boldsymbol{\beta}}_c = &\left[\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} K\left(\frac{(\mathbf{w}_i + \mathbf{w}_j)'\hat{\mathbf{a}}}{h_N}\right)(d_i \mathbf{x}_i - d_j \mathbf{x}_j)(\mathbf{z}_i - \mathbf{z}_j)'\right]^{-1} \\ &\cdot \left[\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} K\left(\frac{(\mathbf{w}_i + \mathbf{w}_j)'\hat{\mathbf{a}}}{h_N}\right)(\mathbf{z}_i - \mathbf{z}_j)'(d_i y_i - d_j y_j)\right], \end{aligned} \tag{8.1.44}$$

where $\mathbf{z}_i$ are the instruments for $d_i \mathbf{x}_i$. In the case when $y$ are unobservable, but the corresponding $\mathbf{x}$ are observable, the natural instrument will be $E(d \mid \mathbf{w}'\mathbf{a})\mathbf{x}$. An efficient method for estimating binary choice models that contain an

intercept term suggested by Chen (2000) can be used to obtain the first-stage estimate of **a**.

## 8.2 AN EXAMPLE – NONRANDOMLY MISSING DATA

### 8.2.1 Introduction

Attrition is a problem in any panel survey. For instance, by 1981, all four of the national longitudinal surveys started in the 1960s had lost at least one-fourth of their original samples. In the Gary income maintenance project, 206 of the sample of 585 black, male-headed households, or 35.2 percent, did not complete the experiment. In Section 11.1 we discuss procedures to handle randomly missing data. However, the major problem in panel data is not simply missing data but also the possibility that they are missing for a variety of self-selection reasons. For instance, in a social experiment such as the New Jersey or Gary negative-income-tax experiment, some individuals may decide that keeping the detailed records that the experiments require is not worth the payment. Also, some may move or may be inducted into the military. In some experiments, persons with large earnings receive no experimental-treatment benefit and thus drop out of the experiment altogether. This attrition may negate the randomization in the initial experiment design. If the probability of attrition is correlated with experimental response, then traditional statistical techniques will lead to biased and inconsistent estimates of the experimental effect. In this section we show how models of limited dependent variables [e.g., see the surveys of Amemiya (1984), Heckman (1976a), and Maddala (1983)] can provide both the theory and computational techniques for analyzing nonrandomly missing data (Griliches, Hall, and Hausman 1978; Hausman and Wise 1979).[3]

### 8.2.2 A Probability Model of Attrition and Selection Bias

Suppose that the structural model is

$$y_{it} = \boldsymbol{\beta}'\mathbf{x}_{it} + v_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{matrix} \tag{8.2.1}$$

where the error term $v_{it}$ is assumed to follow a conventional error-components formulation $v_{it} = \alpha_i + u_{it}$. For ease of exposition, we assume that $T = 2$.

If attrition occurs in the second period, a common practice is to discard those observations for which $y_{i2}$ is missing. But suppose that the probability of

[3] Another example is the analysis of event histories in which responses are at nonequally spaced points in time (e.g., Heckman and Singer 1984; Lancaster 1990). Some people choose to model event histories in discrete time using sequences of binary indicators. Then the subject becomes very much like the discrete panel data analysis discussed in Chapter 7.

observing $y_{i2}$ varies with its value, as well as the values of other variables; then the probability of observing $y_{i2}$ will depend on $v_{i2}$. Least-squares of (8.2.1) based on observed $y$ will lead to biased estimates of the underlying structural parameters and the experimental response.

To formalize the argument, let the indicator variable $d_i = 1$ if $y_{i2}$ is observed in period 2, and $d_i = 0$ if $y_{i2}$ is not observed; in other words, attrition occurs. Suppose that $y_{i2}$ is observed ($d_i = 1$) if the latent variable

$$d_i^* = \gamma y_{i2} + \boldsymbol{\theta}' \mathbf{x}_{i2} + \boldsymbol{\delta}' \mathbf{w}_i + \epsilon_i^* \geq 0, \tag{8.2.2}$$

where $\mathbf{w}_i$ is a vector of variables that do not enter the conditional expectation of $y$ but affect the probability of observing $y$; $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ are vectors of parameters; and $(v_i, \epsilon_i^*)$ are jointly normally distributed. Substituting for $y_{i2}$ leads to the reduced-form specification

$$\begin{aligned} d_i^* &= (\gamma \boldsymbol{\beta}' + \boldsymbol{\theta}') \mathbf{x}_{i2} + \boldsymbol{\delta}' \mathbf{w}_i + \gamma v_{i2} + \epsilon_i^* \\ &= \boldsymbol{\pi}' \mathbf{x}_{i2} + \boldsymbol{\delta}' \mathbf{w}_i + \epsilon_i \\ &= \mathbf{a}' R_i + \epsilon_i, \end{aligned} \tag{8.2.3}$$

where $\epsilon_i = \gamma v_{i2} + \epsilon_i^*$, and $R_i = (\mathbf{x}_{i2}', \mathbf{w}_i')'$, and $\mathbf{a}' = (\boldsymbol{\pi}', \boldsymbol{\delta}')$. We normalize the variance of $\epsilon_i$, $\sigma_\epsilon^2$, equal to 1. Then the probabilities of retention and attrition are probit functions given, respectively, by

$$\begin{aligned} &\text{Prob}(d_i = 1) = \Phi(\mathbf{a}' R_i), \quad \text{and} \\ &\text{Prob}(d_i = 0) = 1 - \Phi(\mathbf{a}' R_i), \end{aligned} \tag{8.2.4}$$

where $\Phi(\cdot)$ is the standard normal distribution function.

Suppose we estimate the model (8.2.1) using only complete observations. The conditional expectation of $y_{i2}$, given that it is observed, is

$$E(y_{i2} \mid \mathbf{x}_{i2}, \mathbf{w}_i, d_i = 1) = \boldsymbol{\beta}' \mathbf{x}_{i2} + E(v_{i2} \mid \mathbf{x}_{i2}, \mathbf{w}_i, d_i = 1). \tag{8.2.5}$$

From $v_{i2} = \sigma_{2\epsilon} \epsilon_i + \eta_i$, where $\sigma_{2\epsilon}$ is the covariance between $v_{i2}$ and $\epsilon_i$, and $\eta_i$ is independent of $\epsilon_i$ (Anderson 1985, Chapter 2), we have

$$\begin{aligned} E(v_{i2} \mid \mathbf{w}_i, d_i = 1) &= \sigma_{2\epsilon} E(\epsilon_i \mid \mathbf{w}_i, d_i = 1) \\ &= \frac{\sigma_{2\epsilon}}{\Phi(\mathbf{a}' R_i)} \int_{-\mathbf{a}' R_i}^{\infty} \epsilon \cdot \frac{1}{\sqrt{2\pi}} e^{-\epsilon^2/2} d\epsilon \\ &= \sigma_{2\epsilon} \frac{\phi(\mathbf{a}' R_i)}{\Phi(\mathbf{a}' R_i)}, \end{aligned} \tag{8.2.6}$$

where $\phi(\cdot)$ denotes the standard normal density function. The last equality of (8.2.6) follows from the formula that the derivative of the standard normal density function $\phi(\epsilon)$ with respect to $\epsilon$ is $-\epsilon \phi(\epsilon)$. Therefore,

$$E(y_{i2} \mid \mathbf{x}_{i2}, \mathbf{w}_i, d_i = 1) = \boldsymbol{\beta}' \mathbf{x}_{i2} + \sigma_{2\epsilon} \frac{\phi(\mathbf{a}' R_i)}{\Phi(\mathbf{a}' R_i)}. \tag{8.2.7}$$

Thus, estimating (8.2.1) using complete observations will lead to biased and inconsistent estimates of $\boldsymbol{\beta}$ unless $\sigma_{2\epsilon} = 0$. To correct for selection bias, one can use either Heckman's (1979) two-stage method (see Section 8.1) or the maximum-likelihood method.

When $d_i = 1$, the joint density of $d_i = 1$, $y_{i1}$, and $y_{i2}$ is given by

$$
\begin{aligned}
f(d_i = 1, y_{i1}, y_{i2}) &= \text{Prob}(d_i = 1 \mid y_{i1}, y_{i2}) f(y_{i1}, y_{i2}) \\
&= \text{Prob}(d_i = 1 \mid y_{i2}) f(y_{i1}, y_{i2}) \\
&= \Phi \left\{ \frac{\mathbf{a}' R_i + \left( \frac{\sigma_{2\epsilon}}{\sigma_u^2 + \sigma_\alpha^2} \right) (y_{i2} - \boldsymbol{\beta}' \mathbf{x}_{i2})}{\left[ 1 - \frac{\sigma_{2\epsilon}^2}{\sigma_u^2 + \sigma_\alpha^2} \right]^{1/2}} \right\} \\
&\quad \cdot [2\pi \sigma_u^2 (\sigma_u^2 + 2\sigma_\alpha^2)]^{-1/2} \\
&\quad \cdot \exp \left\{ -\frac{1}{2\sigma_u^2} \left[ \sum_{t=1}^{2} (y_{it} - \boldsymbol{\beta}' \mathbf{x}_{it})^2 - \frac{\sigma_\alpha^2}{\sigma_u^2 + 2\sigma_\alpha^2} \right. \right. \\
&\quad \left. \left. \cdot \left( \sum_{t=1}^{2} (y_{it} - \boldsymbol{\beta}' \mathbf{x}_{it}) \right)^2 \right] \right\},
\end{aligned}
\tag{8.2.8}
$$

where the first term follows from the fact that the conditional density of $f(\epsilon_i \mid v_{i2})$ is normal, with mean $[\sigma_{2\epsilon}/(\sigma_u^2 + \sigma_\alpha^2)]v_{i2}$ and variance $1 - \sigma_{2\epsilon}^2/(\sigma_u^2 + \sigma_\alpha^2)$. When $d_i = 0$, $y_{i2}$ is not observed and must be "integrated out." In this instance, the joint density of $d_i = 0$, and $y_{i1}$ is given by

$$
\begin{aligned}
f(d_i = 0, y_{i1}) = &\ \text{Prob}\,(d_i = 0 \mid y_{i1}) f(y_{i1}) \\
&= \left\{ 1 - \Phi \left[ \frac{\mathbf{a}' R_i + \frac{\sigma_{1\epsilon}}{\sigma_u^2 + \sigma_\alpha^2} (y_{i1} - \boldsymbol{\beta}' \mathbf{x}_{i1})}{\left[ 1 - \frac{\sigma_{1\epsilon}^2}{\sigma_u^2 + \sigma_\alpha^2} \right]^{1/2}} \right] \right\} \\
&\quad \cdot [2\pi (\sigma_u^2 + \sigma_\alpha^2)]^{-1/2} \\
&\quad \cdot \ \exp \left\{ -\frac{1}{2(\sigma_u^2 + \sigma_\alpha^2)} (y_{i1} - \boldsymbol{\beta}' \mathbf{x}_{i1})^2 \right\}.
\end{aligned}
\tag{8.2.9}
$$

The right-hand side of (8.2.9) follows from the fact that $f(\epsilon_i \mid v_{i1})$ is normal, with mean $[\sigma_{1\epsilon}/(\sigma_u^2 + \sigma_\alpha^2)]v_{i1}$ and variance $1 - \sigma_{1\epsilon}^2/(\sigma_u^2 + \sigma_\alpha^2)$, where $\sigma_{1\epsilon}$ is the covariance between $v_{i1}$ and $\epsilon_i$, which is equal to $\sigma_{2\epsilon} = \sigma_\alpha^2/(\sigma_u^2 + \sigma_\alpha^2)$.

The likelihood function follows from (8.2.8) and (8.2.9). Order the observations so that the first $N_1$ observations correspond to $d_i = 1$, and the remaining

$N - N_1$ correspond to $d_i = 0$; then the log-likelihood function is given by

$$\begin{aligned}
\log L = &- N \log 2\pi - \frac{N_1}{2} \log \sigma_u^2 - \frac{N_1}{2} \log (\sigma_u^2 + 2\sigma_\alpha^2) \\
&- \frac{N - N_1}{2} \log (\sigma_u^2 + \sigma_\alpha^2) \\
&- \frac{1}{2\sigma^2} \sum_{i=1}^{N_1} \left\{ \sum_{t=1}^{2} (y_{it} - \boldsymbol{\beta}' \mathbf{x}_{it})^2 - \frac{\sigma_\alpha^2}{\sigma_u^2 + 2\sigma_\alpha^2} \left[ \sum_{t=1}^{2} (y_{it} - \beta' x_{it}) \right]^2 \right\} \\
&+ \sum_{i=1}^{N_1} \log \Phi \left\{ \frac{\mathbf{a}' R_i + \frac{\sigma_{2\epsilon}}{\sigma_u^2 + \sigma_\alpha^2} (y_{i2} - \boldsymbol{\beta}' \mathbf{x}_{i2})}{\left[ 1 - \frac{\sigma_{2\epsilon}^2}{\sigma_u^2 + \sigma_\alpha^2} \right]^{1/2}} \right\} \qquad (8.2.10) \\
&- \frac{1}{2(\sigma_u^2 + \sigma_\alpha^2)} \sum_{i=N_1+1}^{N} (y_{i1} - \boldsymbol{\beta}' \mathbf{x}_{i1})^2 \\
&+ \sum_{i=N_1+1}^{N} \log \left\{ 1 - \Phi \left[ \frac{\mathbf{a}' R_i + \frac{\sigma_{1\epsilon}}{\sigma_u^2 + \sigma_\alpha^2} (y_{i1} - \boldsymbol{\beta}' \mathbf{x}_{i1})}{\left[ 1 - \frac{\sigma_{1\epsilon}^2}{\sigma_u^2 + \alpha_\alpha^2} \right]^{1/2}} \right] \right\}.
\end{aligned}$$

The critical parameter for attrition bias is $\sigma_{2\epsilon}$. If $\sigma_{2\epsilon} = 0$, so does $\sigma_{1\epsilon}$. The likelihood function (8.2.10) then separates into two parts. One corresponds to the variance-components specification for $y$. The other corresponds to the probit specification for attrition. Thus, if attrition bias is not present, this is identical with the random missing-data situations. Generalized least-squares (GLS) techniques used to estimate (8.2.1) will lead to consistent and asymptotically efficient estimates of the structural parameters of the model.

The Hausman–Wise two-period model of attrition can be extended in a straightforward manner to more than two periods and to simultaneous-equations models with selection bias as discussed in Section 8.2. When $T > 2$, an attrition equation can be specified for each period. If attrition occurs, the individual does not return to the sample; then a series of conditional densities analogous to (8.2.8) and (8.2.9) result. The last period for which the individual appears in the sample gives information on which the random term in the attrition equations is conditioned. For periods in which the individual remains in the sample, an equation like (8.2.8) is used to specify the joint probability of no attrition and the observed values of the dependent variables.

In the case of simultaneous equations models, all the attrition model leads to is simply to add an equation for the probability of observing an individual in the sample. Then the joint density of observing in-sample respondents becomes the product of the conditional probability of the observation being in the sample, given the joint dependent variable $\mathbf{y}$ and the marginal density of $\mathbf{y}$. The joint density of incomplete respondents becomes the product of the conditional probability of the observation being out-of-sample, given the before-dropping-out

values of **y**, and the marginal density of the previous periods' **y**. The likelihood function is simply the product of these two joint densities; see Griliches et al. (1978) for a three-equation model.

The employment of probability equations to specify the status of individuals can be very useful in analyzing the general problems of changing compositions of the sample over time, in particular when changes are functions of individual characteristics. For instance, in addition to the problem of attrition in the national longitudinal surveys' samples of young men, there is also the problem of sample accretion, that is, entrance into the labor force of the fraction of the sample originally enrolled in school. The literature on switching regression models can be used as a basis for constructing behavioral models for analyzing the changing status of individuals over time.[4]

### 8.2.3 Attrition in the Gary Income-Maintenance Experiment

The Gary income-maintenance project focused on the impact of alternative sets of income-maintenance structures on work–leisure decisions. The basic project design was to divide individuals randomly into two groups: "controls" and "experimentals." The controls were not on an experimental treatment plan, but received nominal payments for completing periodic questionnaires. The experimentals were randomly assigned to one of several income-maintenance plans. The experiment had four basic plans defined by an income guarantee and a tax rate. The two guarantee levels were $4,300 and $3,300 for a family of four and were adjusted up for larger families and down for smaller families. The two marginal tax rates were 0.6 and 0.4. Retrospective information of individuals in the experiments was also surveyed for a pre-experimental period (normally just prior to the beginning of the experimental period) so that the behavior of experimentals during the experiment could be compared with their own pre-experimental behavior and also compared with that of the control group to obtain estimates of the effects of treatment plans.

Two broad groups of families were studied in the Gary experiment: black, female-headed households, and black, male-headed households. There was little attrition among the first group, but the attrition among male-headed families was substantial. Of the sample of 334 experimentals used by Hausman and Wise (1979), the attrition rate was 31.1 percent. Among the 251 controls, 40.6 percent failed to complete the experiment.

If attrition is random, as discussed in Section 11.1, it is not a major problem. What matters is that data are missing for a variety of self-selection reasons. In this case it is easy to imagine that attrition is related to endogenous variables. Beyond a break-even point, experimentals receive no benefits from the experimental treatment. The break-even point occurs when the guarantee minus taxes paid on earnings (wage rate times hours worked) is 0. Individuals with high earnings receive no treatment payment and may be much like controls

[4] See Quandt (1982) for a survey of switching regression models.

vis-à-vis their incentive to remain in the experiment. But because high earnings are caused in part by the unobserved random term of the structural equation (8.2.1), attrition may well be related to it.

Hausman and Wise (1979) estimated structural models of earnings with and without correcting for attrition. The logarithm of earnings was regressed against time trend, education, experience, union membership, health status, and the logarithm of non-labor family income. To control for the effects of the treatment, they also used a dummy variable that was 1 if for that period the household was under one of the four basic income-maintenance plans, and 0 otherwise. Because hourly wages for experimentals and controls did not differ, the coefficient of this variable provided a reasonable indicator of the effect of experimental treatment on hours worked.

Because only three observations were available during the experiment, each for a one-month period, they concentrated on a two-period model: a period for the preexperiment average monthly earnings and a period for the average earning of the three monthly observations of the experimental period. Their GLS estimates of the structural parameters that were not corrected for attrition and the maximum-likelihood estimates that incorporated the effects of attrition, (8.2.1) and (8.2.3), are presented in Table 8.1.

The attrition-bias parameter $\sigma_{2\epsilon}/(\sigma_u^2 + \sigma_\alpha^2)$ was estimated to be $-0.1089$. This indicates a small but statistically significant correlation between earnings and the probability of attrition. The estimate of the experimental effect was very close whether or not the attrition bias was corrected for. However, the experimental-effect coefficient did increase in magnitude from $-0.079$ to $-0.082$, an increase of 3.6 percent. Some of the other coefficients showed more pronounced changes. The effect of non-labor family income on earnings (hence hours worked) decreased by 23 percent from the GLS estimates, and the effect of another year of education increased by 43 percent. These results demonstrate that attrition bias was a potentially important problem in the Gary experiment. For other examples, see Ridder (1990), Nijman and Verbeek (1992), and Verbeek and Nijman (1996).

The Hausman–Wise (HW) model assumes that the contemporaneous values affect the probability of responding. Alternatively, the decision on whether to respond may be related to past experiences – if in the first period the effort in responding was high, an individual may be less inclined to respond in the second period. When the probability of attrition depends on lagged but not on contemporaneous variables, and if $v_{it}$ and $\epsilon_i^*$ are mutually independent, then individuals are missing at random (MAR) (Little and Rubin 1987; Rubin 1976) and the missing data are ignorable. (This case is sometimes referred to as selection on observables (e.g., Moffitt, Fitzgerald, and Gottschalk 1997).

Both sets of models are often used to deal with attrition in panel data sets. However, they rely on fundamentally different restrictions on the dependence of the attrition process on time path of the variables and can lead to very different inferences. In a two-period model one cannot introduce dependence on $y_{i2}$ in the MAR model, or dependence on $y_{i1}$ in the HW model without

relying heavily on functional form and distributional assumptions. However, when missing data are augmented by replacing the units who have dropped out with new units randomly sampled from the original population, called refreshment samples by Ridder (1992), it is possible to test between these two types of models nonparametrically as well as to estimate more general models (e.g., Hirano, Imbens, Ridder, and Rubin 2001).

## 8.3 TOBIT MODELS WITH RANDOM INDIVIDUAL EFFECTS

The most typical concern in empirical work using panel data has been the presence of unobserved heterogeneity.[5] Thus, a linear latent response function is often written in the form

$$y_{it}^* = \alpha_i + \boldsymbol{\beta}'\mathbf{x}_{it} + u_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{matrix} \tag{8.3.1}$$

with the error term assumed to be independent of $\mathbf{x}_{it}$ and is i.i.d. over time and across individuals, where the observed value $y_{it}$ equals to $y_{it}^*$ if $y_{it}^* > 0$ and is unobserved for $y_i^* \leq 0$ when data are truncated and is equal to 0 when data are censored. Under the assumption that $\alpha_i$ is randomly distributed with density function $g(\alpha)$ (or $g(\alpha \mid \mathbf{x})$), the likelihood function of the standard Tobit model for the truncated data is of the form

$$\prod_{i=1}^{N} \int \left[ \prod_{t=1}^{T} [1 - F(-\boldsymbol{\beta}'\mathbf{x}_{it} - \alpha_i)]^{-1} f(y_{it} - \boldsymbol{\beta}'\mathbf{x}_{it} - \alpha_i) \right] g(\alpha_i)\, d\alpha_i, \tag{8.3.2}$$

where $f(\cdot)$ denotes the density function of $u_{it}$ and $F(a) = \int_{-\infty}^{a} f(u)\, du$. The likelihood function of the censored data takes the form

$$\prod_{i=1}^{N} \int \left[ \prod_{t \in c_i} F(-\boldsymbol{\beta}'\mathbf{x}_{it} - \alpha_i) \prod_{t \in \bar{c}_i} f(y_{it} - \alpha_i - \boldsymbol{\beta}'\mathbf{x}_{it}) \right] g(\alpha_i)\, d\alpha_i, \tag{8.3.3}$$

where $c_i = \{t \mid y_{it} = 0\}$ and $\bar{c}_i$ denotes its complement. Maximizing (8.3.2) or (8.3.3) with respect to unknown parameters yield consistent and asymptotically normally distributed estimator.

Similarly, for the type II Tobit model we may specify a sample selection equation

$$d_{it}^* = \mathbf{w}_{it}'\mathbf{a} + \eta_i + v_{it}, \tag{8.3.4}$$

with the observed $(y_{it}, d_{it})$ following the rule of $d_{it} = 1$ if $d_{it}^* > 0$ and 0 otherwise as in (8.1.23) and $y_{it} = y_{it}^*$ if $d_{it} = 1$ and unknown otherwise as in (8.1.24). Suppose that the joint density of $(\alpha_i, \eta_i)$ is given by $g(\alpha, \eta)$; then the

[5] In this chapter we consider only the case involving the presence of individual specific effects. For some generalization to the estimation of random coefficient sample selection model, see Chen (1999).

likelihood function of type II Tobit model takes the form

$$\prod_{i=1}^{N}\int [\prod_{t\epsilon c_i} \text{Prob}\,(d_{it}=0 \mid \mathbf{w}_{it},\alpha_i)\prod_{t\epsilon \bar{c}_i} \text{Prob}\,(d_{it}=1 \mid \mathbf{w}_{it},\alpha_i)$$
$$\cdot\, f(y_{it} \mid \mathbf{x}_{it},\mathbf{w}_{it},\alpha_i,\eta_i,d_{it}=1)]g(\alpha_i,\eta_i)d\alpha_i\, d\eta_i \tag{8.3.5}$$
$$=\prod_{i=1}^{N}\int [\prod_{t\epsilon c_i} \text{Prob}\,(d_{it}=0 \mid \mathbf{w}_{it},\alpha_i)\prod_{t\epsilon \bar{c}_i} \text{Prob}\,(d_{it}=1 \mid \mathbf{w}_{it},\eta_i,\alpha_i,y_{it},\mathbf{x}_{it})$$
$$\cdot\, f(y_{it} \mid \mathbf{x}_{it},\alpha_i)]g(\alpha_i,\eta_i)d\alpha_i d\eta_i$$

Maximizing the likelihood function (8.3.2), (8.3.3), or (8.3.5) with respect to unknown parameters yields consistent and asymptotically normally distributed estimator of $\boldsymbol{\beta}$ when either $N$ or $T$ or both tend to infinity. However, the computation is quite tedious even with a simple parametric specification of the individuals effects $\alpha_i$ and $\eta_i$ because it involves multiple integration.[6] Neither is a generalization of the Heckman (1976a) two-stage estimator easily implementable (e.g., Nijman and Verbeek 1992; Ridder 1990; Vella and Verbeek 1999; Wooldridge 1999). Moreover, both the MLE and the Heckman two-step estimators are sensitive to the exact specification of the error distribution. However, if the random effects $\alpha_i$ and $\eta_i$ are independent of $\mathbf{x}_i$, then the Robinson (1988b) and Newey (2009) estimators (8.1.33) and (8.1.38) can be applied to obtain consistent and asymptotically normally distributed estimators of $\boldsymbol{\beta}$. Alternatively, one may ignore the randomness of $\alpha_i$ and $\eta_i$ and apply the Honoré (1992) fixed-effects trimmed least-squares or least absolute deviation estimator for the panel data censored and truncated regression models or the Kyriazidou (1997) two-step semi parametric estimator for the panel data sample selection model to estimate $\boldsymbol{\beta}$ (see Section 8.4).

## 8.4 FIXED-EFFECTS ESTIMATOR

### 8.4.1 Pairwise Trimmed Least-Squares and Least Absolute Deviation Estimators for Truncated and Censored Regressions

When the effects are fixed and if $T \to \infty$, the MLE of $\boldsymbol{\beta}'$ and $\alpha_i$ are straightforward to implement and are consistent. However panel data are often characterized by having many individuals observed over few time periods, the MLE, in general, will be inconsistent as described in Chapter 7. In this section we consider the pairwise trimmed least-squares (LS) and least absolute deviation (LAD) estimators of Honoré (1992) for panel data censored and truncated regression models that are consistent without the need to assume a parametric form for the disturbances $u_{it}$, nor homoskedasticity across individuals.

[6] A potentially computationally attractive alternative is to simulate the integrals, see Gourieroux and Monfort (1996), Keane (1994), Richard (1996), or Chapter 12, Section 12.4.

Table 8.1. *Parameter estimates of the earnings-function structural model with and without a correction for attrition*

| Variables | With attrition correction: maximum likelihood estimates (standard errors) | | Without attrition correction: Generalized-least-squares estimates (standard errors): earnings-function parameters |
|---|---|---|---|
| | Earnings-function parameters | Attrition parameters | |
| Constant | 5.8539<br>(0.0903) | −0.6347<br>(0.3351) | 5.8911<br>(0.0829) |
| Experimental effect | −0.0822<br>(0.0402) | 0.2414<br>(0.1211) | −0.0793<br>(0.0390) |
| Time trend | 0.0940<br>(0.0520) | —[a]<br>— | 0.0841<br>(0.0358) |
| Education | 0.0209<br>(0.0052) | −0.0204<br>(0.0244) | 0.0136<br>(0.0050) |
| Experience | 0.0037<br>(0.0013) | −0.0038<br>(0.0061) | 0.0020<br>(0.0013) |
| Nonlabor income | −0.0131<br>(0.0050) | 0.1752<br>(0.0470) | −0.0115<br>(0.0044) |
| Union | 0.2159<br>(0.0362) | 1.4290<br>(0.1252) | 0.2853<br>(0.0330) |
| Poor health | −0.0601<br>(0.0330) | 0.2480<br>(0.1237) | −0.0578<br>(0.0326) |
| | $\hat{\sigma}_u^2 = 0.1832$ (0.0057)<br>$\frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_u^2 + \hat{\sigma}_\alpha^2} = 0.2596$ (0.0391) | $\frac{\hat{\sigma}_{2\epsilon}}{\hat{\sigma}_u^2 + \hat{\sigma}_\alpha^2} = -0.1089$ (0.0429) | $\hat{\sigma}_u^2 = 0.1236$<br>$\frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_u^2 + \hat{\sigma}_\alpha^2} = 0.2003$ |

[a] Not estimated.

*Source:* Hausman and Wise (1979, Table IV).

*8.4.1.1 Truncated Regression*

We assume a model of (8.3.1) and (8.1.3) except that now the individual effects are assumed fixed. The disturbance $u_{it}$ is again assumed to be independently distributed over $i$ and independently, identically distributed (i.i.d) over $t$ conditional on $\mathbf{x}_i$ and $\alpha_i$.

We note that where data are truncated or censored, first differencing does not eliminate the individual specific effects from the specification. To see this, suppose that the data are truncated. Let

$$y_{it} = E(y_{it} \mid \mathbf{x}_{it}, \alpha_i, y_{it} > 0) + \epsilon_{it}, \tag{8.4.1}$$

where

$$E(y_{it} \mid \mathbf{x}_{it}, \alpha_i, y_{it} > 0) = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + E(u_{it} \mid u_{it} > -\alpha_i - \mathbf{x}_{it}'\boldsymbol{\beta}). \tag{8.4.2}$$

Since $\mathbf{x}_{it} \neq \mathbf{x}_{is}$, in general,

$$\begin{aligned} &E(y_{it} \mid \mathbf{x}_{it}, \alpha_i, y_{it} > 0) - E(y_{is} \mid \mathbf{x}_{is}, \alpha_i, y_{is} > 0) \\ &\quad = (\mathbf{x}_{it} - \mathbf{x}_{is})'\boldsymbol{\beta} + E(u_{it} \mid u_{it} > -\alpha_i - \mathbf{x}_{it}'\boldsymbol{\beta}) \\ &\qquad - E(u_{is} \mid u_{is} > -\alpha_i - \mathbf{x}_{is}'\boldsymbol{\beta}), \end{aligned} \tag{8.4.3}$$

In other words

$$\begin{aligned} (y_{it} - y_{is}) = (\mathbf{x}_{it} - \mathbf{x}_{is})'\boldsymbol{\beta} &+ E(u_{it} \mid u_{it} > -\alpha_i - \mathbf{x}_{it}'\boldsymbol{\beta}) \\ &- E(u_{is} \mid u_{is} > -\alpha_i - \mathbf{x}_{is}'\boldsymbol{\beta}) + (\epsilon_{it} - \epsilon_{is}). \end{aligned} \tag{8.4.4}$$

The truncation correction term, $E(u_{it} \mid u_{it} > -\alpha_i - \mathbf{x}_{it}'\boldsymbol{\beta})$, which is a function of the individual-specific effects $\alpha_i$, remains after first differencing. However, we may eliminate the truncation correction term through first differencing if we restrict our analysis to observations where $y_{it} > (\mathbf{x}_{it} - \mathbf{x}_{is})'\boldsymbol{\beta}$ and $y_{is} > -(\mathbf{x}_{it} - \mathbf{x}_{is})'\boldsymbol{\beta}$. To see this, suppose that $(\mathbf{x}_{it} - \mathbf{x}_{is})'\boldsymbol{\beta} < 0$, then

$$\begin{aligned} &E(y_{is} \mid \alpha_i, \mathbf{x}_{it}, \mathbf{x}_{is}, y_{is} > -(\mathbf{x}_{it} - \mathbf{x}_{is})'\beta) \\ &\quad = \alpha_i + \mathbf{x}_{is}'\boldsymbol{\beta} + E(u_{is} \mid u_{is} > -\alpha_i - \mathbf{x}_{is}'\boldsymbol{\beta} \\ &\qquad - (\mathbf{x}_{it} - \mathbf{x}_{is})'\boldsymbol{\beta}). \end{aligned} \tag{8.4.5}$$

Since $u_{it}$ conditional on $\mathbf{x}_i$ and $\alpha_i$ is assumed to be i.i.d.,

$$E(u_{it} \mid u_{it} > -\alpha_i - \mathbf{x}_{it}'\boldsymbol{\beta}) = E(u_{is} \mid u_{is} > -\alpha_i - \mathbf{x}_{it}'\beta). \tag{8.4.6}$$

Similarly, if $(\mathbf{x}_{it} - \mathbf{x}_{is})'\boldsymbol{\beta} > 0$,

$$\begin{aligned} &E(u_{it} \mid u_{it} > -\alpha_i - \mathbf{x}_{it}'\boldsymbol{\beta} + (\mathbf{x}_{it} - \mathbf{x}_{is})'\boldsymbol{\beta}) \\ &\quad = E(u_{it} \mid u_{it} > -\alpha_i - \mathbf{x}_{is}'\boldsymbol{\beta}) \\ &\quad = E(u_{is} \mid u_{is} > -\alpha_i - \mathbf{x}_{is}'\beta). \end{aligned} \tag{8.4.7}$$

Therefore, by confining our analysis to the truncated observations where $y_{it} > (\mathbf{x}_{it} - \mathbf{x}_{is})'\boldsymbol{\beta}$, $y_{is} > -(\mathbf{x}_{it} - \mathbf{x}_{is})'\boldsymbol{\beta}$, $y_{it} > 0$, $y_{is} > 0$, we have

$$(y_{it} - y_{is}) = (\mathbf{x}_{it} - \mathbf{x}_{is})'\boldsymbol{\beta} + (\epsilon_{it} - \epsilon_{is}), \tag{8.4.8}$$

which no longer involves incidental parameter, $\alpha_i$. Since $E[(\epsilon_{it} - \epsilon_{is}) \mid \mathbf{x}_{it}, \mathbf{x}_{is}] = 0$, applying LS to (8.4.8) will yield consistent estimator of $\boldsymbol{\beta}$.

The idea of restoring symmetry of the error terms of pairwise differencing equation $(y_{it} - y_{is})$ by throwing away observations where $y_{it} < (\mathbf{x}_{it} - \mathbf{x}_{is})'\boldsymbol{\beta}$ and $y_{is} < -(\mathbf{x}_{it} - \mathbf{x}_{is})'\boldsymbol{\beta}$ can be seen by considering the following graphs assuming that $T = 2$. Suppose that the probability density function of $u_{it}$ is of the shape shown on Figure 8.3. Since $u_{i1}$ and $u_{i2}$ are i.i.d. conditional on $(\mathbf{x}_{i1}, \mathbf{x}_{i2}, \alpha_i)$, the probability density of $y_{i1}^*$ and $y_{i2}^*$ conditional on $(\mathbf{x}_{i1}, \mathbf{x}_{i2}, \alpha_i)$ should have the same shape except for the location. The top and bottom figures of Figure 8.4 postulate the probability density of $y_{i1}^*$ and $y_{i2}^*$ conditional on $(\mathbf{x}_{i1}, x_{i2}, \alpha_i)$, respectively, assuming that $\Delta\mathbf{x}_i'\boldsymbol{\beta} < 0$, where $\Delta\mathbf{x}_i = \Delta\mathbf{x}_{i2} = \mathbf{x}_{i2} - \mathbf{x}_{i1}$. The truncated data correspond to those sample points where $y_{it}^*$ or $y_{it} > 0$. Because $\mathbf{x}_{i1}'\boldsymbol{\beta} \neq \mathbf{x}_{i2}'\boldsymbol{\beta}$, the probability density of $y_{i1}$ is different from that of $y_{i2}$. However, the probability density of $y_{i1}^*$ given $y_{i1}^* > -\Delta\mathbf{x}_i'\boldsymbol{\beta}$ (or $y_{i1}$ given $y_{i1} > -\Delta\mathbf{x}_i'\boldsymbol{\beta}$) is identical to the probability density of $y_{i2}^*$ given $y_{i2}^* > 0$ (or $y_{i2}$ given $y_{i2} > 0$) as shown in Figure 8.4. Similarly, if $\Delta\mathbf{x}_i'\boldsymbol{\beta} > 0$, the probability density of $y_{i1}^*$ given $y_{i1}^* > 0$ (or $y_{i1}$ given $y_{i1} > 0$) is identical to the probability density of $y_{i2}^*$ given $y_{i2}^* > \Delta\mathbf{x}_i'\boldsymbol{\beta}$ as shown in Figure 8.5.[7] In other words, in a two-dimensional diagram of $(y_{i1}^*, y_{i2}^*)$ of Figure 8.6 or 8.7, $(y_{i1}^*, y_{i2}^*)$ conditional on $(\mathbf{x}_{i1}, \mathbf{x}_{i2}, \alpha_i)$ is symmetrically distributed around the 45-degree line through $(\mathbf{x}_{i1}'\boldsymbol{\beta} + \alpha_i, \mathbf{x}_{i2}'\boldsymbol{\beta} + \alpha_i)$ or equivalently around the 45-degree line through $(\mathbf{x}_{i1}'\boldsymbol{\beta}, \mathbf{x}_{i2}'\boldsymbol{\beta})$ or $(-\Delta\mathbf{x}_i'\boldsymbol{\beta}, 0)$ as the line $LL'$. Because this is true for any value of $\alpha_i$, the same statement is true for the distribution of $(y_{i1}^*, y_{i2}^*)$ conditional on $(\mathbf{x}_{i1}, \mathbf{x}_{i2})$. When $\Delta\mathbf{x}_i'\boldsymbol{\beta} < 0$, the symmetry of the distribution of $(y_{i1}^*, y_{i2}^*)$ around $LL'$ means that the probability that $(y_{i1}^*, y_{i2}^*)$ falls in the region $A_1 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > -\Delta\mathbf{x}_i'\boldsymbol{\beta}, y_{i2}^* > y_{i1}^* + \Delta\mathbf{x}_i'\boldsymbol{\beta}\}$ equals the probability that it falls in the region $B_1 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > -\Delta\mathbf{x}_i'\boldsymbol{\beta}, 0 < y_{i2}^* < y_{i1}^* + \Delta\mathbf{x}_i'\boldsymbol{\beta}\}$. (Figure 8.6). When $\Delta\mathbf{x}_i'\boldsymbol{\beta} > 0$, the probability that $(y_{i1}^*, y_{i2}^*)$ falls in the region $A_1 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > 0, y_{i2}^* > y_{i1}^* + \Delta\mathbf{x}_i'\boldsymbol{\beta}\}$ equals the probability that it falls in the region $B_1 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > 0, \Delta\mathbf{x}_i'\boldsymbol{\beta} < y_{i2}^* < y_{i1}^* + \Delta\mathbf{x}_i'\boldsymbol{\beta}\}$. (Figure 8.7). That is, points in the regions $A_1$ and $B_1$ are not affected by the truncation. On the other hand, points falling into the region $(0 < y_{i1}^* < -\Delta\mathbf{x}_i'\boldsymbol{\beta}, y_{i2}^* > 0)$ in Figure 8.6 (correspond to points $(y_{i1} < -\Delta\mathbf{x}_i'\boldsymbol{\beta}, y_{i2})$) and $(y_{i1}^* > 0, 0 < y_{i2}^* < \Delta\mathbf{x}_i'\boldsymbol{\beta})$ in Figure 8.7 (correspond to points $(y_{i1}, y_{i2} < \Delta\mathbf{x}_i')b)$) will have to be thrown away to restore symmetry.

[7] I owe this exposition to the suggestion of J.L. Powell.

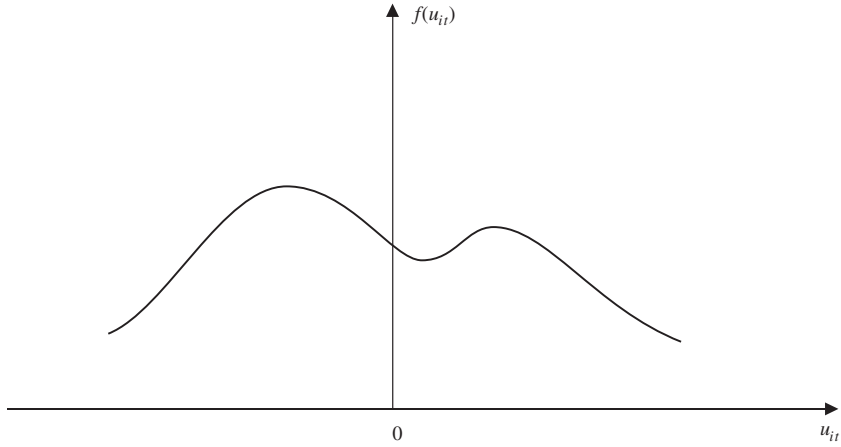


Figure 8.3. Probability density of $u_{it}$.

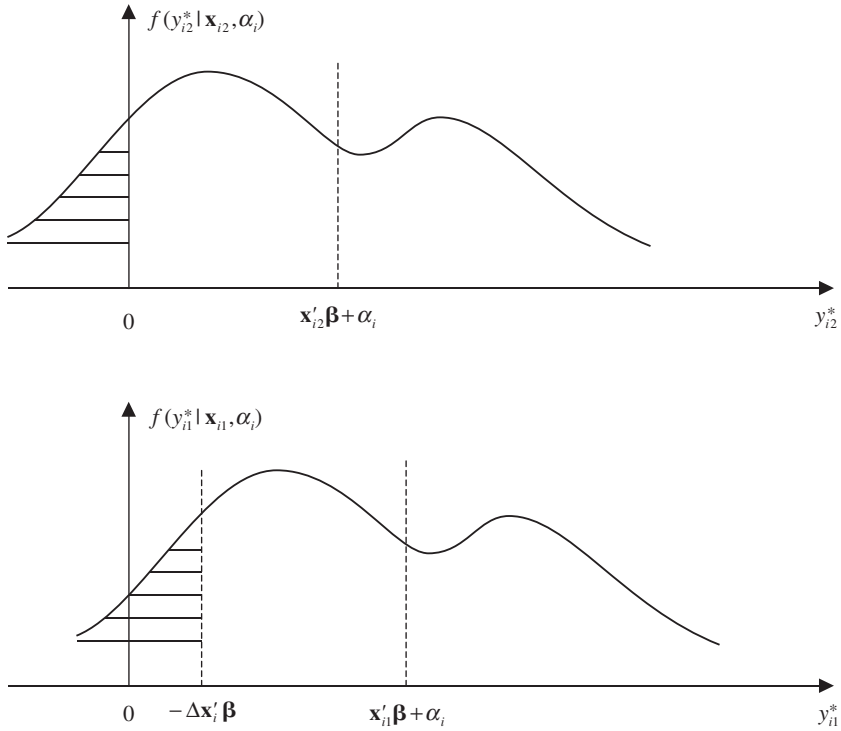


Figure 8.4. Conditional densities of $y_{i1}^*$ and $y_{i2}^*$ given $(x_{i1}, x_{i2}, \alpha_i)$, assuming $\Delta\mathbf{x}_i'\boldsymbol{\beta} < 0$.

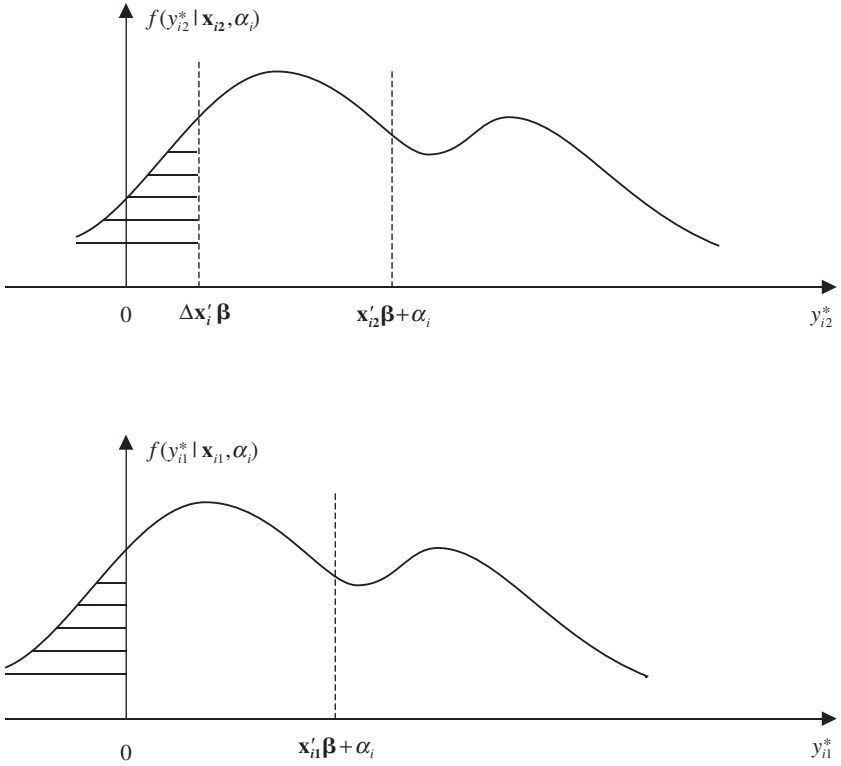


Figure 8.5. Conditional densities of $y_{i1}^*$ and $y_{i2}^*$ given $(x_{i1}, x_{i2}, \alpha_i)$, assuming $\Delta\mathbf{x}_i'\boldsymbol{\beta} > 0$.

Let $C = \{i \mid y_{i1} > -\Delta\mathbf{x}_i'\boldsymbol{\beta}, y_{i2} > \Delta\mathbf{x}_i'\boldsymbol{\beta}\}$, then $(y_{i1} - \mathbf{x}_{i1}'\boldsymbol{\beta} - \alpha_i)$ and $(y_{i2} - \mathbf{x}_{i2}'\boldsymbol{\beta} - \alpha_i)$ for $i \epsilon C$ are symmetrically distributed around 0. Therefore $E[(y_{i2} - y_{i1}) - (\mathbf{x}_{i2} - \mathbf{x}_{i1})'\boldsymbol{\beta} \mid \mathbf{x}_{i1}, \mathbf{x}_{i2}, i \epsilon C] = 0$. In other words,

$$\begin{aligned} &E[\Delta y_i - \Delta\mathbf{x}_i'\boldsymbol{\beta} \mid y_{i1} > -\Delta\mathbf{x}_i'\boldsymbol{\beta}, y_{i2} > \Delta\mathbf{x}_i'\boldsymbol{\beta}] \\ &= E[\Delta y_i - \Delta\mathbf{x}_i'\boldsymbol{\beta} \mid y_{i1}^* > 0, y_{i1}^* > -\Delta\mathbf{x}_i'\boldsymbol{\beta}, y_{i2}^* > 0, y_{i2}^* > \Delta\mathbf{x}_i'\boldsymbol{\beta}] = 0, \end{aligned} \tag{8.4.9a}$$

and

$$E[(\Delta y_i - \Delta\mathbf{x}_i'\boldsymbol{\beta})\Delta\mathbf{x}_i \mid y_{i1} > -\Delta\mathbf{x}_i'\boldsymbol{\beta}, y_{i2} > \Delta\mathbf{x}_i'\boldsymbol{\beta}] = \mathbf{0}, \tag{8.4.9b}$$

where $\Delta y_{i1} = \Delta y_{i2} = y_{i2} - y_{i1}$. However, there could be multiple roots that satisfy (8.4.9b). To ensure a unique solution for $\boldsymbol{\beta}$, Honoré (1992) suggests the trimmed LAD and LS estimators as those $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ that minimize the objective

functions

$$\begin{aligned} Q_N(\boldsymbol{\beta}) &= \sum_{i=1}^{N} [\mid \Delta y_i - \Delta \mathbf{x}_i'\boldsymbol{\beta} \mid 1\{y_{i1} > -\Delta \mathbf{x}_i'\boldsymbol{\beta},\, y_{i2} > \Delta \mathbf{x}_i'\boldsymbol{\beta}\} \\ &\quad + \mid y_{i1} \mid 1\{y_{i1} \geq -\Delta \mathbf{x}_i'\boldsymbol{\beta},\, y_{i2} < \Delta \mathbf{x}_i'\boldsymbol{\beta}\} \\ &\quad + \mid y_{i2} \mid 1\{y_{i1} < -\Delta \mathbf{x}_i'\boldsymbol{\beta},\, y_{i2} \geq \Delta \mathbf{x}_i'\boldsymbol{\beta}\}] \\ &= \sum_{i=1}^{N} \psi(y_{i1}, y_{i2}, \Delta \mathbf{x}_i'\boldsymbol{\beta}), \end{aligned} \tag{8.4.10}$$

and

$$\begin{aligned} R_N(\boldsymbol{\beta}) &= \sum_{i=1}^{N} [(\Delta y_i - \Delta \mathbf{x}_i'\boldsymbol{\beta})^2 1\{y_{i1} \geq -\Delta \mathbf{x}_i'\boldsymbol{\beta},\, y_{i2} > \Delta \mathbf{x}_i'\boldsymbol{\beta}\} \\ &\quad + y_{i1}^2 1\{y_{i1} > -\Delta \mathbf{x}_i'\boldsymbol{\beta},\, y_{i2} < \Delta \mathbf{x}_i'\boldsymbol{\beta}\} \\ &\quad + y_{i2}^2 1\{y_{i1} < -\Delta \mathbf{x}_i'\boldsymbol{\beta},\, y_{i2} > \Delta \mathbf{x}_i'\boldsymbol{\beta}\}] \\ &= \sum_{i=1}^{N} \psi(y_{i1}, y_{i2}, \Delta \mathbf{x}_i'\boldsymbol{\beta})^2, \end{aligned} \tag{8.4.11}$$

respectively. The function $\psi(w_1, w_2, c)$ is defined for $w_1 > 0$ and $w_2 > 0$ by

$$\psi(w_1, w_2, c) = \begin{cases} w_1 & \text{for} \quad w_2 < c, \\ (w_2 - w_1 - c) & \text{for} \quad -w_1 < c < w_2, \\ w_2 & \text{for} \quad c < -w_1, \end{cases}$$

is convex in $c$. The first-order conditions of (8.4.10) and (8.4.11) are the sample analogs of

$$\begin{aligned} E\{P(y_{i1} > -\Delta \mathbf{x}_i'\boldsymbol{\beta},\, y_{i2} > y_{i1} + \Delta \mathbf{x}_i'\boldsymbol{\beta}) - P(y_{i1} > -\Delta \mathbf{x}_i'\boldsymbol{\beta}, \\ \Delta \mathbf{x}_i'\boldsymbol{\beta} < y_{i2} < y_{i1} + \Delta \mathbf{x}_i'\boldsymbol{\beta})]\Delta \mathbf{x}_i\} = \mathbf{0}, \end{aligned} \tag{8.4.12}$$

and

$$\begin{aligned} E\{(\Delta y_i - \Delta \mathbf{x}_i'\boldsymbol{\beta})\Delta \mathbf{x}_i \mid (y_{i1} > -\Delta \mathbf{x}_i'\boldsymbol{\beta},\, \mathbf{y}_{i2} > y_{i1} + \Delta \mathbf{x}_i'\boldsymbol{\beta}) \\ \cup (y_{i1} > -\Delta \mathbf{x}_i'\boldsymbol{\beta},\, \Delta \mathbf{x}_i'\boldsymbol{\beta} < y_{i2} < y_{i1} + \Delta \mathbf{x}_i'\boldsymbol{\beta})\} = 0, \end{aligned} \tag{8.4.13}$$

respectively. Honoré (1992) proves that $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ are consistent and asymptotically normally distributed if the density of $u$ is strictly log-concave. The asymptotic covariance matrix of $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ and $\sqrt{N}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$ may be approximated by

$$\text{Asy Cov}\left(\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right) = \Gamma_1^{-1} V_1 \Gamma_1^{-1}, \tag{8.4.14}$$

and

$$\text{Asy Cov}\left(\sqrt{N}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})\right)=\Gamma_2^{-1}V_2\Gamma_2^{-1}, \tag{8.4.15}$$

where $V_1$, $V_2$, $\Gamma_1$, and $\Gamma_2$ may be approximated by

$$\hat{V}_1=\frac{1}{N}\sum_{i=1}^{N}1\{-y_{i1}<\Delta\mathbf{x}_i'\hat{\boldsymbol{\beta}}<y_{i2}\}\Delta\mathbf{x}_i\Delta\mathbf{x}_i', \tag{8.4.16}$$

$$\hat{V}_2=\frac{1}{N}\sum_{i=1}^{N}1\{-y_{i1}<\Delta\mathbf{x}_i'\tilde{\boldsymbol{\beta}}<y_{i2}\}(\Delta y_i-\Delta\mathbf{x}_i'\tilde{\boldsymbol{\beta}})^2\Delta\mathbf{x}_i\Delta\mathbf{x}_i', \tag{8.4.17}$$

$$\begin{aligned}\hat{\Gamma}_1^{(j,k)}=\frac{1}{h_N}[&\frac{1}{N}\sum_{i=1}^{N}(1\{\Delta y_i<\Delta\mathbf{x}_i'(\hat{\boldsymbol{\beta}}+h_N\mathbf{i}_k)<y_{i2}\}\\&-1\{-y_{i1}<\Delta\mathbf{x}_i(\hat{\boldsymbol{\beta}}+h_N\mathbf{i}_k)<\Delta y_i\})\Delta\mathbf{x}_i^{(j)}\\&+\frac{1}{N}\sum_{i=1}^{N}(-1\{\Delta y_i<\Delta\mathbf{x}_i'\hat{\boldsymbol{\beta}}<y_{i2}\}\\&-1\{-y_{i1}<\Delta\mathbf{x}_i'\hat{\boldsymbol{\beta}}<\Delta y_i\})\Delta\mathbf{x}_i^{(j)}],\end{aligned} \tag{8.4.18}$$

$$\begin{aligned}\hat{\Gamma}_2^{(j,k)}=\frac{1}{h_N}[&\frac{1}{N}\sum_{i=1}^{N}\{-y_{i1}<\Delta\mathbf{x}_i'(\tilde{\boldsymbol{\beta}}+h_N\mathbf{i}_k)<y_{i2}\}\\&\times\left(\Delta\mathbf{y}_i-\Delta\mathbf{x}_i'(\tilde{\boldsymbol{\beta}}+h_N\mathbf{i}_k)\right)\Delta\mathbf{x}_i^{(j)}\\&-\frac{1}{N}\sum_{i=1}^{N}1\{-y_{i1}<\Delta\mathbf{x}_i'\tilde{\boldsymbol{\beta}}<y_{i2}\}(\Delta y_i-\Delta\mathbf{x}_i'\tilde{\boldsymbol{\beta}})\Delta\mathbf{x}_i^{(j)}],\end{aligned} \tag{8.4.19}$$

where $\Gamma_\ell^{(j,k)}$ denotes the $(j,k)$th element of $\Gamma_\ell$, for $\ell=1,2$, $\Delta\mathbf{x}_i^{(j)}$ denotes the $j$th coordinate of $\Delta\mathbf{x}_i$, $\mathbf{i}_k$ is a unit vector with 1 in its $k$th place and $h_N$ decreases to 0 with the speed of $N^{-\frac{1}{2}}$. The bandwidth factor $h_N$ appears in (8.4.18) and (8.4.19) because $\Gamma_\ell$ is a function of densities and conditional expectations of $y$ (Honoré 1992).

### *8.4.1.2 Censored Regression*

When data are censored, observations $\{y_{it},\mathbf{x}_{it}\}$ are available for $i=1,\ldots,N$, $t=1,\ldots,T$, where $y_{it}=\max\{0,y_{it}^*\}$. In other words, $y_{it}$ can now be either 0 or a positive number rather than just a positive number as in the case of truncated data. Of course, we can throw away observations of $(y_{it},\mathbf{x}_{it})$ that correspond to $y_{it}=0$ and treat the censored regression model as the truncated regression model using the methods of Section 8.4.1a. But this will lead to

a loss of information. In the case that data are censored, in addition to the relation (8.4.9a,b), the joint probability of $y_{i1} \leq -\boldsymbol{\beta}'\Delta\mathbf{x}_i$ and $y_{i2} > 0$ is identical to the joint probability of $y_{i1} > -\boldsymbol{\beta}'\Delta\mathbf{x}_i$ and $y_{i2} = 0$, when $\boldsymbol{\beta}'\Delta\mathbf{x}_i < 0$ as shown in Figure 8.6, region $A_2$ and $B_2$, respectively. When $\boldsymbol{\beta}'\Delta\mathbf{x}_i > 0$, the joint probability of $y_{i1} = 0$ and $y_{i2} > \boldsymbol{\beta}'\Delta\mathbf{x}_i$ is identical to the joint probability of $y_{i1} > 0$ and $y_{i2} \leq \boldsymbol{\beta}'\Delta\mathbf{x}_i$ as shown in Figure 8.7. In other words, $(y_{i1}^*, y_{i2}^*)$ conditional on $(\mathbf{x}_{i1}, \mathbf{x}_{i2}, \alpha_i)$ is symmetrically distributed around the 45-degree line through $(\mathbf{x}_{i1}'\boldsymbol{\beta} + \alpha_i, \mathbf{x}_{i2}'\boldsymbol{\beta} + \alpha_i)$ or equivalently around the 45-degree line through $(-\Delta\mathbf{x}_i'\boldsymbol{\beta}, 0)$ as the line $LL'$ in Figure 8.6 or 8.7. Because this is true for any value of $\alpha_i$, the same statement is true for the distribution of $(y_{i1}^*, y_{i2}^*)$ conditional on $(\mathbf{x}_{i1}, \mathbf{x}_{i2})$. When $\Delta\mathbf{x}_i'\boldsymbol{\beta} < 0$, the symmetry of the distribution of $(y_{i1}^*, y_{i2}^*)$ around $LL'$ means that the probability that $(y_{i1}^*, y_{i2}^*)$ falls in the region $A_1 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > -\Delta\mathbf{x}_i'\boldsymbol{\beta}, y_{i2}^* > y_{i1}^* + \Delta\mathbf{x}_i'\boldsymbol{\beta}\}$ equals the probability that it falls in the region $B_1 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > -\Delta\mathbf{x}_i'\boldsymbol{\beta}, 0 < y_{i2}^* < y_{i1}^* + \Delta\mathbf{x}_i'\boldsymbol{\beta}\}$. Similarly, the probability that $(y_{i1}^*, y_{i2}^*)$ falls in the region $A_2 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* < -\Delta\mathbf{x}_i'\boldsymbol{\beta}, y_{i2}^* > 0\}$ equals the probability that it falls in the region $B_2 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > -\Delta\mathbf{x}_i'\boldsymbol{\beta}, y_{i2}^* \leq 0\}$ as shown in Figure 8.6. When $\Delta\mathbf{x}_i'\boldsymbol{\beta} > 0$, the probability that $(y_{i1}^*, y_{i2}^*)$ falls in the region $A_1 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > 0, y_{i2}^* > y_{i1}^* + \Delta\mathbf{x}_i'\boldsymbol{\beta}\}$ equals the probability that it falls in the region $B_1 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > 0, \Delta\mathbf{x}_i'\boldsymbol{\beta} < y_{i2}^* < y_{i1}^* + \Delta\mathbf{x}_i'\boldsymbol{\beta}\}$ and the probability that it falls in the region $A_2 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* \leq 0, y_{i2}^* > \Delta\mathbf{x}_i'\boldsymbol{\beta}\}$ equals the probability that it falls in the region $B_2 = \{(y_{i1}^*, y_{i2}^*) : y_{i1}^* > 0, y_{i2}^* \leq \Delta\mathbf{x}_i'\boldsymbol{\beta}\}$ as in Figure 8.7. Therefore, the probability of $(y_{i1}^*, y_{i2}^*)$ conditional on $(\mathbf{x}_{i1}, \mathbf{x}_{i2})$ falling in $A = (A_1 \cup A_2)$ equals the probability that it falls in $B = (B_1 \cup B_2)$. As neither of these probabilities is affected by censoring, the same is true in the censored sample. This implies that

$$E\left[(1\{(y_{i1}, y_{i2}) \epsilon A\} - 1\{(y_{i1}, y_{i2}) \epsilon B\})\Delta\mathbf{x}_i\right] = \mathbf{0}. \tag{8.4.20}$$

In other words, to restore symmetry of censored observations around their expected values, observations correspond to $(y_{i1} = 0, y_{i2} < \Delta\mathbf{x}_i'\boldsymbol{\beta})$ or $(y_{i1} < -\Delta\mathbf{x}_i'\boldsymbol{\beta}, y_{i2} = 0)$ will have to be thrown away.

By the same argument, conditional on $(\mathbf{x}_{i1}, \mathbf{x}_{i2})$ the expected vertical distance from a $(y_{i1}, y_{i2})$ in $A$ to the boundary of $A$ equals the expected horizontal distance from a $(y_{i1}, y_{i2})$ in $B$ to the boundary of $B$. For $(y_{i1}, y_{i2})$ in $A_1$, the vertical distance to $LL'$ is $(\Delta y_i - \Delta\mathbf{x}_i'\boldsymbol{\beta})$. For $(y_{i1}, y_{i2})$ in $B_1$, the horizontal distance to $LL'$ is $y_{i1} - (y_{i2} - \Delta\mathbf{x}_i'\boldsymbol{\beta}) = -(\Delta y_i - \Delta\mathbf{x}_i'\boldsymbol{\beta})$. For $(y_{i1}, y_{i2})$ in $A_2$, the vertical distance to the boundary of $A_2$ is $y_{i2} - \max(0, \Delta\mathbf{x}_i'\boldsymbol{\beta})$. For $(y_{i1}, y_{i2})$ in $B_2$, the horizontal distance is $y_{i1} - \max(0, -\Delta\mathbf{x}_i'\boldsymbol{\beta})$. Therefore

$$\begin{aligned} E\Big[\Big(&1\{(y_{i1}, y_{i2}) \epsilon A_1\}(\Delta y_i - \Delta\mathbf{x}_i'\boldsymbol{\beta}) + 1\{(y_{i1}, y_{i2}) \epsilon A_2)\}(y_{i2} - \max(0, \Delta\mathbf{x}_i'\boldsymbol{\beta})) \\ &+ 1\{(y_{i1}, y_{i2}) \epsilon B_1\}(\Delta y_i - \Delta\mathbf{x}_i'\boldsymbol{\beta}) - 1\{y_{i1}, y_{i2} \epsilon B_2\}(y_{i1} \\ &- \max(0, -\Delta\mathbf{x}_i'\boldsymbol{\beta}))\Big)\Delta\mathbf{x}_i\Big] = \mathbf{0}. \end{aligned} \tag{8.4.21}$$

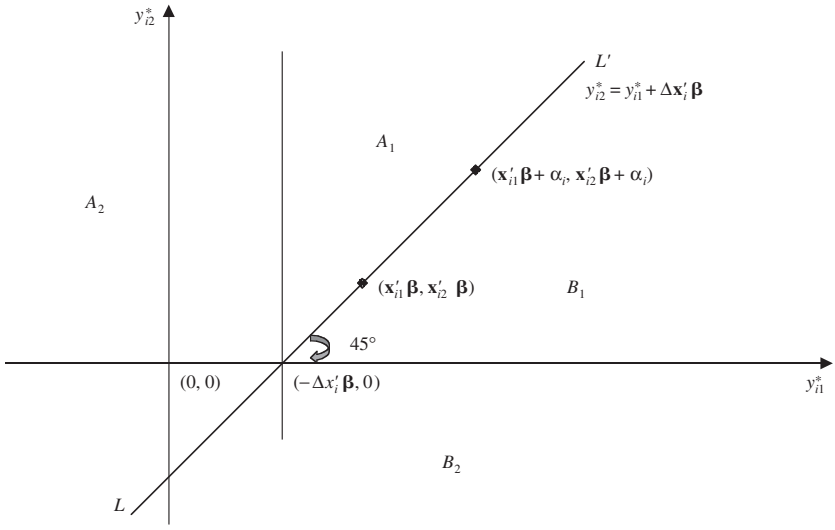


Figure 8.6. The distribution of $(y^*_{i1}, y^*_{i2})$ assuming $\Delta \mathbf{x}'_i \boldsymbol{\beta} < 0$.
$A_1 = \{(y^*_{i1}, y^*_{i2}) : y^*_{i1} > -\Delta \mathbf{x}'_i \boldsymbol{\beta},\ y^*_{i2} > y^*_{i1} + \Delta \mathbf{x}'_i \boldsymbol{\beta}\}$, $A_2 = \{(y^*_{i1}, y^*_{i2}) :$
$y^*_{i1} \leq -\Delta \mathbf{x}'_i \boldsymbol{\beta},\ y^*_{i2} > 0\}$,
$B_1 = \{(y^*_{i1}, y^*_{i2}) : y^*_{i1} > -\Delta \mathbf{x}'_i \boldsymbol{\beta},\ 0 < y^*_{i2} < y^*_{i1} + \Delta \mathbf{x}'_i \boldsymbol{\beta}\}$, $B_2 = \{(y^*_{i1}, y^*_{i2}) :$
$y^*_{i1} > -\Delta \mathbf{x}'_i \boldsymbol{\beta},\ y^*_{i2} \leq 0\}$.

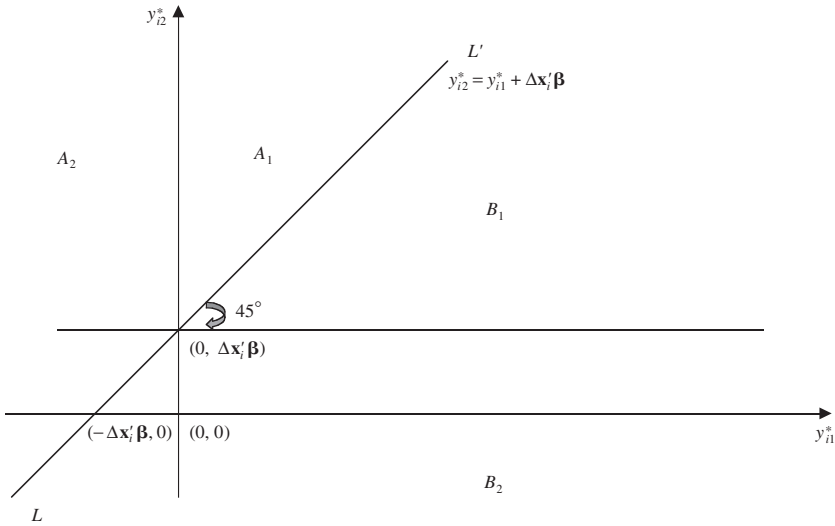


Figure 8.7. The distribution of $(y^*_{i1}, y^*_{i2})$ assuming $\Delta \mathbf{x}'_i \boldsymbol{\beta} > 0$.
$A_1 = \{(y^*_{i1}, y^*_{i2}) : y^*_{i1} > 0,\ y^*_{i2} > y^*_{i1} + \Delta \mathbf{x}'_i \boldsymbol{\beta}\}$, $A_2 = \{(y^*_{i1}, y^*_{i2}) :$
$y^*_{i1} \leq 0,\ y^*_{i2} > \Delta \mathbf{x}'_i \boldsymbol{\beta}\}$,
$B_1 = \{(y^*_{i1}, y^*_{i2}) : y^*_{i1} > 0,\ \Delta \mathbf{x}'_i \boldsymbol{\beta} < y^*_{i2} < y^*_{i1} + \Delta \mathbf{x}'_i \boldsymbol{\beta}\}$, $B_2 = \{(y^*_{i1}, y^*_{i2}) :$
$y^*_{i1} > 0,\ y^*_{i2} \leq \Delta \mathbf{x}'_i \boldsymbol{\beta}\}$.

The pairwise trimmed LAD and LS estimators, $\hat{\boldsymbol{\beta}}^*$ and $\tilde{\boldsymbol{\beta}}^*$, for the estimation of the censored regression model proposed by Honoré (1992) are obtained by minimizing the objective functions

$$\begin{aligned} Q_N^*(\boldsymbol{\beta}) &= \sum_{i=1}^{N} \Big[1 - 1\{y_{i1} \le -\Delta\mathbf{x}_i'\boldsymbol{\beta},\, y_{i2} \le 0\}\Big]\Big[1 - 1\{y_{i2} \le \Delta\mathbf{x}_i'\boldsymbol{\beta},\, y_{i1} \le 0\}\Big] \\ &\quad \cdot \mid \Delta y_i - \Delta\mathbf{x}_i'\boldsymbol{\beta} \mid \qquad (8.4.22) \\ &= \sum_{i=1}^{N} \psi^*(y_{i1}, y_{i2}, \Delta\mathbf{x}_i\boldsymbol{\beta}), \end{aligned}$$

$$\begin{aligned} R_N^*(\boldsymbol{\beta}) &= \sum_{i=1}^{N} \Big\{ \big[\max\{y_{i2}, \Delta\mathbf{x}_i'\boldsymbol{\beta}\} - \max\{y_{i1}, -\Delta\mathbf{x}_i'\boldsymbol{\beta}\} - \Delta\mathbf{x}_i'\boldsymbol{\beta}\big]^2 \\ &\quad - 2 \cdot 1\{y_{i1} < -\Delta\mathbf{x}_i'\boldsymbol{\beta}\}(y_{i1} + \Delta\mathbf{x}_i'\boldsymbol{\beta})y_{i2} \\ &\quad - 2 \cdot 1\{y_{i2} < \Delta\mathbf{x}_i'\boldsymbol{\beta}\}(y_{i2} - \Delta\mathbf{x}_i'\boldsymbol{\beta})y_{i1}\Big\} \qquad (8.4.23) \\ &= \sum_{i=1}^{N} \chi(y_{i1}, y_{i2}, \Delta\mathbf{x}_i'\boldsymbol{\beta}), \end{aligned}$$

where

$$\psi^*(w_1, w_2, c) = \begin{cases} 0, & \text{for} \quad w_1 \le \max\{0, -c\} \quad \text{and} \quad w_2 \le \max\{0, c\}, \\ \mid w_2 - w_1 - c \mid, & \text{otherwise} \end{cases}$$

and

$$\chi(w_1, w_2, c) = \begin{cases} w_1^2 - 2w_1(w_2 - c) & \text{for} \quad w_2 \le c, \\ (w_2 - w_1 - c)^2 & \text{for} \quad -w_1 < c < w_2, \\ w_2^2 - 2w_2(c + w_1) & \text{for} \quad c \le -w_1, \end{cases}$$

which is convex in $c$. The first-order conditions of (8.4.22) and (8.4.23) are the sample analogs of (8.4.20) and (8.4.21), respectively. For instance, when $(y_{i1}, y_{i2}) \in (A_1 \cup B_1)$, the corresponding terms in $R_N^*$ become $(\Delta y_i - \Delta\mathbf{x}_i'\boldsymbol{\beta})^2$. When $(y_{i1}, y_{i2}) \in A_2$, the corresponding terms become $y_{i2}^2 - 2 \times 1\{y_{i1} < -\Delta\mathbf{x}_i'\boldsymbol{\beta}\}\ (y_{i1} + \Delta\mathbf{x}_i'\boldsymbol{\beta})y_{i2}$. When $(y_{i1}, y_{i2}) \in B_2$, the corresponding terms become $y_{i1}^2 - 2 \times 1\{y_{i2} < \Delta\mathbf{x}_i'\boldsymbol{\beta}\}(y_{i2} - \Delta\mathbf{x}_i'\boldsymbol{\beta})y_{i1}$. The partial derivatives of the first term with respect to $\boldsymbol{\beta}$ converges to $E\{[1\{(y_{i1}, y_{i2}) \in A_1\}(\Delta y_i - \Delta\mathbf{x}_i'\beta) + 1\{(y_{i1}, y_{i2} \in B_1\}(\Delta y_i - \Delta\mathbf{x}_i'\boldsymbol{\beta})]\Delta\mathbf{x}_i\}$. The partial derivatives of the second and third terms with respect to $\boldsymbol{\beta}$ yield $-2E\{1[(y_{i1}, y_{i2}) \in A_2]y_{i2}\Delta\mathbf{x}_i - 1[(y_{i1}, y_{i2}) \in B_2]y_{i1}\Delta\mathbf{x}_i\}$. Because $Q_N^*(\boldsymbol{\beta})$ is piecewise linear and convex and $R_N^*(\boldsymbol{\beta})$ is continuously differentiable and convex and twice differentiable except

at a finite number of points, the censored pairwise trimmed LAD and LS estimators, $\hat{\boldsymbol{\beta}}^*$ and $\tilde{\boldsymbol{\beta}}^*$, are computationally simpler than the truncated estimators $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$.

Honoré (1992) shows that $\hat{\boldsymbol{\beta}}^*$ and $\tilde{\boldsymbol{\beta}}^*$ are consistent and asymptotically normally distributed. The asymptotic covariance matrix of $\sqrt{N}(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta})$ is equal to

$$\text{Asy. Cov}\,(\sqrt{N}(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta})) = \Gamma_3^{-1} V_3 \Gamma_3^{-1}, \tag{8.4.24}$$

and of $\sqrt{N}(\tilde{\boldsymbol{\beta}}^* - \boldsymbol{\beta})$ is equal to

$$\text{Asy. Cov}\,(\sqrt{N}(\tilde{\boldsymbol{\beta}}^* - \boldsymbol{\beta})) = \Gamma_4^{-1} V_4 \Gamma_4^{-1}, \tag{8.4.25}$$

where $V_3$, $V_4$, $\Gamma_3$, and $\Gamma_4$ may be approximated by

$$\begin{aligned}\hat{V}_3 = \frac{1}{N}\sum_{i=1}^{N} 1\Big\{ &\Big[\Delta\mathbf{x}_i'\hat{\boldsymbol{\beta}}^* < \Delta y_i,\, y_{i2} > \max\,(0,\, \Delta\mathbf{x}_i'\hat{\boldsymbol{\beta}}^*)\Big] \\ &\cup \Big[\Delta y_i < \Delta\mathbf{x}_i'\hat{\boldsymbol{\beta}}^*,\, y_{i1} > \max\,(0,\, -\Delta\mathbf{x}_i'\hat{\boldsymbol{\beta}}^*)\Big]\Big\}\Delta\mathbf{x}_i\Delta\mathbf{x}_i',\end{aligned} \tag{8.4.26}$$

$$\begin{aligned}\hat{V}_4 = \frac{1}{N}\sum_{i=1}^{N} \Big[ &y_{i2}^2 1\{\Delta\mathbf{x}_i'\tilde{\boldsymbol{\beta}}^* \le -y_{i1}\} + y_{i1}^2 1\{y_{i2} \le \Delta\mathbf{x}_i'\tilde{\boldsymbol{\beta}}^*\} \\ &+ (\Delta y_i - \Delta\mathbf{x}_1'\tilde{\beta}^*)^2 1\{-y_{i1} < \Delta\mathbf{x}_i'\tilde{\boldsymbol{\beta}}^* < y_{i2}\}\Big]\Delta\mathbf{x}_i\Delta\mathbf{x}_i',\end{aligned} \tag{8.4.27}$$

$$\begin{aligned}\hat{\Gamma}_3^{(j,k)} = \frac{-1}{h_N}\Bigg\{ &\frac{1}{N}\sum_{i=1}^{N}\Big[1\{y_{i2} > 0,\, y_{i2} > y_{i1} + \Delta\mathbf{x}_i'(\hat{\boldsymbol{\beta}}^* + h_N\mathbf{i}_k)\} \\ &- 1\{y_{i1} > 0,\, y_{i1} > y_{i2} - \Delta\mathbf{x}_i'(\hat{\boldsymbol{\beta}}^* + h_N\mathbf{i}_k)\}\Big]\Delta\mathbf{x}_i^{(j)} \\ &- \frac{1}{N}\sum_{i=1}^{N}\Big[1\{y_{i2} > 0,\, y_{i2} > y_{i1} + \Delta\mathbf{x}_i'\hat{\boldsymbol{\beta}}^*\} \\ &- 1\{y_{i1} > 0,\, y_{i1} > y_{i2} - \Delta\mathbf{x}_i\hat{\boldsymbol{\beta}}^*\}\Big]\Delta\mathbf{x}_i^{(j)}\Bigg\},\end{aligned} \tag{8.4.28}$$

and

$$\hat{\Gamma}_4 = \frac{1}{N}\sum_{i=1}^{N} 1\{-y_{i1} < \Delta\mathbf{x}_i'\tilde{\boldsymbol{\beta}}^* < y_{i2}\}\Delta\mathbf{x}_i\Delta\mathbf{x}_i'. \tag{8.4.29}$$

where $\mathbf{i}_k$ is a unit vector with 1 in its $k$th place and $h_N$ decreases to 0 at the speed of $N^{-\frac{1}{2}}$.

Both the truncated and censored estimators are presented assuming that $T = 2$. They can be easily modified to cover the case where $T > 2$. For instance,

(8.4.23) can be modified to be the estimator

$$\tilde{\boldsymbol{\beta}}^* = \arg\min \sum_{i=1}^{N} \sum_{t=2}^{T} \chi(y_{i,t-1}, y_{it}, (\mathbf{x}_{it} - \mathbf{x}_{it-1})\boldsymbol{\beta}) \tag{8.4.30}$$

when $T > 2$.

### 8.4.2 A Semiparametric Two-Step Estimator for the Endogenously Determined Sample Selection Model

In this subsection we consider the estimation of the endogenously determined sample selection model in which the sample selection rule is determined by the binary response model (8.3.4) or (8.1.22) for the linear regression model (8.3.1) where $y_{it} = y_{it}^*$ if $d_{it} = 1$ and unknown if $d_{it} = 0$ as in (8.1.24). We assume that both (8.3.1) and (8.3.4) contain unobserved fixed individual-specific effects $\alpha_i$ and $\eta_i$ that may be correlated with the observed explanatory variables in an arbitrary way. Following the spirit of Heckman (1979) two-step estimation procedure for the parametric model, Kyriazidou (1997) proposes a two-step semiparametric method for estimating the main regression of interest (8.3.4). In the first step, the unknown coefficients of the "selection" equation (8.3.4), $\mathbf{a}$, are consistently estimated by some semiparametric method. In the second step, these estimates are substituted into the equation of interest (8.3.1) conditional on $d_{it} = 1$ and estimate it by a weighted least-squares method. The fixed effect from the main equation is eliminated by taking time differences on the observed $y_{it}$. The selection effect is eliminated by conditioning time differencing of $y_{it}$ and $y_{is}$ on those observations where $\mathbf{w}_{it}'\hat{\mathbf{a}} \simeq \mathbf{w}_{is}'\hat{\mathbf{a}}$ because the magnitude of the selection effect is the same if the impact of the observed variables determining selection remains the same over time.

We note that without sample selectivity, that is, $d_{it} = 1$ for all $i$ and $t$, or if $u_{it}$ and $v_{it}$ are uncorrelated conditional on $\alpha_i$ and $\mathbf{x}_{it}$, then (8.3.1) and (8.1.24) correspond to the standard variable intercept model for panel data discussed in Chapter 3 with balanced panel or randomly missing data.[8] If $u_{it}$ and $v_{it}$ are correlated, sample selection will arise because $E(u_{it} \mid \mathbf{x}_{it}, \mathbf{w}_{it}, \alpha_i, d_{it} = 1) \neq 0$. Let $\lambda(\cdot)$ denote the conditional expectation of $u$ conditional on $d = 1$, $\mathbf{x}$, $\mathbf{w}$, $\alpha$ and $\eta$, then (8.3.1) and (8.1.24) conditional on $d_{it} = 1$ can be written as

$$y_{it} = \alpha_i + \boldsymbol{\beta}'\mathbf{x}_{it} + \lambda(\eta_i + \mathbf{w}_{it}'\mathbf{a}) + \epsilon_{it}, \tag{8.4.31}$$

where $E(\epsilon_{it} \mid \mathbf{x}_{it}, d_{it} = 1) = 0$. The form of the selection function $\lambda(\cdot)$ is derived from the joint distribution of $u$ and $v$. For instance, if $u$ and $v$ are bivariate normal, then we have the Heckman sample selection correction of $\lambda(\eta_i + \mathbf{a}'\mathbf{w}_{it}) = \dfrac{\sigma_{uv}}{\sigma_v} \dfrac{\phi\left(\frac{\eta_i + \mathbf{w}_{it}'\mathbf{a}}{\sigma_v}\right)}{\Phi\left(\frac{\eta_i + \mathbf{w}_{it}'\mathbf{a}}{\sigma_v}\right)}$. Therefore, in the presence of sample selection

[8] Linear panel data with randomly missing data will be discussed in Chapter 11, Section 11.1.

or attrition with short panels, regressing $y_{it}$ on $\mathbf{x}_{it}$ using only the observed information is invalidated by two problems – first, the presence of the unobserved effects $\alpha_i$ which introduces the incidental parameter problem and second, the "selection bias" arising from the fact that

$$E(u_{it} \mid \mathbf{x}_{it}, d_{it} = 1) = \lambda(\eta_i + \mathbf{w}'_{it}\mathbf{a}).$$

The presence of individual-specific effects in (8.3.1) is easily solved by time differencing those individuals that are observed for two time periods $t$ and $s$, that is, who have $d_{it} = d_{is} = 1$. However, the sample selectivity factors are not eliminated by time differencing. But conditional on given $i$, if $(u_{it}, v_{it})$ are stationary and $\mathbf{w}'_{it}\mathbf{a} = \mathbf{w}'_{is}\mathbf{a}$, $\lambda(\eta_i + \mathbf{w}_{it}\mathbf{a}) = \lambda(\eta_i + \mathbf{w}'_{is}\mathbf{a})$. Then the difference of (8.4.31) between $t$ and $s$ if both $y_{it}$ and $y_{is}$ are observable no longer contains the individual specific effects, $\alpha_i$, and the selection factor $\lambda(\eta_i + \mathbf{w}'_{it}\mathbf{a})$,

$$\Delta y_{its} = y_{it} - y_{is} = (\mathbf{x}_{it} - \mathbf{x}_{is})'\boldsymbol{\beta} + (\epsilon_{it} - \epsilon_{is}) = \Delta\mathbf{x}'_{its}\boldsymbol{\beta} + \Delta\boldsymbol{\epsilon}_{its}. \qquad (8.4.32)$$

As shown by Ahn and Powell (1993) if $\lambda$ is a sufficiently "smooth" function, and $\hat{\mathbf{a}}$ is a consistent estimator of $\mathbf{a}$, observations for which the difference $(\mathbf{w}_{it} - \mathbf{w}_{is})'\hat{\mathbf{a}}$ is close to 0 should have $\lambda_{it} - \lambda_{is} \simeq 0$. Therefore, Kyriazidou (1997) generalizes the pairwise difference concept of Ahn and Powell (1993) and proposes to estimate the fixed-effects sample selection models in two steps: In the first step, estimate $\mathbf{a}$ by either the Andersen (1970) and Chamberlain (1980) conditional maximum-likelihood approach or the Horowitz (1992) and Lee (1999) smoothed version of the Manski (1975) maximum score method discussed in Chapter 7. In the second step, the estimated $\hat{\mathbf{a}}$ is used to estimate $\boldsymbol{\beta}$ based on pairs of observations for which $d_{it} = d_{is} = 1$ and for which $(\mathbf{w}_{it} - \mathbf{w}_{is})'\hat{\mathbf{a}}$ is "close" to 0. This last requirement is operationalized by weighting each pair of observations with a weight that depends inversely on the magnitude of $(\mathbf{w}_{it} - \mathbf{w}_{is})'\hat{\mathbf{a}}$, so that pairs with larger differences in the selection effects receive less weight in the estimation. The Kyriazidou (1997) estimator takes the form:

$$\hat{\boldsymbol{\beta}}_K = \left\{ \sum_{i=1}^{N} \frac{1}{T_i - 1} \sum_{1 \le s < t \le T_i} (\mathbf{x}_{it} - \mathbf{x}_{is})(\mathbf{x}_{it} - \mathbf{x}_{is})' K\left[\frac{(\mathbf{w}_{it} - \mathbf{w}_{is})'\hat{\mathbf{a}}}{h_N}\right] d_{it} d_{is} \right\}^{-1}$$
$$\cdot \left\{ \sum_{i=1}^{N} \frac{1}{T_i - 1} \sum_{1 \le s < t < T_i} (\mathbf{x}_{it} - \mathbf{x}_{is})(\mathbf{y}_{it} - \mathbf{y}_{is}) K\left[\frac{(\mathbf{w}_{it} - \mathbf{w}_{is})'\hat{\mathbf{a}}}{h_N}\right] d_{it} d_{is} \right\} \qquad (8.4.33)$$

where $T_i$ denotes the number of positively observed $y_{it}$ for the $i$th individual, $K(.)$ is a kernel density function which tends to 0 as the magnitude of its argument increases, and $h_N$ is a positive constant or bandwidth that decreases to 0 as $N \longrightarrow \infty$. The effect of multiplying the kernel function $K(\cdot)$ is to give more weights to observations with $\frac{1}{h_N}(\mathbf{w}_{it} - \mathbf{w}_{is})'\hat{\mathbf{a}} \simeq 0$ and less weight to those observations that $\mathbf{w}_{it}\hat{\mathbf{a}}$ is different from $\mathbf{w}_{is}\hat{\mathbf{a}}$ so that in the limit only observations with $\mathbf{w}_{it}\mathbf{a} = \mathbf{w}'_{is}\mathbf{a}$ are used in (8.4.33). Under appropriate regularity

conditions (8.4.33) is consistent but the rate of convergence is proportional to $\sqrt{Nh_N}$, much slower than the standard square root of the sample size.

When $T = 2$, the asymptotic covariance matrix of the Kyriazidou (1997) estimator (8.4.33) may be approximated by the Eicker (1963)–White (1980) formulae of the asymptotic covariance matrix of the least-squares estimator of the linear regression model with heteroscedasticity,

$$\left(\sum_{i=1}^{N} \hat{\mathbf{x}}_i \hat{\mathbf{x}}_i'\right)^{-1} \sum_{i=1}^{N} \hat{\mathbf{x}}_i \hat{\mathbf{x}}_i' \Delta\hat{e}_i^2 \left(\sum_{i=1}^{N} \hat{\mathbf{x}}_i \hat{\mathbf{x}}_i'\right)^{-1}, \tag{8.4.34}$$

where $\hat{\mathbf{x}}_i = K\left(\frac{\Delta \mathbf{w}_i' \hat{\mathbf{a}}}{h_N}\right)^{1/2} \Delta \mathbf{x}_i (d_{i2} d_{i1})$ and $\Delta\hat{e}_i$ is the estimated residual of (8.4.32).

In the case that only a truncated sample is observed, the first stage estimation of $\hat{\mathbf{a}}$ cannot be implemented. However, a sufficient condition to ensure only observations with $\Delta \mathbf{w}_{its}' \mathbf{a} = 0$ are used is to replace $K\left[\frac{\Delta \mathbf{w}_{its} \hat{\mathbf{a}}}{h_N}\right]$ by a multivariate kernel function $K\left(\frac{\hat{\mathbf{w}}_{it} - \mathbf{w}_{is}}{h_N}\right)$ in (8.4.33). However, the speed of convergence of (8.4.33) to the true $\boldsymbol{\beta}$ will be $\sqrt{Nh_N^k}$, where $k$ denotes the dimension of $\mathbf{w}_{it}$. This is much slower speed than $\sqrt{Nh_N}$ because $h_N$ converges to 0 as $N \longrightarrow \infty$.

## 8.5 AN EXAMPLE: HOUSING EXPENDITURE

Charlier et al. (2001) use Dutch Socio-Economic Panel (SEP) 1987–89 waves to estimate the following endogenous switching regression model for the share of housing expenditure in total expenditure:

$$d_{it} = 1(\mathbf{w}_{it}' \mathbf{a} + \eta_i + \nu_{it} > 0), \tag{8.5.1}$$

$$y_{1it} = \boldsymbol{\beta}_1' \mathbf{x}_{it} + \alpha_{1i} + u_{1it}, \text{ if } d_{it} = 1, \tag{8.5.2}$$

$$y_{2it} = \boldsymbol{\beta}_2' \mathbf{x}_{it} + \alpha_{2i} + u_{2it}, \text{ if } d_{it} = 0, \tag{8.5.3}$$

where $d_{it}$ denotes the tenure choice between owning and renting, with 1 for owners and 0 for renters; $y_{1it}$ and $y_{2it}$ are the budget shares spent on housing for owners and renters, respectively; $\mathbf{w}_{it}$ and $\mathbf{x}_{it}$ are vectors of explanatory variables; $\eta_i$, $\alpha_{1i}$, and $\alpha_{2i}$ are unobserved household specific effects; and $\nu_{it}$, $u_{1it}$, and $u_{2it}$ are the error terms. The budget share spent on housing is defined as the fraction of total expenditure spent on housing. Housing expenditure for renters is just the rent paid by a family. The owner's expenditure on housing consists of net interest costs on mortgages, net rent paid if the land is not owned, taxes on owned housing, costs of insuring the house, opportunity cost of housing equity (which is set at 4% of the value of house minus the mortgage value), and maintenance cost, minus the increase of the value of the house. The explanatory variables considered are the education level of the head of household (DOP), age of the head of the household (AGE), age squared (AGE2), marital status (DMAR), logarithm of monthly family income (LINC), its square (L2INC),

monthly total family expenditure (EXP), logarithm of monthly total family expenditure (LEXP), its square (L2EXP), number of children (NCH), logarithm of constant quality price of rental housing (LRP), logarithm of constant quality price of owner occupied housing after tax (LOP), and LRP-LOP. The variables that are excluded from the tenure choice equation (8.5.1) are DOP, LEXP, L2EXP, LRP, and LOP. The variables excluded from the budget share equations ((8.5.2) and (8.5.3)) are DOP, LINC, L2INC, EXP, NCH, and LRP-LOP.

The random-effects and fixed-effects models with and without selection are estimated. However, because $\mathbf{x}$ include LEXP and L2EXP and they could be endogenous, Charlier, Melenberg, and van Soest (2001) also estimate this model by the instrumental variable (IV) method. For instance, the Kyriazidou (1997) weighted least-squares estimator is modified as:

$$\hat{\boldsymbol{\beta}}_{KN} = \left\{ \sum_{i=1}^{N} \sum_{1 \le s < t \le T_i} (\mathbf{x}_{it} - \mathbf{x}_{is})(\mathbf{z}_{it} - \mathbf{z}_{is})' K \left[ \frac{(\mathbf{w}_{it} - \mathbf{w}_{is})'\hat{\mathbf{a}}}{h_N} \right] d_{it} d_{is} \right\}^{-1}$$
$$\cdot \left\{ \sum_{i=1}^{N} \sum_{1 \le s < t \le T_i} (\mathbf{z}_{it} - \mathbf{z}_{is})(\mathbf{y}_{it} - \mathbf{y}_{is}) K \left[ \frac{(\mathbf{w}_{it} - \mathbf{w}_{is})'\hat{\mathbf{a}}}{h_N} \right] d_{it} d_{is} \right\}, \tag{8.5.4}$$

to take account of the potential endogeneity issue of LEXP and L2EXP, where $\mathbf{z}_{it}$ is a vector of instruments.

Tables 8.2 and 8.3 present the fixed-effects and random-effects estimation results for the budget share equations without and with correction for selection, respectively. The Kyriazidou (1997) estimator is based on the first-stage logit estimation of the tenure choice equation (8.5.1). The random-effects estimator is based on Newey (2009) series expansion method (Charlier, Melenberg, and van Soest 2000). The differences among these different formulations are quite substantial. For instance, the parameters related to AGE, AGE2, LEXP, L2EXP, and the prices are substantially different from their random effects counterparts based on IV. They also lead to very different conclusions on the elasticities of interest. The price elasticities for the average renters and owners are about $-0.5$ in the random-effects model, but are close to $-1$ for owners and $-0.8$ for renters in the fixed-effects models.

The Hausman type specification tests of endogeneity of LEXP and L2EXP are inconclusive. But a test for the presence of selectivity bias based on the difference between the Kyriazidou IV and linear panel data estimates have test statistics of 88.2 for owners and 23.7 for renters, which are significant at the 5 percent level for the $\chi^2$ distribution with 7 degrees of freedom. This indicates that the model that does not allow for correlation between the error terms in the share equations ((8.5.2) and (8.5.3)) and the error term or fixed effect in the selection equation (8.5.1) is probably misspecified.

The Hausman (1978) type specification test of no correlation between the household specific effects and the $\mathbf{x}$'s based on the difference between the

Table 8.2. *Estimation results for the budget share equations without correction for selection (standard errors in parentheses)*[a]

| Variable | Pooled random effects | Pooled IV random effects | Linear model fixed effects | Linear model IV[b] fixed effects |
|---|---|---|---|---|
| *Owners* | | | | |
| Constant | 4.102** (0.238) | 4.939** (0.712) | | |
| AGE | 0.045** (0.009) | 0.029** (0.010) | −0.073 (0.041) | −0.063 (0.044) |
| AGE2 | −0.005** (0.001) | −0.003** (0.001) | 0.009** (0.004) | 0.009* (0.004) |
| LEXP | −0.977** (0.059) | −1.271** (0.178) | −0.769** (0.049) | −1.345** (0.269) |
| L2EXP | 0.052** (0.003) | 0.073** (0.011) | 0.036** (0.003) | 0.070** (0.016) |
| DMAR | 0.036** (0.004) | 0.027** (0.005) | | |
| Dummy87 | | | −0.001 (0.003) | −0.000 (0.004) |
| Dummy88 | | | −0.002 (0.001) | −0.001 (0.002) |
| LOP | 0.068** (0.010) | 0.108** (0.010) | 0.065** (0.016) | 0.050** (0.018) |
| *Renters* | | | | |
| Constant | 2.914** (0.236) | 3.056** (0.421) | | |
| AGE | 0.038** (0.007) | 0.027** (0.007) | 0.114** (0.034) | 0.108** (0.035) |
| AGE2 | −0.004** (0.000) | −0.003** (0.001) | −0.009* (0.004) | −0.009* (0.004) |
| LEXP | −0.772** (0.055) | −0.820** (0.106) | −0.800** (0.062) | −0.653** (0.219) |
| L2EXP | 0.040** (0.003) | 0.045** (0.006) | 0.039** (0.004) | 0.031* (0.014) |
| DMAR | 0.011** (0.002) | 0.001** (0.003) | | |
| Dummy87 | | | −0.004 (0.003) | −0.003 (0.003) |
| Dummy88 | | | −0.002 (0.002) | −0.002 (0.002) |
| LRP | 0.119* (0.017) | 0.112** (0.017) | 0.057** (0.020) | 0.060** (0.020) |

[a] * means significant at the 5 percent level; ** means significant at the 1 percent level.
[b] In IV estimation AGE, AGE2, LINC, L2INC, Dummy87, Dummy88, and either LOP (for owners) or LRP (for renters) are used as instruments.
*Source:* Charlier, Melenberg, and van Soest (2001, Table 3).

Table 8.3. *Estimation results for the budget share equations using panel data models taking selection into account (standard errors in parentheses)*[a]

| Variable | Pooled random effects[b] | Pooled IV random effects[c] | Kyriazidou OLS estimates | Kyriazidou IV[d] estimates |
|---|---|---|---|---|
| *Owners* | | | | |
| Constant | 2.595[e] | 3.370[e] | | |
| AGE | −0.040** (0.013) | −0.020 (0.015) | 0.083 (0.083) | 0.359** (0.084) |
| AGE2 | 0.004** (0.001) | 0.002 (0.001) | −0.008 (0.008) | −0.033** (0.009) |
| LEXP | −0.594** (0.142) | −0.821 (0.814) | −0.766** (0.102) | −0.801** (0.144) |
| L2EXP | 0.026** (0.008) | 0.042 (0.050) | 0.036** (0.006) | 0.036** (0.008) |
| DMAR | 0.006 (0.007) | 0.012 (0.007) | | |
| LOP | 0.126** (0.012) | 0.121** (0.011) | 0.006 (0.030) | 0.001 (0.029) |
| Dummy87 | | | −0.006 (0.007) | −0.013 (0.007) |
| Dummy88 | | | −0.004 (0.004) | −0.008 (0.004) |
| *Renters* | | | | |
| Constant | 2.679[d] | 1.856[d] | | |
| AGE | −0.037** (0.012) | −0.027* (0.012) | 0.127* (0.051) | 0.082 (0.080) |
| AGE2 | 0.004** (0.001) | 0.003* (0.001) | −0.018** (0.006) | −0.014 (0.007) |
| LEXP | −0.601** (0.091) | −0.417 (0.233) | −0.882** (0.087) | −0.898** (0.144) |
| L2EXP | 0.027** (0.005) | 0.016 (0.015) | 0.044** (0.005) | 0.044** (0.009) |
| DMAR | −0.021** (0.005) | −0.019** (0.005) | | |
| LRP | 0.105** (0.016) | 0.106** (0.016) | 0.051 (0.028) | 0.024 (0.030) |
| Dummy87 | | | −0.024** (0.007) | −0.023 (0.013) |
| Dummy88 | | | −0.009* (0.004) | −0.012 (0.007) |

[a] * means significant at the 5 percent level; ** means significant at the 1 percent level.
[b] series approximation using single index ML probit in estimating the selection equation.
[c] IV using AGE, AGE2, LINC, L2INC, DMAR and either LOP (for owners) or LRP (for renters) as instruments.
[d] In IV estimation AGE, AGE2, LINC, L2INC, Dummy87, and Dummy88 are used as instruments.
[e] Estimates include the estimate for the constant term in the series approximation.

*Source:* Charlier, Melenberg, and van Soest (2001, Table 4).

Newey IV and the Kyriazidou IV estimates have test statistics of 232.1 for owners and 37.8 for renters. These are significant at the 5 percent level for the $\chi^2$ distribution with 5 degrees of freedom, thus rejecting the random-effects model that does not allow for correlation between the household-specific effects and the explanatory variables. These results indicate that the random-effects linear panel models or linear panel data models that allow only for very specific selection mechanisms (both of which can be estimated with just the cross-sectional data) are probably too restrictive.

## 8.6 DYNAMIC TOBIT MODELS

### 8.6.1 Dynamic Censored Models

In this section we consider dynamic Tobit models in which the observed $y_{it}$ takes the form[9],

$$y_{it} = \begin{cases} y_{it}^*, & \text{if} \quad y_{it}^* > 0, \\ 0, & \text{if} \quad y_{it}^* \leq 0. \end{cases} \tag{8.6.1}$$

There could be two types of dynamic dependence for $y_{it}^*$ :

$$y_{it}^* = \gamma y_{i,t-1}^* + \boldsymbol{\beta}' \mathbf{x}_{it} + \alpha_i + u_{it}, \tag{8.6.2}$$

or

$$y_{it}^* = \gamma y_{i,t-1} + \boldsymbol{\beta}' \mathbf{x}_{it} + \alpha_i + u_{it}, \tag{8.6.3}$$

where the error term $u_{it}$ is independently distributed over $i$ and independently, identically distributed over $t$ (i.e., we allow Var $(u_{it}) = \sigma_i^2$).

For model (8.6.2), when $y_{i,t-1} = 0$, $y_{i,t-1}^*$ could be any value between $-\infty$ and 0. If there are no individual-specific effects $\alpha_i$ (or $\alpha_i = 0$ for all $i$), panel data actually allow the possibility of ignoring the censoring effects in the lagged dependent variables by concentrating on the subsample where $y_{i,t-1} > 0$. Because if $y_{i,t-1} > 0$, $y_{i,t-1} = y_{i,t-1}^*$, (8.6.1) and (8.6.2) with $\alpha_i = 0$ become

$$\begin{aligned} y_{it}^* &= \gamma y_{i,t-1}^* + \boldsymbol{\beta}' \mathbf{x}_{it} + u_{it} \\ &= \gamma y_{i,t-1} + \boldsymbol{\beta}' \mathbf{x}_{it} + u_{it}. \end{aligned} \tag{8.6.4}$$

Thus, by treating $y_{i,t-1}$ and $\mathbf{x}_{it}$ as predetermined variables that are independent of the error, $u_{it}$, the censored estimation techniques for the static model discussed in Section 8.1 can be applied to the subsample where (8.6.4) holds.

When random individual-specific effects $\alpha_i$ are present in (8.6.2), $y_{is}^*$ and $\alpha_i$ are correlated for all $s$ even if $\alpha_i$ can be assumed to be uncorrelated with $\mathbf{x}_i$. To implement the MLE approach, not only has one to make assumptions on the distribution of individual effects and initial observations but also computation may become unwieldy. To reduce the computational complexity,

[9] See Honoré (1993) for a discussion of the model $y_{it}^* = \gamma y_{i,t-1} + \boldsymbol{\beta}' \mathbf{x}_{it} + \alpha_i + u_{it}$.

Arellano, Bover, and Labeaga (1999) suggest a two-step approach. The first step estimates the reduced form of $y_{it}^*$ by projecting $y_{it}^*$ on all previous $y_{i0}^*, y_{i1}^*, \ldots, y_{i,t-1}^*$ and $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{it}$. The second step estimates $(\gamma, \boldsymbol{\beta}')$ from the reduced form parameters of $y_{it}^*$ equation, $\boldsymbol{\pi}_t$, by a minimum distance estimator of the form (3.8.14). To avoid the censoring problem in the first step, they suggest that for the $i$th individual, only the string $(y_{is}, y_{i,s-1}, \ldots, y_{i0})$, where $y_{i0} > 0, \ldots, y_{i,s-1} > 0$, is used. However, to derive the estimates of $\boldsymbol{\pi}_t$, the conditional distribution of $y_{it}^*$ given $y_{i0}^*, \ldots, y_{i,t-1}^*$ will have to be assumed. Moreover, the reduced form parameters $\boldsymbol{\pi}_t$ are related to $(\gamma, \boldsymbol{\beta}')$ in a highly nonlinear way. Thus, the second-stage estimator is not easily derivable. Therefore, in this section we bypass the issue of fixed or random $\alpha_i$ and discuss only the Honoré (1993) and Hu (1999) trimmed estimator.

For model (8.6.2) if $y_{i,t-1} = 0$ (i.e., $y_{i,t-1}^* < 0$), it is identical to the static model discussed in Section 5. If $y_{it} = 0$, there is no one-to-one correspondence between $u_{it}$ and $y_{it}^*$ given $(y_{i,t-1}, \mathbf{x}_{it}, \alpha_i)$. On the other hand, for model (8.6.3) there is still a one-to-one correspondence between $u_{it}$ and $y_{it}^*$ given $(y_{i,t-1}, \mathbf{x}_{it}, \alpha_i)$ be $y_{i,t-1} = 0$ or $> 0$. Therefore, we may split the observed sample for model (8.6.2) into two groups. For the group where $y_{i,t-1} = 0$, the estimation method discussed in Section 5 can be used to estimate $\boldsymbol{\beta}$. For the group where $y_{i,t-1} \neq 0$, it can be treated just as (8.6.3). However, to estimate $\gamma$ we need to consider the case $y_{i,t-1} > 0$. If we consider the trimmed sample for which $y_{i,t-1} = y_{i,t-1}^* > 0$, then for all practical purposes, the two models are identical.

For ease of demonstrating the symmetry conditions we consider the case when $T = 2$, $y_{i0}^*$ observable, and $y_{i1}^* > 0$, $y_{i0}^* > 0$. In Figures 8.8 and 8.9, let the vertical axis measure the value of $y_{i2}^* - \gamma y_{i1}^* = \tilde{y}_{i2}^*(\gamma)$ and horizontal axis measures $y_{i1}^*$. If $u_{i1}$ and $u_{i2}$ are i.i.d. conditional on $(y_{i0}, \mathbf{x}_{i1}, \mathbf{x}_{i2}, \alpha_i)$, then $y_{i1}^*$ and $y_{i2}^* - \gamma y_{i1}^* = \tilde{y}_{i2}^*(\gamma)$ are symmetrically distributed around the line (1), $\tilde{y}_{i2}^*(\gamma) = y_{i1}^* - \gamma y_{i0}^* + \boldsymbol{\beta}'\Delta\mathbf{x}_{i2}$ (or the 45-degree line through $(\gamma y_{i0} + \boldsymbol{\beta}'\mathbf{x}_{i1} + \alpha_i, \boldsymbol{\beta}'\mathbf{x}_{i2} + \alpha_i)$ or $(\gamma y_{i0} - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2}, 0)$). However, censoring destroys this symmetry. We observe only

$$\begin{aligned} y_{i1} &= \max(0, y_{i1}^*) \\ &= \max(0, \gamma y_{i0} + \boldsymbol{\beta}'\mathbf{x}_{i1} + \alpha_i + u_{i1}), \end{aligned}$$

and

$$y_{i2} = \max(0, \gamma y_{i1}^* + \boldsymbol{\beta}'\mathbf{x}_{i2} + \alpha_i + u_{i2}),$$

or

$$\tilde{y}_{i2}(\gamma) = \max(-\gamma y_{i1}, y_{i2}^* - \gamma y_{i1}).$$

That is, observations for $y_{i1}$ are censored from the left at the vertical axis, and for any $y_{i1} = y_{i1}^* > 0$, $y_{i2} = y_{i2}^* > 0$ implies that $y_{i2}^* - \gamma y_{i1}^* \geq -\gamma y_{i0} + \boldsymbol{\beta}'\Delta\mathbf{x}_{i2}$, and $y_{i2} - \gamma y_{i1}^* > -\gamma y_{i1}^*$. In other words, observations are also censored from below by $\tilde{y}_{i2}(\gamma) = -\gamma y_{i1}$, as line (2) in Figures 8.8 and 8.9. As shown in

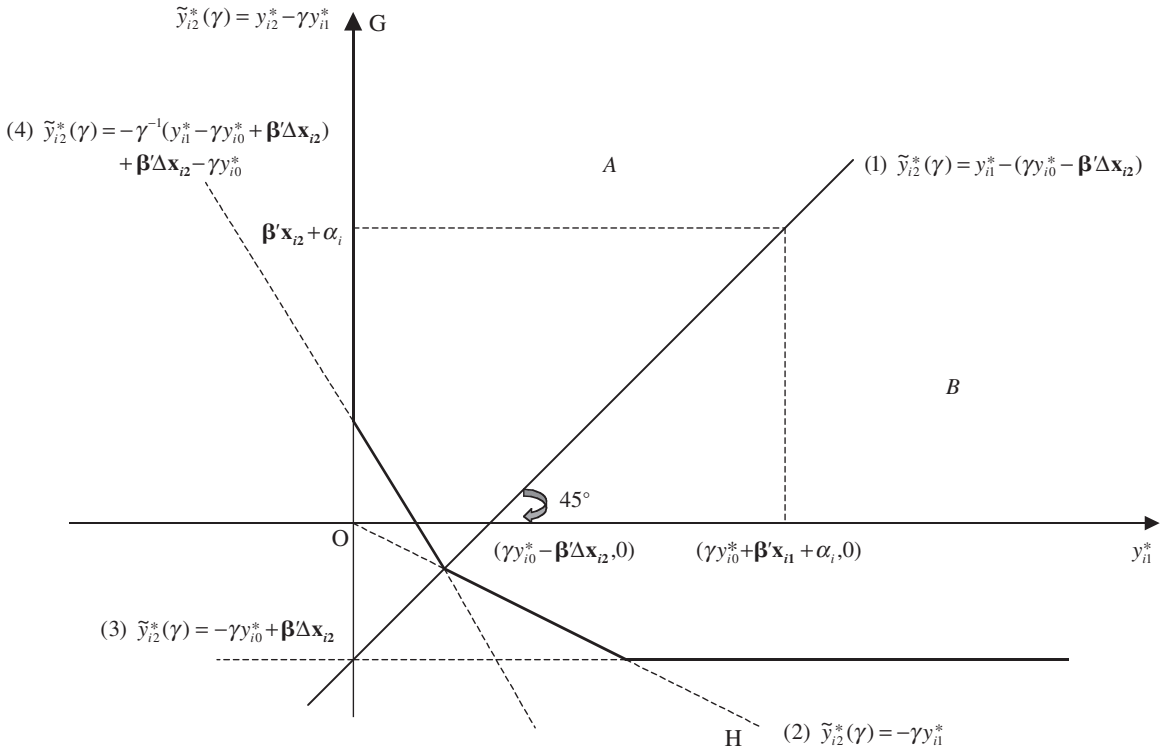


Figure 8.8. $\gamma > 0, \gamma y^*_{i0} - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2} > 0.$

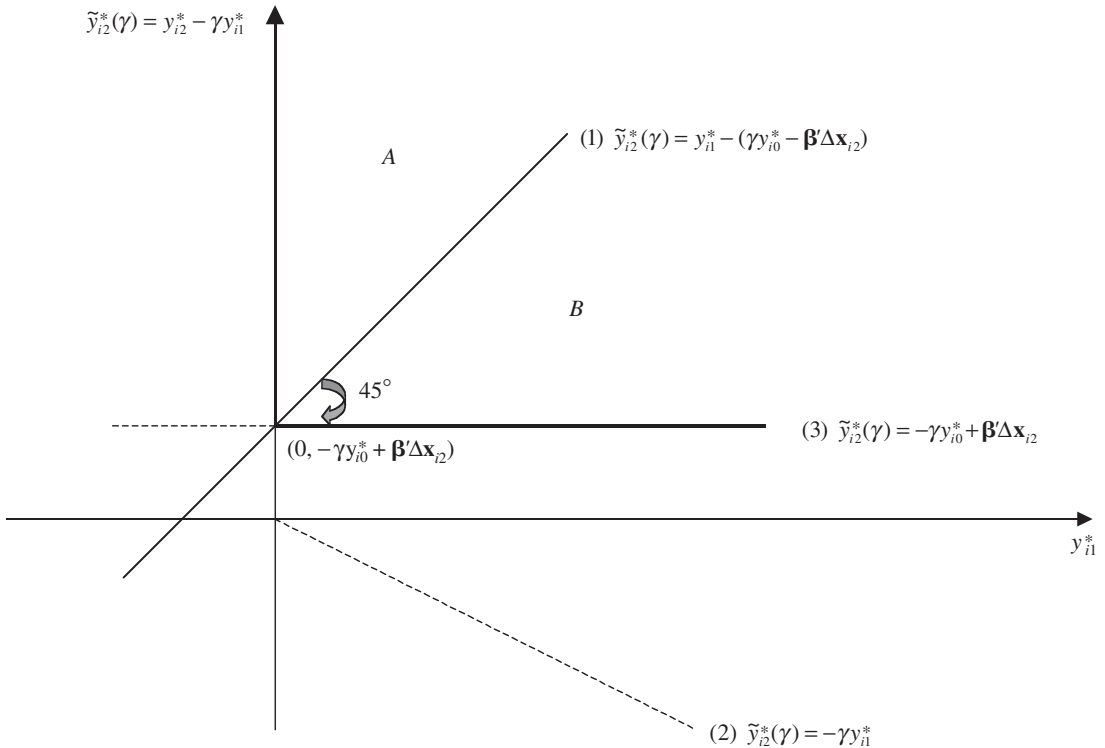


Figure 8.9. $\gamma > 0, \gamma y^*_{i0} - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2} < 0.$

Figure 8.8, the observable ranges of $y_{i1}^*$ and $y_{i2}^* - \gamma y_{i1}^*$ conditional on $(\mathbf{x}_{i1}, \mathbf{x}_{i2}, y_{i0}^*)$ are in the region GOH. The region is not symmetric around the line (1), where we have drawn with $\gamma \geq 0$, $\gamma y_{i0} - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2} > 0$. To restore symmetry, we have to find the mirror images of these two borderlines – the vertical axis and line (2) – around the centerline (1), and then symmetrically truncate observations that fall outside these two new lines.

The mirror image of the vertical axis around line (1) is the horizontal line $\tilde{y}_{i2}^*(\gamma) = -\gamma y_{i0} + \boldsymbol{\beta}'\Delta\mathbf{x}_{i2}$, line (3) in Figure 8.8. The mirror image of line (2) around line (1) has the slope the inverse of line (2), $-\frac{1}{\gamma}$. Therefore, the mirror image of line (2) is the line $\tilde{y}_{i2}^*(\gamma) = -\frac{1}{\gamma}y_{i1}^* + c$, that passes through the intersection of line (1) and line (2). The intersection of line (1) and line (2) is given by $\bar{\tilde{y}}_{i2}^*(\gamma) = \bar{y}_{i1}^* - (\gamma y_{i0} - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2}) = -\gamma\bar{y}_{i1}^*$. Solving for $(\bar{y}_{i1}^*, \bar{\tilde{y}}_{i2}^*(\gamma))$, we have $\bar{y}_{i1}^* = \frac{1}{1+\gamma}(\gamma y_{i0} - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2})$, $\bar{\tilde{y}}_{i2}^*(\gamma) = -\frac{\gamma}{1+\gamma}(\gamma y_{i0} - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2})$. Substituting $\tilde{y}_{i2}^*(\gamma) = \bar{\tilde{y}}_{i2}^*(\gamma)$ and $y_{i1}^* = \bar{y}_{i1}^*$ into the equation $\tilde{y}_{i2}^*(\gamma) = -\frac{1}{\gamma}y_{i1}^* + c$, we have $c = \frac{1-\gamma}{\gamma}(\gamma y_{i0} - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2})$. Thus the mirror image of line (2) is $\tilde{y}_{i2}(\gamma) = -\frac{1}{\gamma}(y_{i1}^* - \gamma y_{i0} + \boldsymbol{\beta}'\Delta\mathbf{x}_{i2}) - (\gamma y_{i0} - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2})$, line (4) in Figure 8.8.

In Figure 8.9 we show the construction of the symmetrical truncation region for the case when $\gamma y_{i0} - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2} < 0$. Because observations are truncated at the vertical axis from the left and at line (2) from below, the mirror image of vertical axis around line (1) is given by line (3). Therefore, if we truncate observations at line (3) from below, then the remaining observations will be symmetrically distributed around line (1).

The observations of $(y_{i1}, \tilde{y}_{i2}(\gamma))$ falling into the northeast direction of the region bordered by the lines (2), (3), and (4) in Figure 8.8 or the vertical axis and line (3) in Figure 8.9 are symmetrically distributed around line (1) (the 45-degree line through $(\gamma y_{i0}^* - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2}, 0)$). Denote the region above the 45-degree line by A and the region below the 45-degree line by B. Then

$$
\begin{aligned}
A \cup B \equiv \Big\{ & (y_{i1}, \tilde{y}_{i2}(\gamma)) : y_{i1} > 0, \tilde{y}_{i2}(\gamma) > -\gamma y_{i1}, y_{i1} > \gamma y_{i0} - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2} - \gamma \\
& \times (\tilde{y}_{i2}(\gamma) + \gamma y_{i0} - \boldsymbol{\beta}\Delta\mathbf{x}_{i2}), \tilde{y}_{i2}(\gamma) > -\gamma y_{i0} + \boldsymbol{\beta}'\Delta\mathbf{x}_{i2} \Big\} \\
= \Big\{ & (y_{i1}, \tilde{y}_{i2}(\gamma)) : y_{i1} > 0, y_{i2} > 0, y_{i1} > \gamma y_{i0} - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2} \\
& - \gamma(\tilde{y}_{i2}(\gamma) + \gamma y_{i0} - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2}), \tilde{y}_{i2}(\gamma) > -\gamma y_{i0} + \boldsymbol{\beta}'\Delta\mathbf{x}_{i2} \Big\}.
\end{aligned}
\tag{8.6.5}
$$

Symmetry implies that conditional on $y_{i0} > 0$, $y_{i1} > 0$, $y_{i2} > 0$ and $\mathbf{x}_{i1}, \mathbf{x}_{i2}$, the probability of an observation falling in region A equals the probability of it

falling in region B. That is

$$E\Big\{(y_{i1}, \tilde{y}_{i2}(\gamma))\epsilon A \cup B\Big\} \cdot \Big[1\Big\{y_{i1} - \tilde{y}_{i2}(\gamma) - \gamma y_{i0} + \boldsymbol{\beta}'\Delta\mathbf{x}_{i2} > 0\Big\} - 1\Big\{y_{i1} - \tilde{y}_{i2}(\gamma) - \gamma y_{i0} + \boldsymbol{\beta}'\Delta\mathbf{x}_{i2} < 0\Big\}\Big] = 0. \tag{8.6.6}$$

Another implication of symmetry is that conditional on $y_{i0} > 0$, $y_{i1} > 0$, $y_{i2} > 0$ and $\mathbf{x}_{i1}$, $\mathbf{x}_{i2}$, the expected vertical distance from a point in region A to the line (1), $\tilde{y}_{i2}(\gamma) - y_{i1} + \gamma y_{i0} - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2}$, equals the expected horizontal distance from a point in region B to that line, $y_{i1} - \tilde{y}_{i2}(\gamma) - \gamma y_{i0} + \boldsymbol{\beta}'\Delta\mathbf{x}_{i2} = -(\tilde{y}_{i2}(\gamma) - y_{i1} + \gamma y_{i0} - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2})$. Therefore,

$$E\Big[1\Big\{(y_{i1}, \tilde{y}_{i2}(\gamma))\epsilon A \cup B\Big\}(y_{i1} - \tilde{y}_{i2}(\gamma) - \gamma y_{i0} + \boldsymbol{\beta}'\Delta\mathbf{x}_{i2})\Big] = 0. \tag{8.6.7}$$

More generally, for any function $\xi(.,.)$ satisfying $\xi(e_1, e_2) = -\xi(e_2, e_1)$ for all $(e_1, e_2)$, we have the orthogonality condition

$$E\Big[1\Big\{(y_{i1}, \tilde{y}_{i2}(\gamma))\epsilon A \cup B\Big\} \cdot \xi(y_{i1} - \gamma y_{i0} + \boldsymbol{\beta}'\Delta\mathbf{x}_{i2}, \tilde{y}_{i2}(\gamma)) \cdot h(y_{i0}, \mathbf{x}_{i1}, \mathbf{x}_{i2})\Big] = 0, \tag{8.6.8}$$

for any function $h(\cdot)$. where

$$\begin{aligned} 1\Big\{(y_{i1}, \tilde{y}_{i2}(\gamma))\epsilon A \cup B\Big\} &\equiv 1\Big\{y_{i0} > 0, y_{i1} > 0, y_{i2} > 0\Big\} \\ &\cdot \Big[1\Big\{\gamma y_{i0} - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2} > 0\Big\} \\ &\cdot 1\Big\{y_{i1} > \gamma y_{i0} - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2} - \gamma(\tilde{y}_{i2}(\gamma) + \gamma y_{i0} - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2})\Big\} \\ &\cdot 1\Big\{\tilde{y}_{i2}(\gamma) > -\gamma y_{i0} + \boldsymbol{\beta}'\Delta\mathbf{x}_{i2}\Big\} + 1\Big\{\gamma y_{i0} - \boldsymbol{\beta}'\Delta\mathbf{x}_{i2} < 0\Big\} \\ &\cdot 1\Big\{\tilde{y}_{i2}(\gamma) > -\gamma y_{i0} + \boldsymbol{\beta}'\Delta\mathbf{x}_{i2}\Big\}\Big]. \end{aligned} \tag{8.6.9}$$

If one chooses $h(\cdot)$ to be a constant, the case $\xi(e_1, e_2)$=sgn $(e_1 - e_2)$ corresponds to (8.6.6) and $\xi(e_1, e_2) = e_1 - e_2$ corresponds to (8.6.7).

If $T \geq 4$, one can also consider any pair of observations $y_{it}$, $y_{is}$ with $y_{i,t-1} > 0$, $y_{it} > 0$, $y_{i,s-1} > 0$ and $y_{is} > 0$. Note that conditional on $\mathbf{x}_{it}$, $\mathbf{x}_{is}$, $(\alpha_i + u_{it})$ and $(\alpha_i + u_{is})$ are identically distributed. Thus, let

$$\begin{aligned} W_{its}(\boldsymbol{\beta}', \gamma) &= \max\ \Big\{0, (\mathbf{x}_{it} - \mathbf{x}_{is})'\boldsymbol{\beta}, y_{it} - \gamma y_{i,t-1}\Big\} - \mathbf{x}'_{it}\boldsymbol{\beta} \\ &= \max\ \Big\{-\mathbf{x}'_{it}\boldsymbol{\beta}, -\mathbf{x}'_{is}\boldsymbol{\beta}, \alpha_i + u_{it}\Big\}, \end{aligned} \tag{8.6.10}$$

and

$$\begin{aligned} W_{ist}(\boldsymbol{\beta}', \gamma) &= \max\ \left\{0, (\mathbf{x}_{is} - \mathbf{x}_{it})'\boldsymbol{\beta}, y_{is} - \gamma y_{i,s-1}\right\} - \mathbf{x}_{is}'\boldsymbol{\beta} \\ &= \max\ \left\{-\mathbf{x}_{is}'\boldsymbol{\beta}, -\mathbf{x}_{it}'\boldsymbol{\beta}, \alpha_i + u_{is}\right\}, \end{aligned} \tag{8.6.11}$$

Then $W_{its}(\boldsymbol{\beta}, \gamma)$ and $W_{ist}(\boldsymbol{\beta}, \gamma)$ are distributed symmetrically around the 45-degree line conditional on $(\mathbf{x}_{it}, \mathbf{x}_{is})$. This suggests the orthogonality condition

$$\begin{aligned} E\left[1\left\{y_{it-1} > 0, y_{it} > 0, y_{i,s-1} > 0, y_{is} > 0\right\} \cdot \xi(W_{its}(\boldsymbol{\beta}', \gamma), W_{ist}(\boldsymbol{\beta}', \gamma))\right. \\ \left.\cdot h(\mathbf{x}_{it}, \mathbf{x}_{is})\right] = 0, \end{aligned} \tag{8.6.12}$$

for any function $h(\cdot)$. When $T \geq 3$, the symmetric trimming procedure (8.6.12) requires weaker assumptions than the one based on three consecutive uncensored observations because the conditioning variables do not involve the initial value $y_{i0}$. However, this approach also leads to more severe trimming.

Based on the orthogonality conditions (8.6.8) or (8.6.12), Hu (1999) suggests a GMM estimator of $\boldsymbol{\theta} = (\boldsymbol{\beta}', \gamma)'$ by minimizing $\mathbf{m}_N(\boldsymbol{\theta})' A_N \mathbf{m}_N(\boldsymbol{\theta})$ where $\mathbf{m}_N(\boldsymbol{\theta})$, is the sample analog of (8.6.8) or (8.6.12), and $A_N$ is a positive definite matrix that converges to a constant matrix A as $N \to \infty$. The GMM estimator will have the limiting distribution of the form

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{\text{GMM}} - \boldsymbol{\theta}\right) \longrightarrow N\left(\mathbf{0}, (\Gamma'\Lambda\Gamma)^{-1}\left[\Gamma' A V A \Gamma\right](\Gamma' A \Gamma)^{-1}\right), \tag{8.6.13}$$

where $\Gamma = \frac{\partial}{\partial \boldsymbol{\theta}} E[\mathbf{m}(\boldsymbol{\theta})]$, $V = E[\mathbf{m}(\boldsymbol{\theta})\mathbf{m}(\boldsymbol{\theta})']$. When the optimal weighting matrix $A = V^{-1}$ is used, the asymptotic covariance matrix of $\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{\text{GMM}} - \boldsymbol{\theta}\right)$ becomes $\left(\Gamma' V^{-1} \Gamma\right)^{-1}$.

However, the true value of $\boldsymbol{\theta}$ is not the only value that satisfies the orthogonality conditions (8.6.6)–(8.6.8) or (8.6.12). For instance, those orthogonality conditions can be trivially satisfied when the parameter values are arbitrarily large. To see this, note that for a given value of $\gamma$, when the value of $\delta_{it} = \mathbf{x}_{it}'\boldsymbol{\beta}$ goes to infinity, the number of observations falling in the (nontruncated) region A∪B in Figures 8.8 and 8.9 approaches 0. Thus, the moment conditions can be trivially satisfied. To overcome this possible lack of identification of GMM estimates based on the minimization of the criterion function, Hu (1999) suggests using a subset of the moments that exactly identify $\boldsymbol{\beta}$ for given $\gamma$ to provide the estimates of $\boldsymbol{\beta}$, then test whether the rest of the moment conditions are satisfied by these estimates for a sequence of $\gamma$ values ranging from 0 to 0.9 with an increment of 0.01. Among the values of $\gamma$ at which the test statistics are not rejected, the one that yields the smallest test statistic is chosen as the estimate of $\gamma$. Hu (1999) uses this estimation method to study earnings dynamics, using matched data from the Current Population Survey and Social Security Administration (CPS-SSA) Earnings Record for a sample of men who were born in 1930–39 and living in the South during the period of 1957–73. The SSA earnings are top-coded at the maximum social security taxable level, namely, $y_{it} = \min(y_{it}^*, c_t)$, where $c_t$ is the Social Security maximum taxable

Table 8.4. *Estimates of AR(1) coefficients of log real annual earnings (in thousands)*[a]

| Linear GMM (assuming no censoring) | | Nonlinear GMM with correction for censoring | |
|---|---|---|---|
| Black | White | Black | White |
| 0.379 | 0.399 | 0.210 | 0.380 |
| (0.030) | (0.018) | (0.129) | (0.051) |

[a] Standard errors in parenthesis.
*Source:* Hu (1999).

earnings level in period $t$. This censoring at the top can be easily translated into censoring at 0 by considering $\tilde{y}_{it} = c_t - y_{it}$, then $\tilde{y}_{it} = \max(0, c_t - y_{it}^*)$.

Table 8.4 presents the estimates of the coefficient of the lagged log real annual earnings coefficient of an AR(1) model based on a sample of 226 black and 1883 white men with and without correction for censoring. When censoring is ignored, the model is estimated by the linear GMM method. When censoring is taken into account, Hu uses an unbalanced panel of observations with positive SSA earnings in three consecutive time periods. The estimated $\gamma$ are very similar for black and white men when censoring is ignored. However, when censoring is taken into account, the estimated autoregressive parameter $\gamma$ is much higher for white men than for black men. The higher persistence of the earnings process for white men than for black men is consistent with the notion that white men had jobs that had better security and were less vulnerable to economic fluctuation than black men in the period 1957–73.

### 8.6.2 Dynamic Sample Selection Models

When the selection rule is endogenously determined as given by (8.2.4) and $y_{it}^*$ is given by (8.6.2) or (8.6.3), with $\mathbf{w}_{it}$ and $\mathbf{x}_{it}$ being nonoverlapping vectors of strictly exogenous explanatory variables (with possibly common elements), the model under consideration has the form:[10]

$$y_{it} = d_{it} y_{it}^*, \tag{8.6.14}$$

$$d_{it} = 1\{\mathbf{w}_{it}'\mathbf{a} + \eta_i + \nu_{it}\}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{matrix} \tag{8.6.15}$$

where $(d_{it}, \mathbf{w}_{it})$ is always observed, and $(y_{it}^*, \mathbf{x}_{it})$ is observed only if $d_{it} = 1$. For notational ease, we assume that $d_{i0}$ and $y_{i0}$ are also observed.

In the static case of $\gamma = 0$, Kyriazidou (1997) achieves the identification of $\boldsymbol{\beta}$ by relying on the conditional pairwise exchangeability of the error vector

[10] The assumption that $\mathbf{x}_{it}$ and $\mathbf{w}_{it}$ do not coincide rules out the censored regression model as a special case of (8.6.14) and (8.6.15).

$(u_{it}, v_{it})$ given the entire path of the exogenous variables $(\mathbf{x}_i, \mathbf{w}_i)$ and the individual effects $(\alpha_i, \eta_i)$. However, the consistency of Kyriazidou estimator (8.4.33) breaks down in the presence of the lagged dependent variable in (8.6.2) or (8.6.3). The reason is the same as in linear dynamic panel data models where first differencing generates nonzero correlation between $y^*_{i,t-1}$ and the transformed error term (see Chapter 4). However, just as in the linear case, estimators based on linear and nonlinear moment conditions on the correlation structure of the unobservables with the observed variables can be used to obtain consistent estimators of $\gamma$ and $\boldsymbol{\beta}$.

Under the assumption that $\{u_{it}, v_{it}\}$ are independently, identically distributed over time for all $i$ conditional on $\boldsymbol{\xi}_i \equiv (\mathbf{w}'_i, \alpha_i, \eta_i, y^*_{i0}, d_{i0})$, where $\mathbf{w}_i = (\mathbf{w}'_{i1}, \ldots, \mathbf{w}_{iT})'$, Kyriazidou (2001) notes that by conditioning on the event that $\Delta\mathbf{w}'_{it}\mathbf{a} = 0$, the following moment conditions hold[11]:

$$E(d_{it}d_{i,t-1}d_{i,t-2}d_{i,t-j}y_{i,t-j}\Delta u_{it} \mid \Delta\mathbf{w}'_{it}\mathbf{a} = 0) = \mathbf{0},\ j = 2, \ldots, t, \tag{8.6.16}$$

and

$$E(d_{is}d_{it}d_{i,t-1}d_{i,t-2}\mathbf{x}_{is}\Delta u_{it} \mid \Delta\mathbf{w}'_{it}\mathbf{a} = 0) = 0, \quad \text{for } t = 2. \ldots, T; s = 1, \ldots, T. \tag{8.6.17}$$

This is because for an individual $i$ when the selection index $\mathbf{w}'_{it}\mathbf{a} = \mathbf{w}'_{i,t-1}\mathbf{a}$, the magnitude of the sample selection effects in the two periods, $\lambda(\eta_i + \mathbf{w}'_{it}\mathbf{a})$ and $\lambda(\eta_i + \mathbf{w}'_{i,t-1}\mathbf{a})$, will also be the same. Thus by conditioning on $\Delta\mathbf{w}'_{it}\mathbf{a} = 0$, the sample selection effects and the individual effects are eliminated by first differencing,

Let $\boldsymbol{\theta} = (\gamma, \boldsymbol{\beta}')'$, $\mathbf{z}'_{it} = (y_{i,t-1}, \mathbf{x}'_{it})$, and

$$m_{1it}(\boldsymbol{\theta}) = d_{it}d_{i,t-1}d_{i,t-2}d_{i,t-j}y_{i,t-j}(\Delta y_{it} - \Delta\mathbf{z}'_{it}\boldsymbol{\theta}), \quad t = 2, \ldots, T; j = 2, \ldots, t, \tag{8.6.18}$$

$$m_{2it,k}(\boldsymbol{\theta}) = d_{is}d_{it}d_{i,t-1}d_{i,t-2}x_{is,k}(\Delta y_{it} - \Delta\mathbf{z}'_{it}\boldsymbol{\theta}), \quad t = 2, \ldots, T; s = 1, \ldots, T; k = 1, \ldots, K. \tag{8.6.19}$$

Kyriazidou (2001) suggests a kernel weighted generalized method of moments estimator (KGMM) that minimizes the following quadratic form:

$$\hat{G}_N(\boldsymbol{\theta})'A_N\hat{G}_N(\boldsymbol{\theta}), \tag{8.6.20}$$

where $A_N$ is a stochastic matrix that converges in probability to a finite nonstochastic limit $A$, $\hat{G}_N(\boldsymbol{\theta})$ is the vector of stacked sample moments with rows

[11] Kyriazidou (2001) shows that these moment conditions also hold if $d^*_{it} = \phi d_{i,t-1} + \mathbf{w}'_{it}\mathbf{a} + \eta_i + v_{it}$.

of the form

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{h_N}K\left(\frac{\Delta\mathbf{w}_{it}'\hat{\mathbf{a}}}{h_N}\right)m_{\ell it}(\boldsymbol{\theta}), \tag{8.6.21}$$

where $K(\cdot)$ is a kernel density function, $\hat{\mathbf{a}}$ is some consistent estimator of $\mathbf{a}$, and $h_N$ is a bandwidth that shrinks to 0 as $N \to \infty$. Under appropriate conditions, Kyriazidou (2001) proves that the KGMM estimator is consistent and asymptotically normal. The rate of convergence is the same as in univariate nonparametric density and regression function estimation, that is, at the speed of $\sqrt{Nh_N}$.

CHAPTER 9

# Cross-Sectionally Dependent Panel Data

Most panel inference procedures discussed so far assume that apart from the possible presence of individual invariant but period varying time-specific effects, the effects of omitted variables are independently distributed across cross-sectional units. Often economic theory predicts that agents take actions that lead to interdependence among themselves. For example, the prediction that risk-averse agents will make insurance contracts allowing them to smooth idiosyncratic shocks implies dependence in consumption across individuals. Cross-sectional units could also be affected by common omitted factors. The presence of cross-sectional dependence can substantially complicate statistical inference for a panel data model. However, properly exploiting the dependence across cross-sectional units in panel data not only can improve the efficiency of parameter estimates, but it may also simplify statistical inference than the situation where only cross-sectional data are available. In Section 9.1 we discuss issues of ignoring cross-sectional dependence. Sections 9.2 and 9.3 discuss spatial and factor approaches for modeling cross-sectional dependence. Section 9.4 discusses cross-sectional mean augmented approach for controlling the impact of cross-sectional dependency. Section 9.5 discusses procedures for testing cross-sectional dependence. Section 9.6 demonstrates that when panel data are cross-sectionally dependent, sometimes it may considerably simplify statistical analysis compared to the case of when only cross-sectional data are available by considering the measurement of treat effects.

## 9.1 ISSUES OF CROSS-SECTIONAL DEPENDENCE

Standard two-way effects models (e.g. (3.6.8)) imply observations across individuals are equal correlated. However, the impacts of common omitted factors could be different for different individuals. Unfortunately, contrary to the observations along the time dimension in which the time label, $t$ or $s$, gives a natural ordering and structure, there is no natural ordering of observations along the cross-sectional dimension. The cross-sectional labeling, $i$ or $j$, is arbitrary.

Let $\mathbf{v}_t = (v_{1t}, \ldots, v_{Nt})'$ be the $N \times 1$ vector of cross-sectionally stacked error vector at time $t$ and the $N \times N$ constant matrix, $\sum = (\sigma_{ij})$ be its

covariance matrix. When $N$ is fixed and $T$ is large, one can estimate the covariance between $i$ and $j$, $\sigma_{ij}$, by $\frac{1}{T}\sum_{t=1}^{T}\hat{v}_{it}\hat{v}_{jt}$ directly, using individual time series regression residuals, $\hat{v}_{it}$ if the conditional variables, $\mathbf{x}_{it}$, are uncorrelated with $v_{it}$. One can then apply the feasible generalized least-squares method (FGLS) or Zellner's (1962) seemingly unrelated regression method (SUR) to estimate the slope coefficients. The FGLS or SUR estimator is consistent and asymptotically normally distributed.

When $T$ is finite, unrestricted $\sum$ cannot be consistently estimated. However, if each row of $\sum$ only has a maximum of $h_N$ elements that are nonzero (i.e., cross-sectional dependence is in a sense "local")[1] and $\frac{h_N}{N} \to 0$ as $N \to \infty$, panel estimators that ignore cross-sectional dependence could still be consistent and asymptotically normally distributed, although they will not be efficient. The test statistics based on the formula ignoring cross-correlations could also lead to severe size distortion (e.g., Breitung and Das 2008). On the other hand, if $\frac{h_N}{N} \to c \neq 0$ as $N \to \infty$, estimators that ignore the presence of cross-sectional dependence could be inconsistent no matter how large $N$ is (e.g., Hsiao and Tahmiscioglu (2008), Phillips and Sul (2007)) if $T$ is finite.[2] To see this, consider the simple regression model,

$$\begin{aligned} y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + v_{it}, \quad & i = 1, \ldots, N \\ & t = 1, \ldots, T, \end{aligned} \tag{9.1.1}$$

where $E(v_{it} \mid \mathbf{x}_{it}) = 0$, $E(v_{it}v_{js}) = 0$ if $t \neq s$ and $E(v_{it}v_{jt}) = \sigma_{ij}$. Let $\sum = (\sigma_{ij})$ be the $N \times N$ covariance matrix of the cross-sectionally stacked error, $\mathbf{v}_t = (v_{1t}, \ldots, v_{Nt})'$. Then the covariance matrix of pooled least-squares estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$, is equal to

$$\mathrm{Cov}\,(\hat{\boldsymbol{\beta}}_{\mathrm{LS}}) = \left[\sum_{t=1}^{T} X_t'X_t\right]^{-1}\left[\sum_{t=1}^{T} X_t'\Sigma X_t\right]\left[\sum_{t=1}^{T} X_t'X_t\right]^{-1}, \tag{9.1.2}$$

where $X_t = (\mathbf{x}_{it}')$ denotes the $N \times K$ cross-sectionally stacked explanatory variables $\mathbf{x}_{it}$ for time period $t$. Since $\sum$ is a symmetrical positive definite matrix, $\sum$ can be decomposed as

$$\Sigma = \vee\Lambda\vee', \tag{9.1.3}$$

where $\wedge$ is a diagonal matrix with the diagonal elements being the eigenvalues of $\Sigma$ and $\vee$ is an orthonormal matrix. If one or more eigenvalues of $\Sigma$ is of order $N$, $\mathbf{X}_t'\Sigma\mathbf{X}_t$ could be of order $N^2$ under the conventional assumption that $\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{it}$ converges to a constant vector. Hence the

$$\mathrm{Cov}\,(\hat{\boldsymbol{\beta}}_{\mathrm{LS}}) = O(\frac{1}{T}).$$

[1] This is equivalent to saying the eigenvalues of $\sum$ are bounded as $N \longrightarrow \infty$. Pesaran and Tosetti (2010) define this case as the "weak dependence."

[2] This is equivalent to saying the eigenvalues of $\sum$ are $O(N)$, which Pesaran and Tosetti (2010) called the "strong dependence."

In other words, the least-squares estimator of $\boldsymbol{\beta}$ converges to a random variable rather than a constant when $T$ is finite and $N$ is large.

## 9.2 SPATIAL APPROACH

### 9.2.1 Introduction

In regional science, correlation across cross-sectional units is assumed to follow a certain spatial ordering, that is, dependence among cross-sectional units is related to location and distance, in a geographic or more general economic or social network space (e.g., Anselin 1988; Anselin and Griffith 1988; Anselin, Le Gallo, Jayet 2008). The neighbor relation is expressed by a so-called spatial weights matrix, $W = (w_{ij})$, an $N \times N$ positive matrix in which the rows and columns correspond to the cross-sectional units, is specified to express the prior strength of the interaction between location $i$ (in the row of the matrix) and location $j$ (column), $w_{ij}$. By convention, the diagonal elements, $w_{ii} = 0$. The weights are often standardized so that the sum of each row, $\sum_{j=1}^{N} w_{ij} = 1$ through row-normalization; for instance, let the $i$th row of $W$, $\mathbf{w}_i' = (d_{i1}, \ldots, d_{iN}) / \sum_{j=1}^{N} d_{ij}$, where $d_{ij} \geq 0$ represents a function of the spatial distance of the $i$th and $j$th units in some (characteristic) space. A side effect of this standardization is that whereas the original weights may be symmetrical, the row-standardized form no longer is.

The spatial weights matrix, $W$, is often included into a model specification to the dependent variable, or to the error term or to both through a so-called *spatial lag operator*. For instance, a *spatial lag* model for the $NT \times 1$ variable $\mathbf{y} = (\mathbf{y}_1', \ldots, \mathbf{y}_N')'$, $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$, may take the form

$$\mathbf{y} = \rho(W \otimes I_T)\mathbf{y} + X\boldsymbol{\beta} + \mathbf{u} \tag{9.2.1}$$

where $X$ and $\mathbf{u}$ denote the $NT \times K$ explanatory variables and $NT \times 1$ vector of error terms, respectively (called the mixed regressive, spatial autoregression model by Anselin (1988) and Ord (1975)). A *spatial error* model may take the form

$$\mathbf{y} = X\boldsymbol{\beta} + \mathbf{v}, \tag{9.2.2}$$

where $\mathbf{v}$ may be specified as in a *spatial autoregressive* form,

$$\mathbf{v} = \theta(W \otimes I_T)\mathbf{v} + \mathbf{u}, \tag{9.2.3}$$

or a *spatial moving average* form,

$$\mathbf{v} = \delta(W \otimes I_T)\mathbf{u} + \mathbf{u}, \tag{9.2.4}$$

and $\mathbf{u}_i' = (u_{i1}, \ldots, u_{iT})$ is assumed to be independently distributed across $i$ with $E\mathbf{u}_i\mathbf{u}_i' = \sigma_u^2 I_T$.

The joint determination of $\mathbf{y}$ for model (9.2.1) or $\mathbf{v}$ for (9.2.2) when $\mathbf{v}$ is given by (9.2.3) is through $[(I_N - \rho W)^{-1} \otimes I_T]$ or $[(I_N - \theta W)^{-1} \otimes I_T]$. Since

$$(I_N - \rho W)^{-1} = I_N + \rho W + \rho^2 W^2 + \cdots, \tag{9.2.5}$$

or

$$(I_N - \theta W)^{-1} = I_N + \theta W + \theta^2 W^2 + \cdots, \tag{9.2.6}$$

to ensure a "distance" decaying effect among the cross-sectional units, $\rho$ and $\theta$ are assumed to have absolute values less than 1.[3]

The *spatial autoregressive* form (9.2.3) implies that the covariance matrix of the $N$ cross-sectional units at time $t$, $\mathbf{v}_t = (v_{1t}, \ldots, v_{Nt})'$ takes the form

$$E\mathbf{v}_t\mathbf{v}_t' = \sigma_u^2[I_N - \theta W]^{-1}[I_N - \theta W']^{-1} = V. \tag{9.2.7}$$

The *spatial moving average* form (9.2.4) implies that the covariance matrix of $\mathbf{v}_t$ takes the form

$$\begin{aligned} E\mathbf{v}_t\mathbf{v}_t' &= \sigma_u^2[I_N + \delta W][I_N + \delta W]' \\ &= \sigma_u^2[I_N + \delta(W + W') + \delta^2 WW'] = \tilde{V}. \end{aligned} \tag{9.2.8}$$

When $W$ is *sparse*, that is, many elements of $W$ are prespecified to be 0, for instance, $W$ could be a block diagonal matrix in which only observations in the same region are considered *neighbors*, and observations across regions are uncorrelated. $W$ can also be a sparse matrix by some neighboring specification, for example, if a district is a spatial unit, some specifications assume that a neighbor for this district is another one which has a common boundary. The spatial moving average form allows the cross-correlations to be "local" (9.2.8). On the other hand, the *spatial autoregressive* form suggests a much wider range of spatial covariance than specified by the nonzero elements of the weights matrix $W$, implying a "global" covariance structure (9.2.7).

Generalizing the spatial approach, Conley (1999) suggests using the notion of "economic distance" to model proximity between two economic agents. The joint distribution of random variables at a set of points is assumed to be invariant to a shift in location and is a function only of the "economic distances" between them. For instance, the population of individuals is assumed to reside in a low dimensional Euclidean space, say $R^2$, with each individual $i$ located at a point $s_i$. The sample then consists of realization of agents' random variables at a collection of locations $\{s_i\}$ inside a sample region. If two agents' locations $s_i$ and $s_j$ are close, then $y_{it}$ and $y_{js}$ may be highly correlated. As the distance

[3] The combination of the row sum of $W$ equal to 1 and $\gamma$ or $\theta$ having absolute value less than 1 implies that the spatial models assume cross-sectional dependence being "weak."

between $s_i$ and $s_j$ grows large, $y_{it}$ and $y_{js}$ approach independence. Under this assumption, the dependence among cross-sectional data can be estimated using methods analogous to time series procedures either parametrically or nonparametrically (e.g., Hall, Fisher, and Hoffman 1992; Priestley 1982; Newey and West 1987).

While the approach of defining cross-sectional dependence in terms of "economic distance" measure allows for more complicated dependence than models with time-specific (or group-specific) effects alone (e.g. Chapter 3, Section 3.6), it still requires that the econometricians have information regarding this "economic distance." In certain urban, environmental, development, growth, and other areas of economics, this information may be available. For instance, in the investigation of peoples' willingness to pay for local public goods, the relevant economic distance may be the time and monetary cost of traveling between points to use these local public goods. Alternatively, if the amenity is air quality, then local weather conditions might constitute the major unobservable common to cross-sectional units in the neighborhood. Other examples include studies of risk sharing in rural developing economies where the primary shocks to individuals in such agrarian economies may be weather related. If so, measures of weather correlation on farms of two individuals could be the proxy for the economic distance between them. In many other situations, prior information such as this may be difficult to come by. However, the combination of the row sum of $W$ equal to 1 and $\delta$ or $\theta$ having absolute value less than 1 implies that the population consists of $N$ cross-sectional units. In other words, the spatial approach is an analysis of the population based on $T$ time series realized observations. If $N$ is considered sample size, then the spatial autoregressive model implies that the cross-sectional dependence is "weak." In other words, each cross-sectional unit is correlated only with a fixed number of other cross-sectional units. Under an assumpton of weak cross-sectional dependence, the covariance estimator (3.2.8) for models with only individual-specific effects or (3.6.13) for models with both individual- and time-specific effects of $\boldsymbol{\beta}$ remains consistent if $T$ is fixed and $N \to \infty$ or if $N$ is fixed and $T$ tends to infinity or both. However, there could be severe size distortion in hypothesis testing if cross-sectional dependence is ignored. Vogelsang (2012) showed that the covariance matrix estimate proposed by Driscoll and Kraay (1998) based on the Newey–West (1987) heteroscedastic autocorrelation (HAC) covariance matrix estimator of cross-sectional averages,

$$T\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}_{it}'\right)^{-1}\hat{\tilde{\Omega}}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}\tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}_{it}'\right)^{-1}, \tag{9.2.9}$$

is robust to heteroscedasticity, autocorrelation and spatial dependence, where $\tilde{\mathbf{x}}_{it} = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)$ if (3.2.8) is used or $\tilde{\mathbf{x}}_{it} = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i - \bar{\mathbf{x}}_t + \bar{\mathbf{x}})$ if (3.6.13) is used, $\bar{\mathbf{x}}_i = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_{it}$, $\bar{\mathbf{x}}_t = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_{it}$, $\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\bar{\mathbf{x}}_i = \frac{1}{T}\sum_{t=1}^{T}\tilde{\mathbf{x}}_t$,

and

$$\hat{\bar{\Omega}} = \frac{1}{T}\left\{\sum_{t=1}^{T}\hat{\mathbf{v}}_t^*\hat{\mathbf{v}}_t^{*\prime} + \sum_{j=1}^{T-1}k\left(\frac{j}{m}\right)\left[\sum_{t=j+1}^{T}\hat{\mathbf{v}}_t^*\hat{\mathbf{v}}_{t-j}^{*\prime} + \sum_{t=j+1}^{T}\hat{\mathbf{v}}_{t-j}^*\hat{\mathbf{v}}_t^{*\prime}\right]\right\}$$

$$\hat{\mathbf{v}}_{it}^* = \tilde{\mathbf{x}}_{it}(\tilde{y}_{it} - \tilde{\mathbf{x}}_{it}'\hat{\boldsymbol{\beta}}_{cv}), \tag{9.2.10}$$

$$\hat{\bar{\mathbf{v}}}_t^* = \frac{1}{N}\sum_{i=1}^{N}\tilde{\mathbf{x}}_{it}(\tilde{y}_{it} - \tilde{\mathbf{x}}_{it}\hat{\boldsymbol{\beta}}_{cv}) = \frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{v}}_{it}^*,$$

$\tilde{y}_{it} = (y_{it} - \bar{y}_i)$ if (3.2.8) is used or $\tilde{y}_{it} = (y_i - \bar{y}_i - \bar{y}_t + \bar{y})$ if (3.6.13) is used, and $k(\frac{j}{m}) = 1 - |\frac{j}{m}|$ if $|\frac{j}{m}| < 1$ and $k(\frac{j}{m}) = 0$ if $|\frac{j}{m}| > 1$, $m$ is an a priori chosen positive constant less than or equal to $T$. The choice of $m$ depends on how strongly an investigator thinks about the serial correlation of the error $u_{it}$.

### 9.2.2 Spatial Error Model

The log-likelihood function for the spatial error model (9.2.2) takes the form

$$-\frac{1}{2}\log\,|\,\Omega\,| - \frac{1}{2}\mathbf{v}'\Omega^{-1}\mathbf{v}, \tag{9.2.11}$$

where

$$\Omega = V \otimes I_T \tag{9.2.12}$$

if $\mathbf{v}$ is a spatial autoregressive form (9.2.3), and

$$\Omega = \tilde{V} \otimes I_T \tag{9.2.13}$$

is $\mathbf{v}$ is a spatial moving average form (9.2.4). Conditional on $\theta$ or $\delta$, the MLE of $\boldsymbol{\beta}$ is just the GLS estimator

$$\hat{\boldsymbol{\beta}} = (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}\mathbf{y}). \tag{9.2.14}$$

When $\Omega$ takes the form of (9.2.12), the log-likelihood function (9.2.11) takes the form

$$\begin{aligned} &T\,\log\,|\,I_N - \theta W\,| - \frac{NT}{2}\log\sigma_u^2 \\ &- \frac{1}{2\sigma_u^2}(\mathbf{y} - X\boldsymbol{\beta})'[(I_N - \theta W)'(I_N - \theta W)\otimes I_T](\mathbf{y} - X\boldsymbol{\beta}). \end{aligned} \tag{9.2.15}$$

Ord (1975) notes that

$$|\,I_N - \theta W\,| = \prod_{j=1}^{N}(1 - \theta\omega_j), \tag{9.2.16}$$

where $\omega_j$ are the eigenvalues of $W$, which are real even $W$ after row normalization is no longer symmetric. Substituting (9.2.16) into (9.2.15), the

log-likelihood values can be evaluated at each possible $(\theta, \boldsymbol{\beta}')$ with an iterative optimization routine. However, when $N$ is large, the computation of the eigenvalues becomes numerically unstable.

### 9.2.3 Spatial Lag Model

For the spatial lag model (9.2.1), the right-hand side, $(W \otimes I_T)\mathbf{y}$, and $\mathbf{u}$ are correlated. When $\mathbf{u} \sim N(\mathbf{0}, \sigma_u^2 I_{NT})$, the log-likelihood function of (9.2.1) is

$$\begin{aligned} & T \log \mid I_N - \rho W \mid -\frac{NT}{2}\log \sigma_u^2 \\ & - \frac{1}{2\sigma_u^2}[\mathbf{y} - \rho(W \otimes I_T)\mathbf{y} - X\boldsymbol{\beta}]'[\mathbf{y} - \rho(W \otimes I_T)\mathbf{y} - X\boldsymbol{\beta}], \mid \rho \mid < 1. \end{aligned} \tag{9.2.17}$$

When $T$ is fixed, the MLE is $\sqrt{N}$ consistent and asymptotically normally distributed under the assumption that $w_{ij}$ are at most of order $h_N^{-1}$, and the ratio $h_N/N \to 0$ as $N$ goes to infinity (Lee (2004)). However, when $N$ is large, just like the MLE for (9.2.11), the MLE for (9.2.1) is burdensome and numerically unstable (e.g., Kelejian and Prucha (2001), Lee (2004)). The $\mid I_N - \rho W \mid$ is similar in form to (9.2.16). A similar iterative optimization routine as that for (9.2.15) can be evaluated at each possible $(\rho, \boldsymbol{\beta}')$. When $N$ is large, the computation of the eigenvalues becomes numerically unstable.

The parameters $(\rho, \boldsymbol{\beta}')$ can also be estimated by the instrumental variables or generalized method of moments estimator (or two-stage least squares estimator) (Kelejian and Prucha 2001),

$$\begin{pmatrix} \hat{\rho} \\ \hat{\boldsymbol{\beta}} \end{pmatrix} = [Z'H(H'H)^{-1}H'Z]^{-1}[Z'H(H'H)^{-1}H'\mathbf{y}], \tag{9.2.18}$$

where $Z = [(W \otimes I_T)\mathbf{y}, X]$ and $H = [(W \otimes I_T)X, X]$. Lee (2003) shows that an optimal instrumental variables estimator is to let $H = [(W \otimes I_T)E\mathbf{y}, X]$, where $E\mathbf{y} = [I_{NT} - \rho(W \otimes I_T)]^{-1}X\boldsymbol{\beta}$. The construction of optimal instrumental variables requires some initial consistent estimators of $\rho$ and $\boldsymbol{\beta}$.

When $w_{ij} = O(N^{-(\frac{1}{2}+\delta)})$, where $\delta > 0$, $E((W \otimes I_T)\mathbf{y}\mathbf{u}') = o(N^{-\frac{1}{2}})$, one can ignore the correlations between $(W \otimes I_T)\mathbf{y}$ and $\mathbf{u}$. Applying the least-squares method to (9.2.1) yields a consistent and asymptotically normally distributed estimator of $(\rho, \boldsymbol{\beta}')$ (Lee 2002). However, if $W$ is "sparse," this condition may not be satisfied. For instance, in Case (1991), "neighbors" refers to households in the same district. Each neighbor is given equal weight. Suppose there are $r$ districts and $m$ members in each district, $N = mr$. Then $w_{ij} = \frac{1}{m}$ if $i$ and $j$ are in the same district and $w_{ij} = 0$ if $i$ and $j$ belong to different districts. If $r \longrightarrow \infty$ as $N \longrightarrow \infty$ and $N$ is relatively much larger than $r$ in the sample, one might regard the condition $w_{ij} = O(N^{-(\frac{1}{2}+\delta)})$ being satisfied. On the other hand, if $r$ is relatively much larger than $m$ or $\lim_{N\to\infty} \frac{r}{m} = c \neq 0$, then $w_{ij} = O(N^{-\frac{1}{2}(N+\delta)})$ cannot hold.

### 9.2.4 Spatial Error Models with Individual-Specific Effects

One can also combine the spatial approach with the error components or fixed effects specification (e.g., Kapoor, Kelejian, and Prucha 2007; Lee and Yu (2010a,b)). For instance, one may generalize the spatial error model by adding the individual-specific effects,

$$\mathbf{y} = X\boldsymbol{\beta} + (I_N \otimes \mathbf{e}_T)\boldsymbol{\alpha} + \mathbf{v}, \tag{9.2.19}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)'$. Suppose $\boldsymbol{\alpha}$ are treated as fixed constants and $\mathbf{v}$ follows a spatial error autoregressive form (9.2.3), the log-likelihood function is of the form (9.2.11) where $\Omega$ is given by (9.2.12), and $\mathbf{v} = (\mathbf{y} - X\boldsymbol{\beta} - (I_N \otimes \mathbf{e}_T)\boldsymbol{\alpha})$. Taking partial derivatives of the log-likelihood function with respect to $\boldsymbol{\alpha}$ and setting it equal to $\mathbf{0}$ yields the MLE estimates of $\boldsymbol{\alpha}$ conditional on $\boldsymbol{\beta}$ and $\theta$. Substituting the MLE estimates of $\boldsymbol{\alpha}$ conditional on $\boldsymbol{\beta}$ and $\theta$ into the log-likelihood function, we obtain the concentrated log-likelihood function

$$\begin{aligned} &-\frac{NT}{2}\log\sigma_u^2 + T\log\mid I_N - \theta W\mid \\ &\quad -\frac{1}{2\sigma_u^2}\tilde{\mathbf{v}}'[(I_N - \theta W)'(I_N - \theta W)\otimes I_T]\tilde{\mathbf{v}}, \end{aligned} \tag{9.2.20}$$

where the element $\tilde{\mathbf{v}}$, $\tilde{v}_{it} = (y_{it} - \bar{y}_i) - (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\boldsymbol{\beta}$, $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$, and $\bar{\mathbf{x}}_i = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_{it}$. In other words, the MLE of $\boldsymbol{\beta}$ is equivalent to first taking the covariance transformation of each $y_{it}$ and $\mathbf{x}_{it}$ to get rid of the individual-specific effects, $\alpha_i$, then maximizing (9.2.20) to obtain the MLE of the spatial error model with fixed individual-specific effects.

The MLE of $\boldsymbol{\beta}$ and $\theta$ are consistent when either $N$ or $T$ or both tend to infinity. However, the MLE of $\boldsymbol{\alpha}$ and $\sigma_u^2$ is consistent only if $T \longrightarrow \infty$. To obtain consistent estimate of $(\beta, \theta, \sigma_u^2)$ with finite $T$ and large $N$, Lee and Yu (2010a,b) suggest maximizing[4]

$$\begin{aligned} &-\frac{N(T-1)}{2}\log\sigma_u^2 + (T-1)\log\mid I_N - \theta W\mid \\ &\quad -\frac{1}{2\sigma_u^2}\tilde{\mathbf{v}}'[(I_N - \theta W)'(I_N - \theta W)\otimes I_T]\tilde{\mathbf{v}}. \end{aligned} \tag{9.2.21}$$

When $\alpha_i$ are treated as random and are independent of $\mathbf{u}$, The $NT \times NT$ covariance matrix of $\mathbf{v}$ takes the form

$$\Omega = \sigma_\alpha^2(I_N \otimes J_T) + \sigma_u^2((B'B)^{-1} \otimes I_T), \tag{9.2.22}$$

if $\alpha_i$ and $u_{it}$ are independent of $X$ and are i.i.d. with mean 0 and variance $\sigma_\alpha^2$ and $\sigma_u^2$, respectively, where $J_T$ is a $T \times T$ matrix with all elements equal to 1,

[4] As a matter of fact, (9.2.21) is derived by the transformation matrix $Q^*$ where $Q^* = [F, \frac{1}{\sqrt{T}}\iota_T]$, where $F$ is the $T \times (T-1)$ eigenvector matrix of $Q = I_T - \frac{1}{T}\mathbf{e}_T\mathbf{e}_T'$ that correspond to the eigenvalues of 1.

$B = (I_N - \theta W)$. Using the results in Wansbeek and Kapteyn (1978), one can show that (e.g., Baltagi et al. 2007)

$$\Omega^{-1} = \sigma_u^{-2} \left\{ \frac{1}{T} J_T \otimes \left[ T\phi I_N + (B'B)^{-1} \right]^{-1} + E_T \otimes B'B \right\}, \tag{9.2.23}$$

where $E_T = I_T - \frac{1}{T} J$ and $\phi = \frac{\sigma_\alpha^2}{\sigma_u^2}$,

$$| \Omega | = \sigma_u^{2NT} \mid T\phi I_N + (B'B)^{-1} \mid \cdot \mid (B'B)^{-1} \mid^{T-1} . \tag{9.2.24}$$

The MLE of $\boldsymbol{\beta}, \theta, \sigma_u^2$, and $\sigma_\alpha^2$ can then be derived by substituting (9.2.23) and (9.2.24) into the log-likelihood function (e.g., Anselin 1988, p. 154).

The FGLS estimator (9.2.14) of the random-effects spatial error model $\boldsymbol{\beta}$ is to substitute initial consistent estimates of $\phi$ and $\theta$ into (9.2.23). Kapoor et al. (2007) propose a method of moments estimation with moment conditions in terms of $(\theta, \sigma_u^2, \tilde{\sigma}^2 = \sigma_u^2 + T\sigma_\alpha^2)$.

### 9.2.5 Spatial Lag Model with Individual-Specific Effects

For the spatial lag model with individual-specific effects,

$$\mathbf{y} = \rho(W \otimes I_T)\mathbf{y} + X\boldsymbol{\beta} + (I_N \otimes \mathbf{e}_T)\boldsymbol{\alpha} + \mathbf{u}. \tag{9.2.25}$$

If $\boldsymbol{\alpha}$ is treated as fixed constants, the log-likelihood function of (9.2.25) is of similar form as (9.2.20)

$$\begin{aligned} & T \log \mid I_N - \rho W \mid - \frac{NT}{2} \log \sigma_u^2 \\ & - \frac{1}{2\sigma_u^2} \{ [\mathbf{y} - \rho(W \otimes I_T)\mathbf{y} - X\boldsymbol{\beta} - (I_N \otimes \mathbf{e}_T)\boldsymbol{\alpha}]' \\ & \qquad [\mathbf{y} - \rho(W \otimes I_T)\mathbf{y} - X\boldsymbol{\beta} - (I_N \otimes \mathbf{e}_T)\boldsymbol{\alpha}] \}. \end{aligned} \tag{9.2.26}$$

The MLE of $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ can be computed similarly as that of (9.2.20).

When $\alpha_i$ are treated as randomly distributed across $i$ with constant variance $\sigma_\alpha^2$ and independent of $X$, then

$$\begin{aligned} & E \left\{ (\mathbf{u} + (I_N \otimes \mathbf{e}_T)\boldsymbol{\alpha})(\mathbf{u} + (I_N \otimes \mathbf{e}_T)\boldsymbol{\alpha})' \right\} \\ & \quad = I_N \otimes V^*, \end{aligned} \tag{9.2.27}$$

where $V^* = \sigma_u^2 I_T + \sigma_\alpha^2 \mathbf{e}_T \mathbf{e}_T'$. The MLE or quasi-MLE for the spatial lag model (9.2.1) can be obtained by maximizing

$$\begin{aligned} & T \log \mid I_N - \rho W \mid - \frac{N(T-1)}{2} \log \sigma_u^2 - \frac{N}{2} \log\left(\sigma_u^2 + T\sigma_\alpha^2\right) \\ & \qquad - \frac{1}{2} (\mathbf{y}^* - X\boldsymbol{\beta})'(I_N \otimes V^{*-1})(\mathbf{y}^* - X\boldsymbol{\beta}), \end{aligned} \tag{9.2.28}$$

where $\mathbf{y}^* = (I_{NT} - \rho(W \otimes I_T)\mathbf{y}$. Conditional on $\rho$, $\sigma_u^2$, and $\sigma_\alpha^2$, the MLE of $\boldsymbol{\beta}$ is the GLS estimator

$$\hat{\boldsymbol{\beta}} = (X'[I_N \otimes V^{*-1}]X)^{-1}(X'(I_N \otimes V^{*-1})(I_{NT} - \rho(W \otimes I_T))\mathbf{y}, \quad (9.2.29)$$

where $V^{*-1}$ is given by (3.3.7). Kapoor et al. (2007) have provided moment conditions to obtain initial consistent estimates $\sigma_u^2$, $\sigma_\alpha^2$, and $\rho$.

One can also combine the random individual-specific effects specification of $\boldsymbol{\alpha}$ with a spatial specification for the error $\mathbf{v}$. For instance, we can let

$$\mathbf{y} = \rho(W_1 \otimes I_T)\mathbf{y} + X\boldsymbol{\beta} + (I_N \otimes \mathbf{e}_T)\boldsymbol{\alpha} + \mathbf{v}, \quad (9.2.30)$$

with

$$\mathbf{v} = \theta(W_2 \otimes I_T)\mathbf{v} + \mathbf{u}, \quad (9.2.31)$$

where $W_1$ and $W_2$ are $N \times N$ spatial weights matrices and $\boldsymbol{\alpha}$ is an $N \times 1$ vector of individual effects. Let $S(\rho) = I_N - \rho W_1$ and $R(\theta) = I_N - \theta W_2$. Under the assumption that $u_{it}$ is independently normally distributed, the log-likelihood function of (9.2.30) takes the form

$$\begin{aligned} \log L = &- \frac{NT}{2} \log \sigma_u^2 + T \log \mid S(\rho) \mid + T \log \mid R(\theta) \mid \\ &- \frac{1}{2}\tilde{\mathbf{v}}^{*'}\tilde{\mathbf{v}}^*, \end{aligned} \quad (9.2.32)$$

where

$$\tilde{\mathbf{v}}^* = [R(\theta) \otimes I_T][(S(\rho) \otimes I_T)\mathbf{y} - X\boldsymbol{\beta} - (I_N \otimes \mathbf{e}_T)\boldsymbol{\alpha}]. \quad (9.2.33)$$

The MLE (or quasi-MLE if $u$ is not normally distributed) can be computed similarly as that of (9.2.20). For details, see Lee and Yu (2010a,b).

### 9.2.6 Spatial Dynamic Panel Data Models

Consider a dynamic panel data model of the form

$$\mathbf{y} = \mathbf{y}_{-1}\gamma + X\boldsymbol{\beta} + (I_N \otimes \mathbf{e}_T)\boldsymbol{\alpha} + \mathbf{v} \quad (9.2.34)$$

where $\mathbf{y}_{-1}$ denotes the $NT \times 1$ vector of $y_{it}$ lagged by one period, $\mathbf{y}_{-1} = (y_{10}, \ldots, y_{1,T-1}, \ldots, y_{N,T-1})'$, $X$ denotes the $NT \times K$ matrix of exogenous variables, $X = (\mathbf{x}_{it}')$, and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)'$ denotes the $N \times 1$ fixed individual-specific effects. If the error term follows a spatial autoregressive form of (9.2.3), even $\mid \gamma \mid < 1$, there could be spatial cointegration if $\gamma + \theta = 1$ while $\gamma \neq 1$ (Yu and Lee (2010)). Yu et al. (2012) show that the MLE of $(\gamma, \theta, \boldsymbol{\beta}, \boldsymbol{\alpha})$ are $\sqrt{NT}$ consistent with $T$ tends to infinity. However, if $\gamma + \theta = 1$, then the asymptotic covariance matrix of the MLE is singular when the estimator is multiplied by the scale factor $\sqrt{NT}$ because the sum of the spatial and dynamic effects converge at a higher rate (e.g., Yu and Lee (2010)).

Yu et al. (2012) also consider the estimation of a dynamic spatial lag model with the spatial-time effect,

$$\begin{aligned}\mathbf{y} =&(\rho W \otimes I_T)\mathbf{y} + \mathbf{y}_{-1}\gamma + (\rho^* W \otimes I_T)\mathbf{y}_{-1} + X\boldsymbol{\beta} \\ &+ (I_N \otimes \mathbf{e}_T)\boldsymbol{\alpha} + \mathbf{v}\end{aligned} \tag{9.2.35}$$

Model (9.2.35) is stable if $\gamma + \rho + \rho^* < 1$ and spatially cointegrated if $\gamma + \rho + \rho^* = 1$ but $\gamma \neq 1$. They develop the asymptotics of (quasi)–MLE when both $N$ and $T$ are large and propose a bias correction formula.

## 9.3 FACTOR APPROACH

Another approach to model cross-sectional dependence is to assume that the error follows a linear factor model,

$$v_{it} = \sum_{j=1}^{r} b_{ij} f_{jt} + u_{it} = \mathbf{b}_i' \mathbf{f}_t + u_{it}, \tag{9.3.1}$$

where $\mathbf{f}_t = (f_{1t}, \ldots, f_{rt})'$ is a $r \times 1$ vector of random factors with mean 0, $\mathbf{b}_i = (b_{i1}, \ldots, b_{ir})'$ is a $r \times 1$ nonrandom factor loading coefficient (to avoid with $u_{it}$ nonseparability), $u_{it}$ represents the effects of idiosyncratic shocks, which is independent of $\mathbf{f}_t$ and is independently distributed across $i$ with constant variance over $t$.

Factor models have been suggested as an effective way of synthesizing information contained in large data sets (e.g., Bai 2003, 2009; Bai and Ng 2002). The conventional time-specific effects model (e.g., Chapter 3) is a special case of (9.3.1) when $r = 1$ and $b_i = b_\ell$ for all $i$ and $\ell$. An advantage of factor model over the spatial approach is that there is no need to prespecify the strength of correlations between units $i$ and $j$.

Let $\mathbf{v}_t = (v_{1t}, \ldots, v_{Nt})'$, then

$$\mathbf{v}_t = B\mathbf{f}_t + \mathbf{u}_t, \tag{9.3.2}$$

where $B = (b_{ij})$ is the $N \times r$ factor loading matrix, and $\mathbf{u}_t = (\mathbf{u}_{1t}, \ldots, \mathbf{u}_{Nt})'$. Then

$$E\mathbf{v}_t\mathbf{v}_t' = B(E\,\mathbf{f}_t\,\mathbf{f}_t')B' + D \tag{9.3.3}$$

where $D$ is an $N \times N$ diagonal covariance matrix of $\mathbf{u}_t$. The covariance between $v_{it}$ and $v_{\ell t}$ is given by

$$Ev_{it}v_{\ell t} = \mathbf{b}_i'(E\,\mathbf{f}_t\,\mathbf{f}_t')\mathbf{b}_\ell \tag{9.3.4}$$

However, $B\mathbf{f}_t = BAA^{-1}\mathbf{f}_t$ for any $r \times r$ nonsingular matrix A. That is, without $r^2$ normalizations, (or prior restrictions) $\mathbf{b}_i$ and $\mathbf{f}_t$ are not uniquely determined. A common normalization is to assume $E\,\mathbf{f}_t\,\mathbf{f}_t' = I_r$. Nevertheless, even with this assumption, it only yields $\frac{r(r+1)}{2}$ restrictions on $B$. $B$ is only identifiable up to an orthonormal transformation, that is, $BCC'B' = BB'$ for any $r \times r$

orthonormal matrix (e.g., Anderson (1985)). Additional $\frac{r(r-1)}{2}$ restrictions are needed, say $B'B$ diagonal.[5]

For given $v_{it}$ and $r$, $B$ and $F$ can be estimated by minimizing

$$\begin{aligned}\tilde{V}(r) &= (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(v_{it}-\mathbf{b}_i'\,\mathbf{f}_t)^2 \\ &= (NT)^{-1}tr[(V-FB')(V'-BF')],\end{aligned} \tag{9.3.5}$$

where $V=(\mathbf{v}_1,\ldots,\mathbf{v}_N)$ is a $T\times N$ matrix with $\mathbf{v}_i=(v_{i1},\ldots,v_{iT})'$, $F=(\mathbf{f}_1,\ldots,\mathbf{f}_r)$ is a $T\times r$ matrix with $\mathbf{f}_j=(f_{j1},\ldots,f_{jT})'$. Taking partial derivatives of (9.3.5) with respect to $B$, setting them equal to 0, and using the normalization $\frac{1}{T}F'F=I_r$, we obtain

$$B=\frac{1}{T}V'F. \tag{9.3.6}$$

Substituting (9.3.6) into (9.3.5), minimizing (9.3.5) is equivalent to maximizing $tr[F'(VV')F]$ subject to $B'B$ being diagonal. Therefore, the $T\times r$ common factor $F=(\mathbf{f}_1,\ldots,\mathbf{f}_r)$ is estimated as $\sqrt{T}$ times the eigenvectors corresponding to the $r$ largest eigenvalues of the $T\times T$ matrix $\sum_{i=1}^{N}\mathbf{v}_i\mathbf{v}_i'$, denoted by $\hat{F}$ Anderson (1985)). Given $\hat{F}$, the factor loading matrix $B$ can be estimated as $\hat{B}=\frac{1}{T}V'\hat{F}$.

To identify $r$, Bai and Ng (2002) note that if $\lim_{N\to\infty}\frac{1}{N}B'B$ converges to a nonsingular $r\times r$ constant matrix $A$, the largest $r$ eigenvalues of (9.3.3) are of order $N$ because the $r$ positive eigenvalues $B'B$ are equal to those of $BB'$. In other words, the $r$ common factors, $\mathbf{f}_t$, practically drive all $N\times 1$ errors, $\mathbf{v}_t$. Therefore, when $r$ is unknown, under the assumption that $\lim_{N\to\infty}\frac{1}{N}B'B=A$, Bai and Ng (2002) suggest using the criterion

$$\min_{k}\ \mathrm{PC}(k)=\hat{V}(k)+kg(N,T), \tag{9.3.7}$$

or

$$\min_{k}\ \mathrm{IC}(k)=ln\hat{V}(k)+kg(N,T), \tag{9.3.8}$$

where $k<\min(N,T)$,

$$\hat{V}(k)=\min_{B^k,\,F^k}\ (NT)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(\hat{v}_{it}-\hat{\mathbf{b}}_i^{k\prime}\hat{\mathbf{f}}_t^{k})^2, \tag{9.3.9}$$

[5] Even in this case uniqueness is only up to a sign change. For instance, $-\mathbf{f}_t$ and $-B$ also satisfy the restrictions. However, the covariance between $v_{it}$ and $v_{jt}$ remains the same, $E(v_{it}v_{jt})=\mathbf{b}_i'\mathbf{b}_j=\mathbf{b}_i^{*\prime}\mathbf{b}_j^{*}=\mathbf{b}_i'CC'\mathbf{b}_j$ for any $r\times r$ orthonormal matrix.

where $\hat{v}_{it}$ is the estimated $v_{it}$, $\hat{\mathbf{f}}_t^k$ denotes the $k$-dimensional estimated factor at time $t$, $\hat{\mathbf{b}}_i^k$ denotes the estimated loading factor for the $i$th individual, and $g(N,T)$ is a penalty function satisfying (1) $g(N,T) \longrightarrow 0$ and (2) $\min\{N,T\}g(N,T) \longrightarrow \infty$ as $N,T \longrightarrow \infty$ to select $r$. The reason for using these criteria is because $\hat{V}(k)$ decreases with $k$ and $\hat{V}(k) - \hat{V}(r)$ converges to a nonzero positive number for $k < r$ and $\hat{V}(k) - \hat{V}(r) \longrightarrow 0$ at certain rate, say $C(N,T)$. Choosing a penalty function $kg(N,T)$ increases with $k$. When $g(N,T)$ diminishes to 0 at a rate slower than $C(N,T)$, the penalty will eventually become dominant and prevent choosing a $k > r$. Bai and Ng (2002) show that $C(N,T) = \min\{N,T\}$ when $u_{it}$ satisfy the stationarity assumption (allowing for weak serial and cross-sectional dependence). Therefore, they propose the specific forms of $g(N,T)$ as $\hat{\sigma}_u^2 \cdot \frac{N+T}{NT} ln\left(\frac{NT}{N+T}\right)$ or $\hat{\sigma}_u^2 \cdot \frac{N+T}{NT} ln\,(\min(N,T))$, etc.[6] They show that when both $N$ and $T \longrightarrow \infty$, the criterion (9.3.8) or (9.3.7) selects $\hat{k} \longrightarrow r$ with probability 1. Moreover, $\hat{\mathbf{f}}_t \longrightarrow \mathbf{f}_t$ if $(\sqrt{T}/N) \longrightarrow \infty$.

To estimate a regression model of the form

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + v_{it}, \qquad \begin{matrix} i = 1,\ldots,N, \\ t = 1,\ldots,T, \end{matrix} \tag{9.3.10}$$

where $v_{it}$ follows (9.3.1), Bai (2009) and Pesaran (2006) suggest the least-squares regression of the model

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{b}_i'\,\mathbf{f}_t + u_{it}, \tag{9.3.11}$$

subject to $\frac{1}{T}\sum_{t=1}^{T} \mathbf{f}_t\,\mathbf{f}_t' = I_r$ and $B'B$ being diagonal. Noting that conditional on $\mathbf{f}_t$, the least squares estimator of $\boldsymbol{\beta}$ is equal to

$$\hat{\boldsymbol{\beta}} = (\sum_{i=1}^{N} X_i' M X_i)^{-1}(\sum_{i=1}^{N} X_i' M \mathbf{y}_i), \tag{9.3.12}$$

where $\mathbf{y}_i$ and $X_i$ denote the stacked $T$ time series observations of $y_{it}$ and $\mathbf{x}_{it}'$, $M = I - F(F'F)^{-1}F'$ where $F$ is the $T \times r$ matrix of $F = (\mathbf{f}_t')$. Conditional on $\boldsymbol{\beta}$, the residual $v_{it}$ is a pure factor structure (9.3.1). The least squares estimator of $F$ is equal to the first $r$ eigenvectors (multiplied by $\sqrt{T}$ due to the restriction $F^{r'}F^r/T = I$) associated with the largest $r$ eigenvalues of the matrix (Anderson (1985)),

$$\frac{1}{N}\sum_{i=1}^{N}(\mathbf{y}_i - X_i\boldsymbol{\beta})(\mathbf{y}_i - X_i\boldsymbol{\beta})', \tag{9.3.13}$$

$$\left[\frac{1}{NT}\sum_{i=1}^{N}(\mathbf{y}_i - X_i\boldsymbol{\beta})(\mathbf{y}_i - X_i\boldsymbol{\beta})'\right]\hat{F} = \hat{F}\Lambda, \tag{9.3.14}$$

[6] When $u_{it}$ exhibit considerable serial correlation and the sample size is not sufficiently large, the Bai and Ng (2002) criterion may overfit (e.g.,Greenaway-McGrevy, Han, and Sul 2012).

where $\Lambda$ is a diagonal matrix that consists of the $r$ largest eigenvalues of (9.3.13) multiplied by $\frac{1}{T}$. Conditional on $(\hat{\boldsymbol{\beta}}, \hat{F})$, (9.3.6) leads to

$$\hat{B}' = T^{-1}\left[\hat{F}'(\mathbf{y}_1 - X_1\hat{\boldsymbol{\beta}}), \ldots, \hat{F}'(\mathbf{y}_N - X_N\hat{\boldsymbol{\beta}})\right] = \frac{1}{T}\hat{V}'\hat{F}. \quad (9.3.15)$$

Iterating between (9.3.12), (9.3.14), and (9.3.15) leads to the least-squares estimator of (9.3.11) as if $\mathbf{f}$ were observable.

When both $N$ and $T$ are large, the least-squares estimator (9.3.12) is consistent and asymptotically normally distributed with covariance matrix

$$\sigma_u^2\left(\sum_{i=1}^{N} X_i' M X_i\right)^{-1} \quad (9.3.16)$$

if $u_{it}$ is independently, identically distributed with mean 0 and constant variance $\sigma_u^2$. However, if $u_{it}$ is heteroscedastic and cross-sectionally or serially correlated, when $\frac{T}{N} \to c \neq 0$ as $N, T \to \infty$, $\sqrt{NT}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is asymptotically biased of the form

$$\left(\frac{T}{N}\right)^{\frac{1}{2}} C + \left(\frac{N}{T}\right)^{\frac{1}{2}} D^*, \quad (9.3.17)$$

where $C$ denotes the bias induced by heteroscadasticity and cross-sectional correlation and $D^*$ denotes the bias induced by serial correlation and heteroscedasticity of $u_{it}$. Bai (2009) has provided the formulas for constructing the bias-corrected estimator.

To estimate a model with both additive and interactive effects,

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \mathbf{b}_i'\,\mathbf{f}_t + u_{it}, \quad (9.3.18)$$

in addition to the normalization conditions, $F'F = I_r$ and $B'B$ diagonal, we also need to impose the restriction $\sum_{i=1}^{N}\alpha_i = 0$, $\sum_{i=1}^{N}\mathbf{b}_i = \mathbf{0}$, $\sum_{t=1}^{T}\mathbf{f}_t = \mathbf{0}$, to obtain a unique solution of $(\beta, \alpha_i, \mathbf{b}_i, \mathbf{f}_t)$ (Bai 2009, p. 1253). Just like the standard fixed-effects estimator (Chapter 3) we can first take individual observations from its time series mean, $\tilde{y}_{it} = y_{it} - \bar{y}_i$, $\tilde{\mathbf{x}}_{it} = \mathbf{x}_{it} - \bar{\mathbf{x}}_i$, to get rid of $\alpha_i$ from (9.3.18), and then iteratively estimate $\boldsymbol{\beta}$ and $F$ by

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N}\tilde{X}_i'\tilde{M}\tilde{X}_i\right)^{-1}\left(\sum_{i=1}^{N}\tilde{X}_i\tilde{M}\tilde{\mathbf{y}}_i\right), \quad (9.3.19)$$

where $\tilde{\mathbf{y}}_i, \tilde{X}_i$ denote the stacked $T$ time series observations of $\tilde{y}_{it}$ and $\tilde{\mathbf{x}}_{it}'$, $\tilde{M} = I - \hat{F}(\hat{F}'\hat{F})^{-1}\hat{F}'$, and $\hat{F}$ is the $T \times r$ matrix consisting of the first $r$ eigenvectors (multiplied by $\sqrt{T}$) associated with the $r$ largest eigenvalues of the matrix

$$\frac{1}{NT}\sum_{i=1}^{N}(\tilde{\mathbf{y}}_i - \tilde{X}_i\hat{\boldsymbol{\beta}})(\tilde{\mathbf{y}}_i - \tilde{X}_i\hat{\boldsymbol{\beta}})'. \quad (9.3.20)$$

After convergent solutions of $\hat{\boldsymbol{\beta}}$ and $\hat{F}$ are obtained, one can obtain $\hat{\alpha}_i$ and $\hat{B}'$ by

$$\hat{\alpha}_i = \bar{y}_i - \bar{\mathbf{x}}_i'\hat{\boldsymbol{\beta}}, \tag{9.3.21}$$

$$\hat{B}' = T^{-1}[\hat{F}'(\tilde{\mathbf{y}}_1 - \tilde{X}_1\hat{\boldsymbol{\beta}}), \ldots, \hat{F}'(\tilde{\mathbf{y}}_N - \tilde{X}_N\hat{\boldsymbol{\beta}})]. \tag{9.3.22}$$

Ahn, Lee, and Schmidt (2001, 2013) have proposed a nonlinear GMM method to estimate a linear panel data model with interactive effects (9.3.1). For ease of exposition, suppose $r = 1$. Let $\theta_t = \frac{f_t}{f_{t-1}}$, then

$$(y_{it} - \theta_t y_{i,t-1}) = \mathbf{x}_{it}'\boldsymbol{\beta} - \mathbf{x}_{i,t-1}'\boldsymbol{\beta}\theta_t + (u_{it} - \theta_t u_{i,t-1}),\ t = 2, \ldots, T. \tag{9.3.23}$$

It follows that

$$E[\mathbf{x}_i(u_{it} - \theta_t u_{i,t-1})] = \mathbf{0} \tag{9.3.24}$$

Let $W_i = I_{T-1} \otimes \mathbf{x}_i$,

$$\ominus = \begin{matrix}(T-1)\times(T-1)\end{matrix}\begin{bmatrix} \theta_2 & 0 & \ldots & \ldots & 0 \\ 0 & \theta_3 & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & 0 \\ 0 & \ldots & \ldots & \ldots & \theta_T \end{bmatrix},$$

$$\tilde{\mathbf{u}}_i = (u_{i2}, \ldots, u_{iT})',\ \tilde{\mathbf{u}}_{i,-1} = (u_{i1}, \ldots, u_{i,T-1})'.$$

Then a GMM estimator of $\boldsymbol{\beta}$ and $\ominus$ can be obtained from the moment conditions,

$$E[W_i(\tilde{\mathbf{u}}_i - \ominus\tilde{\mathbf{u}}_{i,-1})] = \mathbf{0}. \tag{9.3.25}$$

The nonlinear GMM estimator is consistent and asymptotically normally distributed when $N \to \infty$ under fixed $T$ even $u_{it}$ is serially correlated and heteroscedastic. However, the computation can be very cumbersome when $r > 1$. For instance, if $r = 2$, in addition to letting $\theta_t = \frac{f_{it}}{f_{i,t-1}}$, we need to introduce additional parameters $\delta_t = f_{2t} - f_{2,t-1}\theta_t$ and to take the quasi-difference of $(y_{it} - \theta_t y_{i,t-1})$ equation one more time to eliminate the factor error.

**Remark 9.3.1:** The unique determination of $\mathbf{b}_i$ and $E\,\mathbf{f}_t\,\mathbf{f}_t'$ is derived under the assumption that $v_{it}$ are observable. The derivation of the least-squares regression of (9.3.11) is based on the assumption that (9.3.11) is identifiable from $(y_{it}, \mathbf{x}_{it}')$. The identification conditions for (9.3.11) remain to be explored. Neither does it appear feasible to simultaneously estimate $\boldsymbol{\beta}$, $B$ and $\mathbf{f}_t$ when $N$ is large. On the other hand, the two-step procedure of (9.3.12)–(9.3.14) depends on the possibility of getting initially consistent estimator of $\boldsymbol{\beta}$.

## 9.4 GROUP MEAN AUGMENTED (COMMON CORRELATED EFFECTS) APPROACH TO CONTROL THE IMPACT OF CROSS-SECTIONAL DEPENDENCE

The Frisch-Waugh FGLS approach of iteratively estimating (9.3.12) and (9.3.14) (or (9.3.15) and (9.3.16)) may work for the factor approach only if both $N$ and $T$ are large. However, if $N$ is large, the implementation of FGLS is cumbersome. Nevertheless, when $N \longrightarrow \infty$, $\bar{u}_t = \frac{1}{N}\sum_{i=1}^{N} u_{it} \longrightarrow 0$, model (9.2.2) and (9.3.2) (or (9.3.11)) imply that

$$\bar{\mathbf{b}}' \mathbf{f}_t \simeq \bar{y}_t - \bar{\mathbf{x}}_t' \boldsymbol{\beta}, \tag{9.4.1}$$

where $\bar{\mathbf{b}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{b}_i$, $\bar{y}_t = \frac{1}{N}\sum_{i=1}^{N} y_{it}$ and $\bar{\mathbf{x}}_t = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_{it}$. If $\mathbf{b}_i' \mathbf{f}_t = c_i \bar{\mathbf{b}}' \mathbf{f}_t$, for all $t$ or if $\mathbf{f}_t$ can be approximated by linear combinations of $\bar{y}_t$ and $\bar{\mathbf{x}}_t$ ((9.4.1)), instead of estimating $\hat{\mathbf{f}}_t$, Pesaran (2006) suggests a simple approach to filter out the cross-sectional dependence by augmenting (9.3.18) by $\bar{y}_t$ and $\bar{\mathbf{x}}_t$,

$$y_{it} = \mathbf{x}_{it}' \boldsymbol{\beta} + \alpha_i + \bar{y}_t c_i + \bar{\mathbf{x}}_t' \mathbf{d}_i + e_{it}. \tag{9.4.2}$$

The pooled estimator,

$$\hat{\boldsymbol{\beta}}^* = \left( \sum_{i=1}^{N} w_i X_i' M^* X_i \right)^{-1} \left( \sum_{i=1}^{N} w_i X_i' M^* \mathbf{y}_i \right) \tag{9.4.3}$$

is consistent and asymptotically normally distributed when $N \to \infty$ and $T$ either fixed or $\to \infty$, where $w_i = \frac{\sigma_i^2}{\sum_{j=1}^{N} \sigma_j^2}$, $\sigma_j^2 =$ Var $(u_{jt})$, $M^* = (I - H(H'H)^{-1}H')$, $H = (\mathbf{e}, \bar{\mathbf{y}}, \bar{X})$ and $\bar{\mathbf{y}}$, $\bar{X}$ are $T \times 1$ and $T \times K$ stacked $\bar{y}_t$ and $\bar{\mathbf{x}}_t'$, respectively. Pesaran (2006) called (9.4.3) the common correlated effects pooled estimator (CCEP). The limited Monte Carlo studies conducted by Westerlund and Urbain (2012) appear to show that the Pesaran (2006) CCEP estimator of $\boldsymbol{\beta}$ (9.4.3) is less biased than the Bai (2009) iterated least-squares estimator (9.3.12).

Kapetanios, Pesaran, and Yamagata (2011) further show that the cross-sectional average-based method is robust to a wide variety of data-generating processes. For instance, for the error process generated by a multifactor error structure (9.3.1), whether the unobservable common factors $\mathbf{f}_t$ follow $I(0)$ or unit root processes, the asymptotic properties of (9.4.3) remain similar.

**Remark 9.4.1:** The advantage of Pesaran's (2006) cross-sectional mean-augmented approach to take account the cross-sectional dependence is its simplicity. However, there are restrictions on its application. The method works when $\mathbf{b}_i' \mathbf{f}_t = c_i \bar{\mathbf{b}}' \mathbf{f}_t$ for all $t$ or if $\mathbf{f}_t$ can be considered as linear combinations of $\bar{y}_t$ and $\bar{\mathbf{x}}_t$. It is hard to ensure $\mathbf{b}_i' \mathbf{f}_t = c_i \bar{\mathbf{b}}' \mathbf{f}_t$ if $r > 1$. For instance, consider the case that $r = 2$, $\mathbf{b}_i = (1, 1)'$, $\bar{\mathbf{b}} = (2, 0)'$, $\mathbf{f}_t = (1, 1)'$, then $\mathbf{b}_i' \mathbf{f}_t = \bar{\mathbf{b}}' \mathbf{f}_t = 2$. However, if $\mathbf{f}_s = (2, 0)'$, then $\mathbf{b}_i' \mathbf{f}_s = 2$ while $\bar{\mathbf{b}}' \mathbf{f}_s = 4$. If $\mathbf{b}_i' \mathbf{f}_t = c_{it} \bar{\mathbf{b}}' \mathbf{f}_t$, (9.4.2) does not approximate (9.3.11) and (9.4.3) is not consistent if $\mathbf{f}_t$ is

correlated with $\mathbf{x}_{it}$. If $\mathbf{b}_i' \mathbf{f}_t = c_{it}\bar{\mathbf{b}}' f_t$, additional assumptions are needed to approximate $\mathbf{b}_i' \mathbf{f}_t$. For instance, Pesaran (2006) assumes that

$$\mathbf{x}_{it} = \Gamma_i \mathbf{f}_t + \boldsymbol{\epsilon}_{it}, \tag{9.4.4}$$

$$E(\boldsymbol{\epsilon}_{it} u_{it}) = \mathbf{0}. \tag{9.4.5}$$

Then

$$\begin{aligned} \mathbf{z}_{it} = \begin{pmatrix} y_{it} \\ \mathbf{x}_{it} \end{pmatrix} &= \begin{pmatrix} \boldsymbol{\beta}'\Gamma_i + \mathbf{b}_i' \\ \Gamma_i \end{pmatrix} \mathbf{f}_t + \begin{pmatrix} \boldsymbol{\beta}'\boldsymbol{\epsilon}_{it} + u_{it} \\ \boldsymbol{\epsilon}_{it} \end{pmatrix} \\ &= C_i \mathbf{f}_t + \mathbf{e}_{it}. \end{aligned} \tag{9.4.6}$$

It follows that

$$\bar{\mathbf{z}}_t = \frac{1}{N} \sum_{i=1}^{N} \mathbf{z}_{it} = \bar{C} \mathbf{f}_t + \bar{\mathbf{e}}_t, \tag{9.4.7}$$

where $\bar{C} = \frac{1}{N} \sum_{i=1}^{N} C_i$, $\bar{\mathbf{e}}_t = \begin{pmatrix} \boldsymbol{\beta}'(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\epsilon}_{it}) + \bar{u}_t \\ \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\epsilon}_{it} \end{pmatrix}$. If $r \leq k+1$, $\bar{C}$ is of rank $r$ and $\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\epsilon}_{it} \longrightarrow \mathbf{0}$ (or $\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\epsilon}_{it} \longrightarrow$ a constant vector) as $N \longrightarrow \infty$, then

$$\mathbf{f}_t \simeq (\bar{C}'\bar{C})^{-1} \bar{C}' \bar{\mathbf{z}}_t. \tag{9.4.8}$$

Then model (9.3.11) is formally identical to (9.4.2) when $(C_i, \mathbf{d}_i') = \mathbf{b}_i'(\bar{C}'\bar{C})^{-1}\bar{C}'$.

However, under (9.4.4), (9.4.5), and the additional assumption that

$$\text{Cov}\,(\Gamma_i, \mathbf{b}_i) = \mathbf{0}, \tag{9.4.9}$$

one can simply obtain a consistent estimator of $\boldsymbol{\beta}$ by adding time dummies to (9.1.1). The least-squares dummy variable estimator of $\boldsymbol{\beta}$ is equivalent to the within (time) estimator of (see Chapter 3, Section 3.2)

$$(y_{it} - \bar{y}_t) = (\mathbf{x}_{it} - \bar{\mathbf{x}}_t)'\boldsymbol{\beta} + (v_{it} - \bar{v}_t), \tag{9.4.10}$$

where

$$v_{it} - \bar{v}_t = (\mathbf{b}_i - \bar{\mathbf{b}})' \mathbf{f}_t + (u_{it} - \bar{u}_t),$$

$$\mathbf{x}_{it} - \bar{\mathbf{x}}_t = (\Gamma_i - \bar{\Gamma}) \mathbf{f}_t + (\boldsymbol{\epsilon}_{it} - \bar{\boldsymbol{\epsilon}}_t),$$

$$\bar{y}_t = \frac{1}{N} \sum_{i=1}^{N} y_{it}, \bar{\mathbf{x}}_t = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_{it}, \bar{\mathbf{b}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{b}_i, \bar{\Gamma} = \frac{1}{N} \sum_{i=1}^{N} \Gamma_i,$$

$$\bar{u}_t = \frac{1}{N} \sum_{i=1}^{N} u_{it},$$

and $\bar{\boldsymbol{\epsilon}}_t = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\epsilon}_{it}$. Under (9.4.9),

$$\begin{aligned}&\text{Cov}\,(\mathbf{x}_{it} - \bar{\mathbf{x}}_t,\, v_{it} - \bar{v}_t)\\ &\quad = E\left\{(\Gamma_i - \bar{\Gamma})(\mathbf{b}_i - \bar{\mathbf{b}})'\right\}\ \text{Cov}\,(\mathbf{f}_t)\,\text{Cov}\,(\boldsymbol{\epsilon}_{it}, u_{it}) = \mathbf{0}. \qquad (9.4.11)\end{aligned}$$

Therefore, as $N \to \infty$, the least-squares estimator of (9.4.10),

$$\hat{\boldsymbol{\beta}}_{cv} = \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it} - \bar{\mathbf{x}}_t)(\mathbf{x}_{it} - \bar{\mathbf{x}}_t)'\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it} - \bar{\mathbf{x}}_t)(y_{it} - \bar{y}_t)\right] \qquad (9.4.12)$$

is consistent and asymptotically normally distributed with covariance matrix

$$\text{Cov}\,(\hat{\boldsymbol{\beta}}_{cv}) = \sigma_u^2\left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it} - \bar{\mathbf{x}}_t)(\mathbf{x}_{it} - \bar{\mathbf{x}}_t)'\right]^{-1}. \qquad (9.4.13)$$

(Coakley, Fuertes, and Smith 2006; Sarafidis and Wansbeek 2012). However, if (9.4.9) does not hold, (9.4.12) exhibits large bias and large size distortion (Sarafidis and Wansbeek 2012).

## 9.5 TEST OF CROSS-SECTIONAL INDEPENDENCE

Many of the conventional panel data estimators that ignore cross-sectional dependence are inconsistent even when $N \to \infty$ if $T$ is finite. Modeling cross-sectional dependence is much more complicated than modeling time series dependence. So is the estimation of panel data models in the presence of cross-sectional dependence. Therefore, it could be prudent to first test cross-sectional independence and only embark on estimating models with cross-sectional dependence if the tests reject the null hypothesis of no cross-sectional dependence.

### 9.5.1 Linear Model

Consider a linear model,

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + v_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T. \end{matrix} \qquad (9.5.1)$$

The spatial approach assumes a known correlation pattern among cross-sectional units, $W$. Under the null of cross-sectional independence, $\theta = 0$ for any $W$. Therefore, a test for spatial effects is a test of the null hypothesis $H_0 : \theta = 0$ (or $\delta = 0$). Burridge (1980) derives the Lagrange multiplier test statistic for model (9.2.2) or (9.2.3),

$$\tau = \frac{[\hat{\mathbf{v}}'(W \otimes I_T)\hat{\mathbf{v}}/(\hat{\mathbf{v}}'\hat{\mathbf{v}}/NT)]^2}{tr[(W^2 \otimes I_T) + (W'W \otimes I_T)]} \qquad (9.5.2)$$

which is $\chi^2$ distributed with one degree of freedom, where $\hat{\mathbf{v}} = \mathbf{y} - X\boldsymbol{\beta}$.

For error component spatial autoregressive model (9.2.19), Anselin (1988) derived the Lagrangian multiplier (LM) test statistic for $H_0 : \theta = 0$,

$$\tau^* = \frac{\left[\frac{1}{\sigma_u^2}\hat{\mathbf{v}}^{*'}(W \otimes I_T + \hat{k}(T\hat{k} - 2)\mathbf{e}_T\mathbf{e}_T')\hat{\mathbf{v}}^*\right]}{P}, \tag{9.5.3}$$

which is asymptotically $\chi^2$ distributed with one degree of freedom, where $\mathbf{v}^* = \mathbf{y} - X\tilde{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^N X_i' V^{*-1} X_i\right)^{-1}\left(\sum_{i=1}^N X_i' V^{*-1}\mathbf{y}_i\right)$, the usual error component estimator, $\hat{k} = \hat{\sigma}_\alpha^2\left[\hat{\sigma}_u^2\left(1 + T\frac{\hat{\sigma}_\alpha^2}{\hat{\sigma}_u^2}\right)\right]^{-1}$, and $P = (T^2\hat{k}^2 - 2\hat{k} + T)(tr W^2 + tr W'W)$. Baltagi et al. (2007) consider various combination of error components and the spatial parameter test. Kelejian and Prucha (2001), and Pinkse (2000) have suggested tests of cross-sectional dependence based on the spatial correlation analogue of the Durbin–Watson/Box-Pierce tests for time series correlations.

Breusch and Pagan (1980) derived an LM test statistic for cross-sectional dependence:

$$LM = T\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\hat{\rho}_{ij}^2, \tag{9.5.4}$$

where $\hat{\rho}_{ij}$ is the estimated sample cross-correlation coefficient between the least-squares residuals $\hat{v}_{it}$ and $\hat{v}_{jt}$, where $\hat{v}_{it} = y_{it} - \mathbf{x}_{it}'\hat{\boldsymbol{\beta}}_i$, and $\hat{\boldsymbol{\beta}}_i = (X_i'X_i)^{-1}X_i\mathbf{y}_i$. When $N$ is fixed and $T \to \infty$, (9.5.4) converges to a $\chi^2$ distribution with $\frac{N(N-1)}{2}$ degrees of freedom under the null of no cross-sectional dependence. When $N$ is large, the scaled Lagrangian multiplier statistic (SLM),

$$\text{SLM} = \sqrt{\frac{2}{N(N-1)}}LM \tag{9.5.5}$$

is asymptotically normally distributed with mean 0 and variance 1.

Many panel data sets have $N$ much larger than $T$. Because $E(T\hat{\rho}_{ij}^2) \neq 0$ for all $T$, SLM is not properly centered. In other words, when $N > T$, the SLM tends to overeject, often substantially.

To correct for the bias in large $N$ and finite $T$ panels, Pesaran et al. (2008) propose a bias-adjusted LM test,

$$\text{LM}_B = \sqrt{\frac{2}{N(N-1)}}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\frac{(T-k)\hat{\rho}_{ij}^2 - \mu_{ij}}{w_{ij}}, \tag{9.5.6}$$

where $\mu_{ij} = E[(T-k)\hat{\rho}_{ij}^2]$, $w_{ij}^2 = \text{Var}\,[(T-k)\hat{\rho}_{ij}^2]$, and $k$ is the dimension of $\mathbf{x}_{it}$. They show that (9.5.6) is asymptotically normally distributed with mean 0 and variance 1 for all $T > k + 8$. The exact expressions for $\mu_{ij}$ and $w_{ij}^2$ when $\mathbf{x}_{it}$ is strictly exogenous and $v_{it}$ are normally distributed are given by Pesaran et al. (2008).

Because the adjustment of the finite sample bias of the LM test is complicated, Pesaran (2004) suggests a CD test statistic for cross-sectional dependence:

$$\text{CD} = \sqrt{\frac{2T}{N(N-1)}}\left(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\hat{\rho}_{ij}\right). \tag{9.5.7}$$

When both $N$ and $T \to \infty$, the CD test converges to a normal distribution with mean 0 and variance 1 under the null of cross-sectional independence conditional on $\mathbf{x}$. The CD test can also be applied to the linear dynamic model:

$$y_{it} = \gamma y_{i,t-1} + \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + u_{it}. \tag{9.5.8}$$

The Monte Carlo simulations conducted in Pesaran (2004) shows that the estimated size is very close to the nominal level for any combinations of $N$ and $T$ considered. However, the CD test has power only if $\frac{1}{N}\sum_{i=1}^{N}\rho_{ij} \neq 0$. On the other hand, the LM test has power even if the average of the correlation coefficient is equal to 0 as long as some pairs, $\hat{\rho}_{ij} \neq 0$.

As an alternative, Sarafidis, Yamagata, and Robertson (SYR) (2009) proposed a Sargan's (1958) difference test based on the GMM estimator of (9.5.8). As shown in Chapter 4, $\boldsymbol{\theta}' = (\gamma, \boldsymbol{\beta}')$ can be estimated by the GMM method (4.3.47). SYR suggest to split $W_i$ into two separate sets of instruments,

$$W_{1i}' = \begin{bmatrix} y_{i0} & 0 & 0 & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & y_{i0} & y_{i1} & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & y_{i0} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & y_{i0} & \cdot & \cdot & y_{i,T-2} \end{bmatrix}, \tag{9.5.9}$$

and

$$W_{2i}' = \begin{bmatrix} \mathbf{x}_i' & \mathbf{0}' & \mathbf{0}' & \cdot & \cdot \\ \mathbf{0}' & \mathbf{x}_i' & \mathbf{0}' & \cdot\cdot & \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \mathbf{x}_i' \end{bmatrix}, \tag{9.5.10}$$

where $\mathbf{x}_i' = (\mathbf{x}_{i1}', \ldots, \mathbf{x}_{iT}')$, $W_{1i}'$ is $(T-1) \times T(T-1)/2$, $W_{2i}'$ is $(T-1) \times KT(T-1)$ and $\mathbf{x}_{it}$ is strictly exogenous.[7]

Under the null of no cross-sectional dependence, both sets of moment conditions

$$E[\mathbf{W}_{1i}\Delta\mathbf{u}_i] = \mathbf{0}, \tag{9.5.11}$$

[7] If $\mathbf{x}_{it}$ is predetermined rather than strictly exogenous, a corresponding $W_2$ can be constructed as

$$W_2 = \begin{bmatrix} \mathbf{x}_1', & \mathbf{0}' & \mathbf{0}' & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \mathbf{x}_1' & \mathbf{x}_2' & \mathbf{0}' & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \mathbf{x}_1' & \cdot & \mathbf{x}_{T-1}' \end{bmatrix}$$

and

$$E[\mathbf{W}_{2i}\Delta\mathbf{u}_i] = \mathbf{0}, \tag{9.5.12}$$

hold. However, if there exists cross-sectional dependence, (9.5.11) may not hold. For instance, suppose $u_{it}$ can be decomposed into the sum of two components, the impact of $r$ time-varying common omitted factors and an idiosyncratic component, $\epsilon_{it}$,

$$u_{it} = \mathbf{b}_i' \mathbf{f}_t + \epsilon_{it}. \tag{9.5.13}$$

For simplicity, we assume $\epsilon_{it}$ is independently distributed over $i$ and $t$. Then the first difference of $u_{it}$,

$$\Delta u_{it} = \mathbf{b}_i'\Delta \mathbf{f}_t + \Delta\epsilon_{it}, \tag{9.5.14}$$

and

$$\begin{aligned} y_{it} = \frac{1-\gamma^t}{1-\gamma}\alpha_i + \gamma^t y_{i0} + \sum_{j=0}^{t-1}\gamma^j \mathbf{x}_{i,t-j}'\boldsymbol{\beta} \\ + \mathbf{b}_i' \sum_{j=0}^{t-1}\gamma^j \mathbf{f}_{t-j} + \sum_{j=0}^{t-1}\gamma^j \epsilon_{i,t-j}. \end{aligned} \tag{9.5.15}$$

Under the assumption that $\mathbf{f}_t$ are nonstochastic and bounded but $\mathbf{b}_i$ are random with mean $\mathbf{0}$ and covariance $E\mathbf{b}_i\mathbf{b}_i' = \sum_b$, $E(y_{i,t-j}\Delta u_{it})$ is not equal to 0, for $j = 2, \ldots, t$. Therefore, SYR suggest estimating $\gamma$ and $\boldsymbol{\beta}$ by (4.3.45) first using both (9.5.11) and (9.5.12) moment conditions, denoted by $(\hat{\gamma}, \hat{\boldsymbol{\beta}}')$, construct estimated residuals $\Delta\mathbf{u}_i$ by $\Delta\hat{u}_i = \Delta\mathbf{y}_i - \Delta y_{i,-1}\hat{\gamma} - \Delta X_i\hat{\boldsymbol{\beta}}$, where $\Delta\mathbf{y}_i = (\Delta y_{i2}, \ldots, \Delta y_{iT})'$, $\Delta\mathbf{y}_{i,-1} = (\Delta y_{i1}, \ldots, \Delta y_{i,T-1})'$ and $\Delta X_i = (\Delta\mathbf{x}_{i1}, \ldots, \Delta\mathbf{x}_{iT})'$. Then estimate $(\gamma, \boldsymbol{\beta}')$ using moment conditions (9.5.12) only,

$$\begin{aligned} \begin{pmatrix} \tilde{\gamma} \\ \tilde{\boldsymbol{\beta}} \end{pmatrix} = \left\{ \left[ \sum_{i=1}^N \begin{pmatrix} \Delta\mathbf{y}_{i,-1}' \\ \Delta X_i' \end{pmatrix} W_{2i} \right] \hat{\Omega}^{-1} \left[ \sum_{i=1}^N W_{2i}'(\Delta\mathbf{y}_{i,-1}, \Delta X_i) \right] \right\}^{-1} \\ \cdot \left\{ \left[ \sum_{i=1}^N \begin{pmatrix} \Delta\mathbf{y}_{i,-1}' \\ \Delta X_i' \end{pmatrix} W_{2i} \right] \hat{\Omega}^{-1} \left[ \sum_{i=1}^N W_{2i}'\Delta\mathbf{y}_i \right] \right\}, \end{aligned} \tag{9.5.16}$$

where $\hat{\Omega}^{-1} = N^{-1}\sum_{i=1}^N W_{2i}'\Delta\hat{\mathbf{u}}_i\Delta\hat{\mathbf{u}}_i'W_{2i}$. Under the null of cross-sectional independence both estimators are consistent. Under the alternative, $(\hat{\gamma}, \hat{\boldsymbol{\beta}}')$ may not be consistent but (9.5.16) remains consistent. Therefore, SYR, following

the idea of Sargan (1958) and Hansen (1982), suggest using the test statistic

$$N^{-1}\left(\sum_{i=1}^{N}\Delta\hat{\mathbf{u}}_i'W_i\right)\hat{\Psi}^{-1}\left(\sum_{i=1}^{N}W_i'\Delta\hat{\mathbf{u}}_i\right) - N^{-1}\left(\sum_{i=1}^{N}\Delta\tilde{\mathbf{u}}_i'W_{2i}\right)\tilde{\Psi}^{-1}\left(\sum_{i=1}^{N}W_{2i}'\Delta\tilde{\mathbf{u}}_i\right) \tag{9.5.17}$$

where $\Delta\tilde{\mathbf{u}}_i = \Delta\mathbf{y}_i - \Delta\mathbf{y}_{i,-1}\tilde{\gamma} - \Delta X_i\tilde{\boldsymbol{\beta}}$, $\hat{\Psi} = \frac{1}{N}\sum_{i=1}^{N}W_i'\Delta\hat{\mathbf{u}}_i\Delta\hat{\mathbf{u}}_i'W_i$ and $\tilde{\Psi} = \frac{1}{N}\sum_{i=1}^{N}W_{2i}'\Delta\tilde{\mathbf{u}}_i\Delta\tilde{\mathbf{u}}_i'W_{2i}$. SYR show that under the null of cross-sectional independence, (9.5.17) converges to a $\chi^2$ distribution with $\frac{T(T-1)}{2}(1+K)$ degrees of freedom as $N\to\infty$.

The advantage of the SYR test is that the test statistic (9.5.17) has power even if $\sum_{j=1}^{N}\rho_{ij}=0$. Monte Carlo studies conducted by SYR show that the test statistic (9.5.17) performs well if the cross-sectional dependence is driven by nonstochastic $\mathbf{f}_t$ but stochastic $\mathbf{b}_i$. However, if the cross-sectional dependence is driven by fixed $\mathbf{b}_i$ and stochastic $\mathbf{f}_t$, then the test statistic is unlikely to have power because $E(\Delta y_{i,t-j}\Delta u_i)=0$ if $\mathbf{f}_t$ is independently distributed over time.[8]

### 9.5.2 Limited Dependent-Variable Model

Many limited dependent-variable models take the form of relating observed $y_{it}$ to a latent $y_{it}^*$, (e.g., Chapters 7 and 8),

$$y_{it}^* = \mathbf{x}_{it}'\boldsymbol{\beta} + v_{it}, \tag{9.5.18}$$

through a link function $g(\cdot)$

$$y_{it} = g(y_{it}^*). \tag{9.5.19}$$

For example, in the binary choice model,

$$g(y_{it}^*) = I(y_{it}^* > 0), \tag{9.5.20}$$

and in the Tobit model,

$$g(y_{it}^*) = y_{it}^* I(y_{it}^* > 0), \tag{9.5.21}$$

where I(A) is an indicator function that takes the value 1 if $A$ occurs and 0 otherwise.

There is a fundamental difference between the linear model and limited dependent-variable model. There is a one-to-one correspondence between $v_{it}$

[8] $E(\Delta y_{i,t-j}\Delta u_{it})$ is not equal to 0 if $\mathbf{f}_t$ is serially correlated. However, if $\mathbf{f}_t$ is serially correlated, then $u_{it}$ is serially correlated and $y_{i,t-j}$ is not a legitimate instrument if the order of serially correlation is greater than $j$. Laggerd $y$ can be legitimate instruments only if $E(\Delta u_{it}y_{i,t-s})=0$. Then the GMM estimator of (4.3.47) will have to be modified accordingly.

and $y_{it}$ in the linear model, but not in limited dependent variable model. The likelihood for observing $\mathbf{y}_t = (y_{it}, \ldots, y_{Nt})'$,

$$P_t = \int_{A(\mathbf{v}_t|\mathbf{y}_t)} f(\mathbf{v}_t) d\mathbf{v}_t, \tag{9.5.22}$$

where $A(\mathbf{v}_t \mid \mathbf{y}_t)$ denotes the region of integration of $\mathbf{v}_t = (v_{it}, \ldots, v_{Nt})'$ which is determined by the realized $\mathbf{y}_t$ and the form of the link function. For instance, in the case of probit model, $A(\mathbf{v}_t \mid \mathbf{y}_t)$ denotes the region $(a_{it} < v_{it} < b_{it})$, where $a_{it} = -\mathbf{x}_{it}'\boldsymbol{\beta}$, $b_{it} = \infty$ if $y_{it} = 1$ and $a_{it} = -\infty$, $b_{it} = -\mathbf{x}_{it}'\boldsymbol{\beta}$ if $y_{it} = 0$.

Under the assumption that $v_{it}$ is independently normally distributed across $i$, Hsiao, Pesaran, and Picks (2012) show that the Lagrangian multiplier test statistic of cross-sectional independence takes an analogous form:

$$LM = T \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \tilde{\rho}_{ij}^2, \tag{9.5.23}$$

where

$$\tilde{\rho}_{ij} = \frac{T^{-1} \sum_{t=1}^{T} \tilde{v}_{it} \tilde{v}_{jt}}{\sqrt{T^{-1} \sum_{t=1}^{T} \tilde{v}_{it}^2} \sqrt{T^{-1} \sum_{t=1}^{T} \tilde{v}_{jt}^2}}, \tag{9.5.24}$$

and $\tilde{v}_{it} = E(v_{it} \mid y_{it})$, the conditional mean of $v_{it}$ given $y_{it}$. For instance, in the case of probit model,

$$\tilde{v}_{it} = \frac{\phi(\mathbf{x}_{it}'\boldsymbol{\beta})}{\Phi(\mathbf{x}_{it}'\boldsymbol{\beta})[1 - \Phi(\mathbf{x}_{it}'\boldsymbol{\beta})]} [y_{it} - \Phi(\mathbf{x}_{it}'\boldsymbol{\beta})]. \tag{9.5.25}$$

In the case of the Tobit model

$$\tilde{v}_{it} = (y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta}) I(y_{it} > 0) - \sigma_i \frac{\phi(\frac{\mathbf{x}_{it}'\boldsymbol{\beta}}{\sigma_i})}{\Phi(-\frac{\mathbf{x}_{it}'\boldsymbol{\beta}}{\sigma_i})} [1 - I(y_{it} > 0)], \tag{9.5.26}$$

where $\sigma_i^2 = \text{Var}\ (v_{it})$, $\phi(\cdot)$ and $\Phi(\cdot)$ denote standard normal and integrated standard normal. Under the null of cross-sectional independence, (9.5.23) converges to a $\chi^2$ distribution with $\frac{N(N-1)}{2}$ degrees of freedom if $N$ is fixed and $T \to \infty$. When $N$ is also large

$$\sqrt{\frac{2}{N(N-1)}} LM \tag{9.5.27}$$

is asymptotically standard normally distributed.

When $N$ is large and $T$ is finite, the LM test statistically is not centered properly. However, for the nonlinear model, the bias correction factor is not easily derivable. Hsiao et al. (2012) suggest constructing Pesaran (2006) CD statistic using $\tilde{v}_{it}$.

Sometimes, the deviation of $\tilde{v}_{it}$ is not straightforward for a nonlinear model. Hsiao et al. (2012) suggest replacing $\tilde{v}_{it}$ by

$$v_{it}^{*} = y_{it} - E(y_{it} \mid \mathbf{x}_{it}) \tag{9.5.28}$$

in the construction of an LM or CD test statistic. Monte Carlo experiments conducted by Hsiao et al. (2012) show that there is very little difference between the two procedures to construct CD tests.

### 9.5.3 An Example – A Housing Price Model of China

Mao and Shen (2013) consider China's housing price model using 30 provincial-level quarterly data from the second quarter of 2001 to the fourth quarter of 2012 of the logarithm of seasonally adjusted real house price, $y_{it}$, as a linear function of the logarithm of seasonally adjusted real per capita wage income ($x_{1it}$); the logarithm of real long-term interest rate ($x_{2it}$); and the logarithm of the urban population ($x_{3it}$). Table 9.1 provides Mao and Shen (2013) estimates of the mean group estimator $\hat{\boldsymbol{\beta}} = \frac{1}{N}\sum_{i=1}^{N} \hat{\boldsymbol{\beta}}_i$ for the cross sectionally independent heterogeneous model (MG),

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta}_i + v_{it}; \tag{9.5.29}$$

the Pesaran (2006) common correlated effects heterogeneous model (CCEMG),

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta}_i + \bar{y}_t c_i + \bar{\mathbf{x}}_t'\mathbf{d}_i + v_{it}; \tag{9.5.30}$$

and the homogeneous common correlated effects model (CCEP),

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \bar{y}_t \tau_i + \bar{\mathbf{x}}_t'\mathbf{d}_i + v_{it}. \tag{9.5.31}$$

It can be seen from the results in Table 9.1 that (1) the estimated slope coefficients, $\boldsymbol{\beta}$, are very sensitive to the adjustment (CCEMG or CCEP) or nonadjustment of cross-sectional dependence, and (2) the suggested approach to control the impact of cross-sectional dependence works only if the observed data satisfy the assumptions underlying the approach. (See Remark 9.4.1 for the limitation of augmenting regression models by the cross-sectional mean.) As one can see from Table 9.1, the Pesaran (2004) CD tests (9.5.7) of the residuals of the (9.5.30) and (9.5.31) indicate that significant cross-sectional dependence remains. It is only by further adjusting the common correlated effects model residuals by a spatial model with the spatial weight matrix specified in terms of the geometric distance between region $i$ and $j$ that Mao and Shen (2013) can achieve cross-sectional independence to their model.

Table 9.1. *Common correlated effects estimation*

| | MG | | | CCEMG | | | CCEP | | |
|---|---|---|---|---|---|---|---|---|---|
| x1 | 1.088‡ | 1.089‡ | 0.979‡ | 0.264 | 0.313† | 0.308$^{+}$ | 0.388† | 0.467‡ | 0.449‡ |
| | (0.058) | (0.056) | (0.114) | (0.176) | (0.173) | (0.170) | (0.169) | (0.165) | (0.170) |
| x2 | – | −0.003 | −0.052 | – | −6.453 | 4.399 | – | −4.796 | 4.387 |
| | – | (0.058) | (0.057) | – | (2.927) | (2.839) | – | (3.943) | (3.401) |
| x3 | – | – | 0.718 | – | – | −0.098 | – | – | 0.104 |
| | – | – | (0.484) | – | – | (0.552) | – | – | (0.130) |
| *CD* | 28.15‡ | 30.39‡ | 27.64‡ | −4.257‡ | −.4173‡ | −4.073‡ | −4.521‡ | −4.494‡ | −4.518‡ |

Symbols $^{+}$, †, and ‡ denote that the corresponding stastics are significant at 10%, 5%, and 1% level respectively. The values in parentheses are corresponding standard errors.
*Source:* Mao and Shen (2013, Table V).

## 9.6 A PANEL DATA APPROACH FOR PROGRAM EVALUATION

### 9.6.1 Introduction

Individuals are often given "treatments," such as a drug trial, a training program, and so forth. If it is possible to simultaneously observe the same person in the treated and untreated states, then it is fairly straightforward to isolate the treatment effects in question. When it is not possible to simultaneously observe the same person in the treated and untreated states or the assignment of individuals to treatment is nonrandom, treatment effects could confound with the factors that would make people different on outcome measures or with the sample selection effects or both.

In this section we first review some basic approaches for measuring treatment effects with cross-sectional data, and then we show how the availability of panel data can substantially simplify the inferential procedure.

### 9.6.2 Definition of Treatment Effects

For individual $i$, let $(y_i^{0*}, y_i^{1*})$ be the potential outcomes in the untreated and treated state. Suppose the outcomes can be decomposed as the sum of the effects of observables, $\mathbf{x}$, $m_j(\mathbf{x})$, and unobservables, $\epsilon_j$; $j = 0, 1$, in the form,

$$y_i^{0*} = m_0(\mathbf{x}_i) + \epsilon_i^0, \tag{9.6.1}$$

$$y_i^{1*} = m_1(\mathbf{x}_i) + \epsilon_i^1, \tag{9.6.2}$$

where $\epsilon_i^0$ and $\epsilon_i^1$ are the 0 mean unobserved random variables, assumed to be independent of $\mathbf{x}_i$. The treatment effect for individual $i$ is defined as

$$\Delta_i = y_i^{1*} - y_i^{0*}. \tag{9.6.3}$$

The average treatment effect (ATE) (or the mean impact of treatment if people were randomly assigned to the treatment)[9] is defined as

$$\begin{aligned} \Delta^{\text{ATE}} &= E[y_i^{1*} - y_i^{0*}] = E\left\{[m_1(\mathbf{x}) - m_0(\mathbf{x})] + (\epsilon^1 - \epsilon^0)\right\} \\ &= E[m_1(\mathbf{x}) - m_0(\mathbf{x})]. \end{aligned} \tag{9.6.4}$$

Let $d_i$ be the dummy variable indicating an individual's treatment status with $d_i = 1$ if the $i$th individual receives the treatment and 0 otherwise. The effect of treatment on the treated (TT) (or the mean impact of treatment of those who received treatment compared to what they would have been in the absence of

[9] See Heckman (1997), Heckman and Vytacil (2001), and Imbens and Angrist (1994) for the definitions of the marginal treat effect (MTE) and the local average treatment effect (LATE).

treatment) is defined as

$$\begin{aligned}\Delta^{\mathrm{TT}} &= E(y^{1*} - y^{0*} \mid d = 1) \\ &= E[m_1(\mathbf{x}) - m_0(\mathbf{x}) \mid d = 1] + E[\epsilon^1 - \epsilon^0 \mid d = 1].\end{aligned} \tag{9.6.5}$$

Similarly, we can define the effect of treatment on untreated group as

$$\begin{aligned}\Delta^{\mathrm{TUT}} &= E(y^{1*} - y^{0*} \mid d = 0) \\ &= E[m_1(\mathbf{x}) - m_0(\mathbf{x}) \mid d = 0] + E[\epsilon^1 - \epsilon^0 \mid d = 0].\end{aligned} \tag{9.6.6}$$

The ATE is of interest if one is interested in the effect of treatment for a randomly assigned individual or population mean response to treatment. The TT is of interest if the same selection rule for treatment continues in the future. The relation between ATE and TT is given by

$$\Delta^{\mathrm{ATE}} = \text{Prob}\,(d = 1)\Delta^{\mathrm{TT}} + \text{Prob}\,(d = 0)\Delta^{\mathrm{TUT}}. \tag{9.6.7}$$

If $E[m_1(\mathbf{x}) - m_0(\mathbf{x}) \mid d = 1] = E[m_1(\mathbf{x}) - m_0(\mathbf{x}) \mid d = 0] = E[m_1(\mathbf{x}) - m_0(\mathbf{x})]$ and $E[\epsilon^1 - \epsilon^0 \mid d = 1] = E[\epsilon^1 - \epsilon^0 \mid d = 0] = E[\epsilon^1 - \epsilon^0]$, then $\Delta^{\mathrm{ATE}} = \Delta^{\mathrm{TT}} = \Delta^{\mathrm{TUT}}$.

If we simultaneously observe $y_i^{0*}$ and $y_i^{1*}$ for a given $i$, then ATE and TT can be easily measured. However, for a given $i$, the observed outcome, $y_i$ is either $y_i^{0*}$ or $y_i^{1*}$, not both,

$$y_i = d_i y_i^{1*} + (1 - d_i) y_i^{0*}. \tag{9.6.8}$$

If we measure the treatment effect by comparing the mean difference between those receiving the treatment (the treatment group) and those not receiving the treatment (control group), $\frac{1}{n_d}\sum_{i\in\psi} y_i$ and $\frac{1}{N-n_d}\sum_{i\in\bar{\psi}} y_i$, where $\psi = \{i \mid d_i = 1\}$, $\bar{\psi}_i = \{i \mid d_i = 0\}$ and $n_d = \sum_{i=1}^{N} d_i$,

$$\begin{aligned}&\frac{1}{n_d}\sum_{i\in\psi} y_i - \frac{1}{N-n_d}\sum_{i\in\bar{\psi}} y_i \longrightarrow E(y \mid d = 1) - E(y \mid d = 0) \\ &= \{E[m_1(x) - m_0(x) \mid d = 1]\} + \{E[m_0(x) \mid d = 1] - \\ &\qquad E[m_0(x) \mid d = 0]\} + \{E(\epsilon^1 \mid d = 1) - E(\epsilon^0 \mid d = 0)\}.\end{aligned} \tag{9.6.9}$$

The average difference between the treatment group and control group ((9.6.9)) is the sum of three components, the treatment effect of the treated, $\Delta^{\mathrm{TT}}$, $E[m_1(x) - m_0(x) \mid d = 1]$, the effects of confounding variables being different between the treatment group and control group, $E[m_0(x) \mid d = 1] - E[m_0(x) \mid d = 0]$, and the participation (or selection effects, $E(\epsilon^1 \mid d = 1) - E(\epsilon^0 \mid d = 0)$). If participation of treatment is random, then $E(\epsilon^1 \mid d = 1) = E(\epsilon^1) = E(\epsilon^0 \mid d = 0) = E(\epsilon^0) = 0$. If $f(x \mid d = 1) = f(x \mid d = 0) = f(x)$, then $E[m_1(x) - m_0(x) \mid d = 1] = E[m_1(x) - m_0(x)]$

and $E[m_0(x) \mid d = 1] - E[m_0(x) \mid d = 0] = 0$, (9.6.9) provides an unbiased measure of $\Delta^{ATE}(\equiv \Delta^{TT})$.[10]

In an observational study, the treatment group and control group are often drawn from different populations (e.g., LaLonde 1986; Dehejia and Wahba 1999). For instance, the treatment group can be drawn from welfare recipients eligible for a program of interest while the control group can be drawn from a different population. If there are systematic differences between the treatment group and comparison group in observed and unobserved characteristics that affect outcomes, estimates of treatment effects based on the comparison of the difference between $\frac{1}{n_d}\sum_{i\epsilon\psi} y_i - \frac{1}{N-n_d}\sum_{i\epsilon\bar{\psi}} y_i$ are distorted. The distortion can come from either one or both of the following two sources:

(1) Selection bias due to observables, $E\{m_0(\mathbf{x}) \mid d = 1\} \neq E\{m_0(\mathbf{x}) \mid d = 0\}$ (or $E\{m_1(\mathbf{x}) \mid d = 1\} \neq E\{m_1(\mathbf{x}) \mid d = 0\}$), that is, bias due to differences in observed (conditional) variables between the two groups.

(2) Selection bias due to unobservables, that is, bias due to differences in unobserved characteristics between the two groups, $E(\epsilon^0 \mid d = 1) \neq E(\epsilon^0 \mid d = 0)$, (and $E(\epsilon^1 \mid d = 1) \neq E(\epsilon^1 \mid d = 0)$).

A variety of matching and statistical-adjustment procedures have been proposed to take account of discrepancies in observed and unobserved characteristics between treatment and control group members (e.g., Heckman and Robb 1985; Heckman, Ichimura, and Todd 1998; LaLonde 1986; Rosenbaum and Rubin 1983). We shall first review methods for the analysis of cross-sectional data, and then discuss the panel data approach.

### 9.6.3 Cross-Sectional Adjustment Methods

#### *9.6.3.1 Parametric Approach*

Suppose $y_i^{1*}$ and $y_i^{0*}$ ((9.6.1) and (9.6.2)) can be specified parametrically. In addition, if the participation of treatment is assumed to be a function of

$$d_i^* = h(\mathbf{z}_i) + v_i, \tag{9.6.10}$$

where

$$d_i = \begin{cases} 1, & \text{if } d_i^* > 0, \\ 0, & \text{if } d_i^* \leq 0, \end{cases} \tag{9.6.11}$$

and $\mathbf{z}$ denote the factors determining the selection equation that may overlap with some or all elements of $\mathbf{x}$. With a known joint distribution of $f(\epsilon^1, \epsilon^0, v)$, the mean response functions $m_1(\mathbf{x})$, $m_0(\mathbf{x})$ can be consistently estimated by the maximum-likelihood method and

$$\hat{ATE}(\mathbf{x}) = \hat{m}_1(\mathbf{x}) - \hat{m}_0(\mathbf{x}) \tag{9.6.12}$$

(e.g., Damrongplasit, Hsiao, and Zhao 2010).

[10] Similarly, we can write $E(y \mid d = 1) - E(y \mid d = 0) = \Delta^{\text{TUT}} + \{E(m_1(\mathbf{x}) \mid d = 1) - E(m_1(\mathbf{x}) \mid d = 0\} + \{E(\epsilon^1 \mid d = 1) - E(\epsilon^1 \mid d = 0)\}$.

*9.6.3.2 Nonparametric Approach*

If $m_1(\mathbf{x})$ and $m_0(\mathbf{x})$ are unspecified, they can be estimated by nonparametric methods provided that conditional on a set of confounding variables, say $\mathbf{x}$, the distributions of $(y^{1*}, y^{0*})$ are independent of $d$ (or $d^*$). In other words, conditional on $\mathbf{x}$, there is no selection on unobservables (conditional independence),

$$E(y^{1*} \mid d, \mathbf{x}) = E(y^{1*} \mid \mathbf{x}), \tag{9.6.13}$$

$$E(y^{0*} \mid d, \mathbf{x}) = E(y^{0*} \mid \mathbf{x}). \tag{9.6.14}$$

Then conditional on $\mathbf{x}$, the average treat effect, ATE($\mathbf{x}$),

$$\begin{aligned} \text{ATE}(\mathbf{x}) &= E(y^{1*} - y^{0*} \mid \mathbf{x}) \\ &= E(y \mid d = 1, \mathbf{x}) - E(y \mid d = 0, \mathbf{x}) \\ &= E(y^{1*} \mid \mathbf{x}) - E(y^{0*} \mid \mathbf{x}) \end{aligned} \tag{9.6.15}$$

*9.6.3.2 (i) Matching Observables in Terms of Propensity Score Method (or Selection on Observables Adjustment)*

However, if the dimension of $\mathbf{x}$ is large, the nonparametric method may suffer from "the curse of dimensionality." As a dimension reduction method, Rosenbaum and Rubin (1983) have suggested a propensity score method to match the observable characteristics of the treatment group and the control group. The Rosenbaum and Rubin (1983) propensity score methodology supposes unit $i$ has observable characteristics $\mathbf{x}_i$. Let $P(\mathbf{x}_i)$ be the probability of unit $i$ having been assigned to treatment, called the propensity score in statistics and choice probability in econometrics, defined as $P(\mathbf{x}_i) = \text{Prob}\ (d_i = 1 \mid \mathbf{x}_i) = E(d_i \mid \mathbf{x}_i)$. Assume that $0 < P(\mathbf{x}_i) < 1$ for all $\mathbf{x}_i$,[11] and Prob $(d_1, \ldots, d_N \mid \mathbf{x}_1, \ldots, \mathbf{x}_N) = \prod_{i=1}^{N} P(\mathbf{x}_i)^{d_i}[1 - P(\mathbf{x}_i)]^{1-d_i}$, for $i = 1, \ldots, N$. If the treatment assignment is ignorable given $\mathbf{x}$, then it is ignorable given $P(\mathbf{x})$; that is,

$$\{(y_i^{1*}, y_i^{0*}) \perp d_i \mid \mathbf{x}_i\} \Longrightarrow \{(y_i^{1*}, y_i^{0*}) \perp d_i \mid P(\mathbf{x}_i)\}, \tag{9.6.16}$$

where $\perp$ denotes orthogonality.

To show (9.6.16) holds, it is sufficient to show that

$$\begin{aligned} &\text{Prob}\ (d = 1 \mid y^{0*}, y^{1*}, P(\mathbf{x})) \\ &\quad = \text{Prob}\ (d = 1 \mid P(\mathbf{x})) \\ &\quad = P(\mathbf{x}) = \text{Prob}\ (d = 1 \mid \mathbf{x}) = \text{Prob}\ (d = 1 \mid y^{0*}, y^{1*}, \mathbf{x}) \end{aligned} \tag{9.6.17}$$

[11] The assumption that $0 < P(\mathbf{x}_i) < 1$ guarantees that for each $\mathbf{x}_i$, we obtain observations in both the treated and untreated states. This assumption can be relaxed as long as there are $\mathbf{x}$ such that $0 < P(\mathbf{x}) < 1$.

Eq. (9.6.17) follows from applying the ignorable treatment assignment assumption to

$$\begin{aligned}
&\text{Prob}\,(d = 1 \mid y^{0*}, y^{1*}, P(\mathbf{x})) \\
&\quad = E_x\left\{\text{Prob}\,(d = 1 \mid y_0^*, y_1^*, \mathbf{x}) \mid y^{0*}, y^{1*}, P(\mathbf{x})\right\} \\
&\quad = E_x\left\{\text{Prob}\,(d = 1 \mid \mathbf{x}) \mid y^{0*}, y^{1*}, P(\mathbf{x})\right\} \qquad (9.6.18) \\
&\quad = E_x\left\{P(\mathbf{x}) \mid y^{0*}, y^{1*}, P(\mathbf{x})\right\} \\
&\quad = E_x\{P(\mathbf{x}) \mid P(\mathbf{x})\} = P(\mathbf{x}),
\end{aligned}$$

where $E_x$ denotes taking the expectation with respect to $\mathbf{x}$.

It follows from (9.6.16) that

$$\mathbf{x}_i \perp d_i \mid P(\mathbf{x}_i). \qquad (9.6.19)$$

To prove (9.6.19), it is sufficient to show that

$$\text{Prob}\,(d = 1 \mid \mathbf{x}, P(\mathbf{x})) = \text{Prob}\,(d = 1 \mid P(\mathbf{x})). \qquad (9.6.20)$$

Equation (9.6.20) follows from $\text{Prob}(d = 1 \mid \mathbf{x}, P(\mathbf{x})) = \text{Prob}\,(d = 1 \mid \mathbf{x}) = P(\mathbf{x})$ and

$$\begin{aligned}
\text{Prob}\,(d = 1 \mid P(\mathbf{x})) &= E_x\{\text{Prob}\,(d = 1 \mid \mathbf{x}, P(\mathbf{x})) \mid P(\mathbf{x})\} \\
&= E_x\{P(\mathbf{x}) \mid P(\mathbf{x})\} = P(x).
\end{aligned}$$

Equation (9.6.19) implies that the conditional density of $\mathbf{x}$ given $d$ and $P(\mathbf{x})$,

$$f(\mathbf{x} \mid d = 1, P(\mathbf{x})) = f(\mathbf{x} \mid d = 0, P(\mathbf{x})) = f(\mathbf{x} \mid P(\mathbf{x})). \qquad (9.6.21)$$

In other words, Equation (9.6.19) implies that if a subclass of units or a matched treatment–control pair is homogeneous in $P(\mathbf{x})$, then the treated and control units in that subclass or matched pair will have the same distribution of $\mathbf{x}$. In other words, at any value of a propensity score, the mean difference between the treatment group and control group is an unbiased estimate of the average treatment effect at that value of the propensity score if treatment assignment is ignorable.

$$\Delta(P(\mathbf{x}))^{\text{TT}} = E\{E(y \mid d = 1, P(\mathbf{x})) - E(y \mid d = 0, P(\mathbf{x})) \mid d = 1\}, \qquad (9.6.22)$$

where the outer expectation is over the distribution of $\{P(\mathbf{x}) \mid d = 1\}$.

The attraction of propensity score matching method is that in (9.6.15) we condition on $\mathbf{x}$ (intuitively, to find observations with similar covariates), while in (9.6.22) we condition just on the propensity score because (9.6.22) implies that observations with the same propensity score have the same distribution of the full vector of covariates, $\mathbf{x}$. Equation (9.6.19) asserts that conditional on $P(\mathbf{x})$, the distribution of covariates should be the same across the treatment and comparison groups. In other words, conditional on the propensity score,

each individual has the same probability of assignment to treatment as in a randomized experiment. Therefore, the estimation of average treatment effect for the treated[12] can be done in two steps. The first step involves the estimation of propensity score parametrically or nonparametrically (e.g., see Chapter 7). In the second step, given the estimated propensity score, one can estimate $E\{y \mid P(\mathbf{x}), d = j\}$ for $j = 0, 1$, take the difference between the treatment and control groups, then weight these by the frequency of treated observations or frequency of (both treated and untreated) observations in each stratum to get an estimate of TT or ATE ($E\{E[y \mid d = 1, P(\mathbf{x})] - E[y \mid d = 0, P(\mathbf{x})]\} = E\{E[y_1 - y_0 \mid P(\mathbf{x})]\}$), where the outer expectation is with respect to the propensity score, $P(\mathbf{x})$. For examples of using this methodology to evaluate the effects of training programs in nonexperimental studies, see Dehejia and Wahba (1999), and LaLonde (1986), Liu, Hsiao, Matsumoto, and Chou (2009), etc.

#### *9.6.3.2 (ii) Regression Discontinuity Design*

Let $\mathbf{x}_i = (w_i, \mathbf{q}_i')$ be $k$ covariates, where $w_i$ is a scalar and $\mathbf{q}_i$ is a $(k-1) \times 1$ vector. Both $w_i$ and $\mathbf{q}_i$ are not affected by the treatment. The basic idea behind the regression discontinuity (RD) design is that assignment to the treatment is determined, either completely or partly, by the value of a predictor $w_i$ being on either side of a fixed threshold. This predictor, $w_i$, (together with $\mathbf{q}_i$), also affects the potential outcomes.

For notational ease, we shall assume $\mathbf{q}_i = 0$ for this subsection. In the sharp RD (SRD) designs, it is assumed that all units with the values of $w$ at least $c$ are assigned to the treatment group and participation is mandatory for these individuals, and with values of $w$ less than $c$ are assigned to the control groups and members of these group are not eligible for the treatment, then

$$\begin{aligned} \text{ATE}(c) &= \lim_{w \downarrow c} E(y \mid w) - \lim_{w \uparrow c} E(y \mid w), \\ &= E(y^1 - y^0 \mid w = c) \end{aligned} \tag{9.6.23}$$

(This approach although assumes unconfoundedness of Rosenbaum and Rubin (1983), however, it violates $0 < P(d = 1 \mid \mathbf{x}) < 1$).

This approach assumes either

(1) $E(y^0 \mid w)$ and $E(y^1 \mid w)$ are continuous in $w$ or (2) $F_{y^0|w}(y \mid w)$ and $F_{y^1|w}(y \mid w)$ are continuous in $w$ for all $y$.

In the fuzzy RD (FRD), we allow

$$\lim_{w \downarrow c} \text{Prob}(d = 1 \mid w) \neq \lim_{w \uparrow c} \text{Prob}(d = 1 \mid w), \tag{9.6.24}$$

then

$$\text{ATE}(c) = \frac{\lim_{w \downarrow c} E(y \mid w) - \lim_{w \uparrow c} E(y \mid w)}{\lim_{w \downarrow c} P(d = 1 \mid w) - \lim_{w \uparrow c} P(d = 1 \mid w)}. \tag{9.6.25}$$

[12] The measurement is of interest if future selection criteria for treatment are like past selection criteria.

Let

$$\lim_{w\downarrow c} \text{Prob}(d = 1 \mid w) - \lim_{w\uparrow c} \text{Prob}\,(d = 1 \mid w) = \bigtriangledown. \tag{9.6.26}$$

$$P = \lim_{w\uparrow c} \text{Prob}\,(d = 1 \mid w)$$

Then

$$\begin{aligned} &\lim_{w\downarrow c} E(y \mid w) - \lim_{w\uparrow c} E(y \mid w) \\ &= \left\{(P + \bigtriangledown)Ey^1 - (1 - P - \bigtriangledown)Ey^0\right\} - [PEy^1 + (1 - p)Ey^0] \\ &= \bigtriangledown E[y^1 - y^0] \end{aligned} \tag{9.6.27}$$

Both the SRD and FRD designs provide only estimates of the ATE for a subpopulation with $w_i = c$. The designs do not allow the estimation of the overall ATE.

Let $\psi = \{i \mid w_i < c\}$ and $\bar{\psi} = \{i \mid w_i \geq c\}$, then for the SRD, we may estimate the ATE($c$) by the kernel method,

$$\widehat{\text{ATE}}(c) = \frac{\sum\limits_{i\in\bar{\psi}} y_i K(\frac{w_i - c}{h}) - \sum\limits_{i\in\psi} y_i K(\frac{w_i - c}{h})}{\sum\limits_{i\in\bar{\psi}} K(\frac{w_i - c}{h}) - \sum\limits_{i\in\psi} K(\frac{w_i - c}{h}).}, \tag{9.6.28}$$

where $K(\cdot)$ is a kernel function satisfying $K(0) \neq 0$, $K(v) \to 0$ as $v \to \pm\infty$. Or use the Fan and Gijbels (1992) local linear regression approach,

$$\min_{\alpha_0,\beta_0} \sum_{i:c-h<x_i<c} (y_i - \alpha_0 - \beta_0(w_i - c))^2 \tag{9.6.29}$$

and

$$\min_{\alpha_1,\beta_1} \sum_{i:c\leq x_i<c+h} (y_i - \alpha_1 - \beta_1(w_i - c))^2 \tag{9.6.30}$$

Since $E(y^1 \mid w = c) = \hat{\alpha}_1 + \beta_1(c - c) = \hat{\alpha}_1$ and $E(y^0 \mid w = c) = \hat{\alpha}_0 + \hat{\beta}_0(c - c) = \hat{\alpha}_0$, therefore

$$\widehat{\text{ATE}}(c) = \hat{\alpha}_1 - \hat{\alpha}_0. \tag{9.6.31}$$

For FRD,

$$\widehat{\text{ATE}}(c) = \frac{\hat{\alpha}_1 - \hat{\alpha}_0}{\hat{\gamma}_1 - \hat{\gamma}_0}, \tag{9.6.32}$$

where $(\hat{\gamma}_1, \hat{\delta}_1)$ is the solution of

$$\min \sum_{i:c \le x_i < c+h} (d_i - \gamma_1 - \delta_1(w_i - c))^2 \tag{9.6.33}$$

and $(\hat{\gamma}_0, \hat{\delta}_0)$ is the solution of

$$\min \sum_{i:c-h \le x_i < c} (d_i - \gamma_0 + \delta_0(w_i - c))^2. \tag{9.6.34}$$

(For a survey of RD, see Imbens, and Lemieux 2008.)

#### *9.6.3.3 Summary of Cross-Sectional Approaches*

The advantages of the parametric approach are that it can simultaneously take account of both selection on observables and selection on unobservables. It can also estimate the impact of each explanatory variable. The disadvantage is that it needs to impose both functional form and distributional assumptions. If the prior information is inaccurate, the resulting inferences are misleading. The advantages of the nonparametric approach are that there is no need to impose any assumption on the conditional mean functions or the effects of unobservables. The disadvantages are that some sort of conditional independence assumption have to hold conditional on some confounding variables. Hence it only takes into account the issues of selection on observables; neither is it feasible to estimate the impact of each observable factor. In other words, the advantages of the parametric approaches are the disadvantages of nonparametric approach. The disadvantages of the parametric approach are the advantages of the nonparametric approach.

### 9.6.4 Panel Data Approach

Panel data contains information over time for a number of individuals. Some of the observed individuals could be receiving treatment for part of the observed periods and no treatment for the rest of the observed periods. Some could be receiving treatment and some no treatment for the whole sample periods. The information on interindividual differences and intraindividual dynamics could lessen the restrictions imposed on the adjustment approaches using cross-sectional data alone.

#### *9.6.4.1 Parametric Approach*

One of the common assumptions using cross-sectional data is to assume that the observable factors, $\mathbf{x}$, are orthogonal to the impact of unobservable factors, $\epsilon^0$ and $\epsilon^1$ (e.g., (9.6.1) and (9.6.2)), even if it allows the joint dependence of $(\epsilon^1, \epsilon^0, d)$. However, the impact of unobservable factors could be correlated with observable factors, $\mathbf{x}$. Panel data allow us to control the correlation between $(\epsilon^0, \epsilon^1)$ and $\mathbf{x}$, in addition to the correlation between $(\epsilon^0, \epsilon^1)$ and $d$. For

instance, suppose that the outcome equations and participation equation are of the form

$$y_{it}^{1*} = \mathbf{x}_{it}'\boldsymbol{\beta}_1 + \epsilon_{it}^1, \tag{9.6.35}$$

$$y_{it}^{0*} = \mathbf{x}_{it}'\boldsymbol{\beta}_0 + \epsilon_{it}^0, \tag{9.6.36}$$

$$d_{it} = 1(\mathbf{x}_{it}'\gamma + v_{it} > 0), \tag{9.6.37}$$

where

$$\epsilon_{it}^j = \alpha_i^j + u_{it}^j, \quad j = 0, 1, \tag{9.6.38}$$

and $u_{it}^j$ is i.i.d. with mean 0 and constant variance. If the correlations between $\epsilon_{it}^j$ and $d_{it}$ are not confined to the individual specific components, $\alpha_i^j$ with $d_{it}$, but also the individual time-varying component $u_{it}^j$ so $E(u_{it}^j v_{it}) \neq 0$, the panel data fixed-effects sample selection estimators of Kyriazidou (1997), Honoré (1992), etc. (as summarized in Chapter 8.4) can be used to control the impact of unobserved heterogeneity, $\alpha_i^1, \alpha_i^0$, and estimate the treatment effects (e.g., Hsiao, Shen, Wang, and Weeks (2007, 2008)).[13]

### *9.6.4.2 Nonparametric Approach*

#### *9.6.4.2 (i) Difference-in-Difference Method*

As discussed in Section 9.6.3, one of the critical assumption using the nonparametric approach is to assume conditional independence between the outcomes $(y_1, y_0)$ and participation, $d$, conditional on $\mathbf{x}$. To avoid the issue of the curse of dimensionality, Rosenbaum and Rubin (1983) propose a propensity score matching method. However, the propensity score matching adjustment to control the bias induced by selection on observables depends critically on the correct specification of the propensity score, Prob $(d_i = 1 \mid \mathbf{x}_i)$. With panel data, one can avoid the specification of the propensity score, Prob $(d_i = 1 \mid \mathbf{x}_i)$ if under the assumption that there is no selection bias and the impacts due to changes in $x$ over time are the same between the treatment group (those who received the treatment) and the control group (those who did not receive the treatment) through a difference-in-difference method (Imbens and Angrist 1994).

Assume a panel begins with all the individuals in the control group (i.e., no treatment). At some time during the sample span, some individuals received treatment at time $t$ and no treatment at time $s$, and some individuals neither received treatment at time $t$, nor at time $s$. Let $\psi = \{i \mid d_{it} = 1, d_{is} = 0\}$, and $\Psi = \{i \mid d_{it} = 0, d_{is} = 0\}$. Then the difference-in-difference estimate of the

[13] See Heckman and Hotz (1989) for other types of model specification tests.

ATE is

$$\begin{aligned}\widehat{\text{ATE}} = &[E(y_{it} \mid i \in \psi) - E(y_{is} \mid i \in \psi)] \\ &- [E(y_{it} \mid i \in \Psi) - E(y_{is} \mid i \in \Psi)].\end{aligned} \tag{9.6.39}$$

For instance, the Northern Territory in Australia considered marijuana use a criminal act in 1995, but decriminalized it in 1996.[14] The Australian National Drug Strategy Household Surveys provide information about marijuana smoking behavior for residents of New Territories, New South Wales, Queensland Victoria, and Tasmania in 1995 and 2001; in all of them except New Territories (NT) it was nondecriminalized over this period. The percentage of smokers in NT in 1995 was 0.2342 and in 2001 it was 0.2845. The percentages of residents in nondecriminalized states in 1995 were 0.1423 and 0.1619 in 2001. The difference-in-difference estimate of the impact of decriminalization on marijuana usage is to raise the probability of smoking by

$$\begin{aligned}&\{(0.2845 - 0.2342) - (0.1619 - 0.1423)\} \\ &= 0.0503 - 0.0196 = 0.0307.\end{aligned} \tag{9.6.40}$$

#### *9.6.4.2 (ii) Predicting Counterfactuals Using Control Group Information*

The difference-in-difference method will provide a valid measurement of treatment effects under fairly restrictive assumptions. Namely, (1) there is no selection effect $E(\epsilon_i^0 \mid d_i) = E(\epsilon_i^0) = E(\epsilon_i^1) = E(\epsilon_i^1 \mid d_i) = 0$; (2) the marginal impacts of $\mathbf{x}$ are the same for those receiving treatment and not receiving treatment, $\frac{\partial E(y^{1*}|\mathbf{X})}{\partial \mathbf{X}} = \frac{\partial E(y^{0*}|\mathbf{X})}{\partial \mathbf{X}}$; and (3) changes in $\mathbf{x}$ for those in the treatment group and control group are the same, $E\{(x_{jt} - \mathbf{x}_{js}) \mid d_j = 1)\} = E\{(\mathbf{x}_{it} - \mathbf{x}_{is}) \mid d_i = 0\}$. However, with panel data, it is possible to relax these restrictive assumptions and still allow us to measure the treatment effects through the exploitation of correlations across individuals. Moreover, it also allows the treatment effects to vary over time.

Hsiao, Ching, and Wan (2012) propose to exploit the correlations across cross-sectional units to construct the counterfactuals. They assume the correlations across cross-sectional units are due to some common omitted factors. Decompose the outcomes of individual unit $i$ into the sum of two components, the impact of $K$ common factors that affect all individuals, $\mathbf{f}_t$, and an idiosyncratic component, $\alpha_i + \epsilon_{it}$, where $\alpha_i$ is fixed and $\epsilon_{it}$ is random with $E(\epsilon_{it}) = 0$ and $E(\epsilon_{it}\epsilon_{js}) = 0$ if $i \neq j$. The impact of common factors, $\mathbf{f}_t$, on individuals can be different for different individuals,

$$y_{it} = \alpha_i + \mathbf{b}_i' \mathbf{f}_t + \epsilon_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T. \end{matrix} \tag{9.6.41}$$

[14] Decriminalization does not mean smoking or possession of small amounts of marijuana is legal. It is still an offense to use or grow marijuana. An individual caught must pay a fine within a specified period to be eligible for the reduced penalty involving no criminal record or imprisonment (e.g., Damrongplasit and Hsiao 2009).

Then the contemporaneous covariance between $y_{it}$ and $y_{jt}$ is given by

$$\text{Cov}\,(y_{it}, y_{jt}) = \mathbf{b}_i' E(\mathbf{f}_t \mathbf{f}_t')\mathbf{b}_j. \tag{9.6.42}$$

Stacking the $N \times 1$ $y_{it}$ into a vector,

$$\mathbf{y}_t = B\,\mathbf{f}_t + \boldsymbol{\alpha} + \boldsymbol{\epsilon}_t, \tag{9.6.43}$$

where $\mathbf{y}_t = (y_{1t}, \ldots, y_{Nt})'$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)'$, $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \ldots, \epsilon_{Nt})'$, and $B$ is the $N \times K$ factor loading matrix, $B = (\mathbf{b}_1, \ldots, \mathbf{b}_N)'$. Suppose all $N$ units did not receive the treatment for $t = 1, \ldots, T_1$, that is, $\mathbf{y}_t = \mathbf{y}_t^{0*}$, but from time period $T_1 + 1$ onwards, the first unit received treatment, $y_{1t} = y_{1t}^{1*}$, $t = T_1 + 1, \ldots, T$, while the rest of individuals did not, $y_{it} = y_{it}^{0*}$, $t = 1, \ldots, T$ for $i = 2, \ldots, N$. As long as

$$E(\epsilon_{it} \mid d_{1t}) = 0, \quad i = 2, \ldots, N, \tag{9.6.44}$$

one can write

$$\begin{aligned} y_{1t}^{0*} &= E(y_{1t}^{0*} \mid \tilde{\mathbf{y}}_t) + \eta_{1t}, \quad t = 1, \ldots, T, \\ &= a + \mathbf{c}'\tilde{\mathbf{y}}_t + \eta_{1t}, \end{aligned} \tag{9.6.45}$$

where $\tilde{\mathbf{y}}_t = (y_{2t}, \ldots, y_{Nt})'$ and $E(\eta_{1t} \mid \tilde{\mathbf{y}}_t) = 0$. It is shown by Hsiao, Ching, and Wan (2012) that minimizing

$$\frac{1}{T_1}(\mathbf{y}_1^0 - \mathbf{e}a - Y\mathbf{c})'A(\mathbf{y}_1^0 - \mathbf{e}a - Y\mathbf{c}) \tag{9.6.46}$$

yields consistent estimates of $a$ and $\mathbf{c}$, where $\mathbf{y}_1^0 = (y_{11}, \ldots, y_{1T_1})'$, $\mathbf{e}$ is a $T_1 \times 1$ vector of 1's, $Y$ is a $T_1 \times (N-1)$ matrix of $T_1$ time series observations of $\tilde{\mathbf{y}}_t$, and $A$ is a $T_1 \times T_1$ positive definite matrix. From the estimates $(\hat{a}, \hat{\mathbf{c}}')$, one can construct the predicted value of the first unit in the absence of treatment, $y_{1t}^{0*}$, by

$$\hat{y}_{1t}^{0*} = \hat{a} + \hat{\mathbf{c}}'\tilde{\mathbf{y}}_t, \quad t = T_1 + 1, \ldots, T. \tag{9.6.47}$$

The treatment effect on the first unit can then be estimated by

$$\hat{\Delta}_{1t} = y_{1t} - \hat{y}_{1t}^{0*}, \quad t = T_1 + 1, \ldots, T. \tag{9.6.48}$$

The construction of the standard error of $\hat{y}_{1t}^{0*}$, $\sigma_{y_{1t}^0}$, follows from the standard prediction error formula. For instance, if $\eta_{1t}$ is independently, identically distributed over time, then

$$\sigma_{y_{1t}^0}^2 = \sigma_{\eta_1}^2[1 + (1, \tilde{\mathbf{y}}_t')(Y'Y)^{-1}(1, \tilde{\mathbf{y}}_t')']. \tag{9.6.49}$$

Hence, the confidence band for $\Delta_{1t}$ can be easily constructed as

$$\hat{\Delta}_{1t} \pm c\sigma_{y_{1t}^0}, \tag{9.6.50}$$

where $c$ is chosen by the desired confidence level.

Cross-sectional data provide measurement of policy intervention as a once-and-for-all impact. Panel data allow the policy impact to be evolutionary. If $\Delta_{1t}$

is serially correlated, but stationary, one can further model the time-varying treatment effects by an autoregressive moving average model using Box–Jenkins (1970) methodology

$$a(L)\Delta_{1t} = \mu + \theta(L)\eta_t \tag{9.6.51}$$

where $L$ is the lag operator, $\eta_t$ is an i.i.d. process with 0 mean and constant variance, and the roots of $\theta(L) = 0$ lie outside the unit circle. If the roots of $a(L) = 0$ all lie outside the unit circle, the treatment effect is stationary, and the long-term treatment effect is

$$\Delta_1 = a(L)^{-1}\mu = \mu^*. \tag{9.6.52}$$

Alternatively, one can estimate the long-run impact by taking the simple average of $\hat{\Delta}_{1t}$. When both $T_1$ and $(T - T_1)$ go to infinity,

$$\operatorname*{plim}_{(T-T_1)\to\infty} \frac{1}{T - T_1} \sum_{t=T_1+1}^{T} \hat{\Delta}_{1t} = \Delta_1 \tag{9.6.53}$$

The variance of (9.6.53) can be approximated by the heteroscedastic-autocorrelation consistent (HAC) estimator of Newey and West (1987).

Condition (9.6.44) makes no claim about the relationship between $d_{1t}$ and $\epsilon_{1t}$. They can be correlated. All we need is that the $j$th unit's idiosyncratic components $\epsilon_{jt}$ are independent of $d_{1t}$ for $j \neq 1$. The approach can be viewed as a "measurement without theory" approach or a nonparametric approach.

The parameters, $a$ and $\mathbf{c}$ can be obtained by regressing $y_{1t}$ on $y_{it}, i = 2, \ldots, N$, for $t = 1, \ldots, T_1$. Often $N$ is large. Using more or all cross-sectional units improves the within-sample fit, but does not necessarily yield more accurate post-sample prediction. One way to select the best combination of cross-sectional units to generate predicted $y_{1t}^{0*}$ for $t = T_1 + 1, \ldots, T$ is to use one of the model selection criteria (e.g., AIC (Akaike (1973)), AICC (Hurvich and Tsai (1989)) or BIC (Schwarz (1978)). For instance, Hsiao, Ching, and Wan (2012) suggest the following two-step procedure:

Step 1: Selection the best predictor for $y_{1t}^*$ using $j$ cross-sectional units out of $(N - 1)$ cross-sectional units, denoted by $M(j)^*$ by $R^2$, for $j = 1, \ldots, N - 1$.

Step 2: From $M(1)^*, M(2)*, \ldots, M(N-1)^*$, choose $M(m)^*$ in terms of some model selection criterion.

#### *9.6.4.3 An Example – Measuring the Impact of the Closer Economic Partnership Arrangement on Hong Kong*

Hong Kong signed Closer Economic Partnership Arrangement (CEPA) with Mainland China in June 2003 and started implementing its arrangement in January 2004. The CEPA aims to strengthen the linkage between Mainland China and Hong Kong by allowing Chinese citizens to enter Hong Kong as

Table 9.2. *AICC selected model using data for the period 1993Q1–2003Q4*

| | Beta | Std | $T$ |
|---|---|---|---|
| Constant | −0.0019 | 0.0037 | −0.524 |
| Austria | −1.0116 | 0.1682 | −6.0128 |
| Italy | −0.3177 | 0.1591 | −1.9971 |
| Korea | 0.3447 | 0.0469 | 7.3506 |
| Mexico | 0.3129 | 0.051 | 6.1335 |
| Norway | 0.3222 | 0.0538 | 5.9912 |
| Singapore | 0.1845 | 0.0546 | 3.3812 |
| $R^2 = 0.931$ | | | |
| AICC = −378.9427 | | | |

*Source:* Hsiao et al. (2012, Table 20).

individual tourists and liberalizing trade in services, enhancing cooperation in the area of finance, and promoting trade and investment facilitation and mutual recognition of professional qualifications. The implementation of CEPA started on January 1, 2004, where 273 types of Hong Kong products could be exported to the Mainland tariff free; another 713 types on January 1, 2005; 261 on January 1, 2006; and a further 37 on January 2007. Chinese citizens residing in selected cities are also allowed to visit Hong Kong as individual tourists, from 4 cities in 2003 to 49 cities in 2007, covering all 21 cities in Guangdong province.

Hsiao, Ching, and Wan (2012) tried to assess the impact of economic integration of Hong Kong with Mainland China on Hong Kong's economy by comparing what actually happened to Hong Kong's real GDP growth rates with what would have been if there were no CEPA with Mainland China in 2003. More specifically, they analyzed how these events have changed the growth rate of Hong Kong.

Because Hong Kong, by comparison, is tiny relative to other regions, Hsiao et al. (2012) believe that whatever happened in Hong Kong will have no bearing on other countries. In other words, they expect (9.6.44) to hold. Therefore, they use quarterly real growth rate of Australia, Austria, Canada, China, Denmark, Finland, France, Germany, Indonesia, Italy, Japan, Korea, Malaysia, Mexico, Netherlands, New Zealand, Norway, Philippines, Singapore, Switzerland, Taiwan Thailand, UK, and US to predict the quarterly real growth rate of Hong Kong in the absence of intervention. All the nominal GDP and CPI are from Organisation for Economic Co-operation and Development (OECD) Statistics, International Financial Statistics, and the CEIC database.

Using the AICC criterion, the countries selected are Austria, Italy, Korea, Mexico, Norway, and Singapore. Ordinary least-squares (OLS) estimates of the weights are reported in Table 9.2. Actual and predicted growth path from 1993Q1 to 2003Q4 are plotted in Figure 9.1. The availability of more

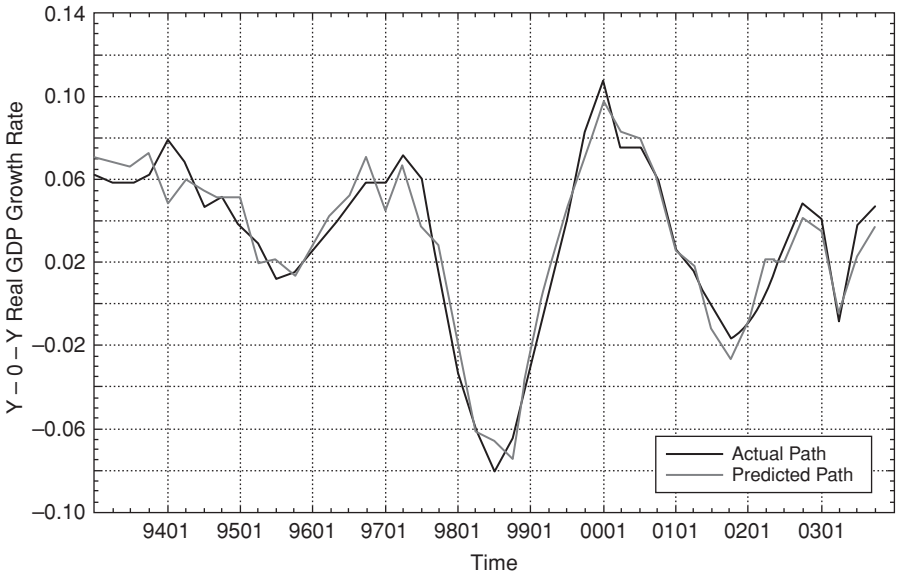


Figure 9.1. Actual and AICC predicted real GDP growth rate from 1993Q1 to 2003Q4. *Source:* Hsiao et al. (2012, Fig. 7).

Table 9.3. *Treatment effect for economic integration 2004Q1–2008Q1 based on AICC selected model*

| | Actual | Control | Treatment |
|---|---|---|---|
| Q1-2004 | 0.077 | 0.0493 | 0.0277 |
| Q2-2004 | 0.12 | 0.0686 | 0.0514 |
| Q3-2004 | 0.066 | 0.0515 | 0.0145 |
| Q4-2004 | 0.079 | 0.0446 | 0.0344 |
| Q1-2005 | 0.062 | 0.0217 | 0.0403 |
| Q2-2005 | 0.071 | 0.0177 | 0.0533 |
| Q3-2005 | 0.081 | 0.0333 | 0.0477 |
| Q4-2005 | 0.069 | 0.029 | 0.04 |
| Q1-2006 | 0.09 | 0.0471 | 0.0429 |
| Q2-2006 | 0.062 | 0.0417 | 0.0203 |
| Q3-2006 | 0.064 | 0.025 | 0.039 |
| Q4-2006 | 0.066 | 0.0009 | 0.0651 |
| Q1-2007 | 0.055 | −0.0101 | 0.0651 |
| Q2-2007 | 0.062 | 0.0092 | 0.0528 |
| Q3-2007 | 0.068 | 0.0143 | 0.0537 |
| Q4-2007 | 0.069 | 0.0508 | 0.0182 |
| Q1-2008 | 0.073 | 0.0538 | 0.0192 |
| MEAN | 0.0726 | 0.0323 | 0.0403 |
| STD | 0.0149 | 0.0213 | 0.016 |
| T | 4.8814 | 1.5132 | 2.5134 |

*Source:* Hsiao et al. (2012, Table 21).

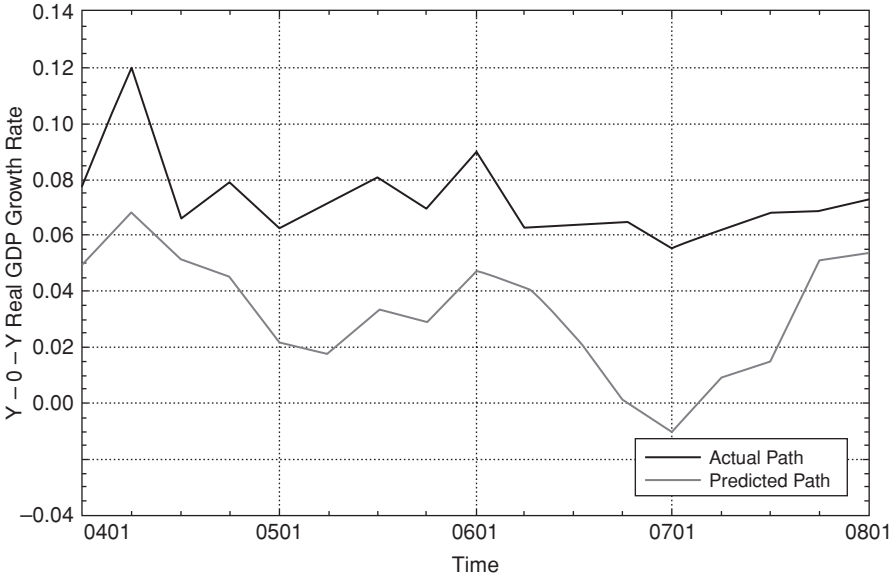


Figure 9.2. AICC – Actual and counterfactual real GDP growth rate from 2004Q1 to 2008Q1. *Source:* Hsiao et al. (2012, Fig. 8).

preintervention period data appears to allow more accurate estimates of the country weights and better tracing of the preintervention path. The estimated quarterly treatment effects are reported in Table 9.3. The actual and predicted counterfactual for the period 2004Q1 to 2008Q1 are presented in Figure 9.2.

Table 9.4. *AIC selected model using data for the period 1993Q1–2003Q4*

| | Beta | Std | $T$ |
|---|---|---|---|
| Constant | −0.003 | 0.0042 | −0.7095 |
| Austria | −1.2949 | 0.2181 | −5.9361 |
| Germany | 0.3552 | 0.233 | 1.5243 |
| Italy | −0.5768 | 0.1781 | −3.2394 |
| Korea | 0.3016 | 0.0587 | 5.1342 |
| Mexico | 0.234 | 0.0609 | 3.8395 |
| Norway | 0.2881 | 0.0562 | 5.1304 |
| Switzerland | 0.2436 | 0.1729 | 1.4092 |
| Singapore | 0.2222 | 0.0553 | 4.0155 |
| Philippines | 0.1757 | 0.1089 | 1.6127 |
| $R^2 = 0.9433$ | | | |
| AIC $= -385.7498$ | | | |

*Source:* Hsiao et al. (2012, Table 22).

Table 9.5. *AIC–Treatment effect for economic integration 2004Q1–2008Q1 based on AIC selected model*

| | Actual | Control | Treatment |
|---|---|---|---|
| Q1-2004 | 0.077 | 0.0559 | 0.0211 |
| Q2-2004 | 0.12 | 0.0722 | 0.0478 |
| Q3-2004 | 0.066 | 0.0446 | 0.0214 |
| Q4-2004 | 0.079 | 0.0314 | 0.0476 |
| Q1-2005 | 0.062 | 0.0121 | 0.0499 |
| Q2-2005 | 0.071 | 0.0126 | 0.0584 |
| Q3-2005 | 0.081 | 0.0314 | 0.0496 |
| Q4-2005 | 0.069 | 0.0278 | 0.0412 |
| Q1-2006 | 0.09 | 0.0436 | 0.0464 |
| Q2-2006 | 0.062 | 0.0372 | 0.0248 |
| Q3-2006 | 0.064 | 0.0292 | 0.0348 |
| Q4-2006 | 0.066 | 0.0122 | 0.0538 |
| Q1-2007 | 0.055 | 0.0051 | 0.0499 |
| Q2-2007 | 0.062 | 0.0279 | 0.0341 |
| Q3-2007 | 0.068 | 0.0255 | 0.0425 |
| Q4-2007 | 0.069 | 0.0589 | 0.0101 |
| Q1-2008 | 0.073 | 0.062 | 0.011 |
| Mean | 0.0726 | 0.0347 | 0.0379 |
| Std | 0.0149 | 0.0193 | 0.0151 |
| $T$ | 4.8814 | 1.7929 | 2.5122 |

*Source:* Hsiao et al. (2012, Table 23).

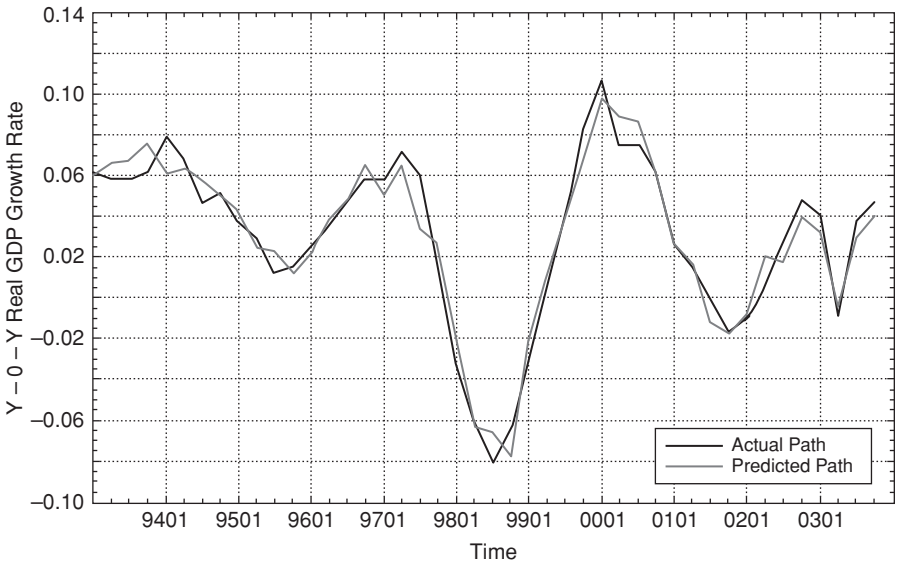


Figure 9.3. Actual and AIC predicted real GDP growth rate from 1993Q1 to 2003Q4. *Source:* Hsiao et al. (2012, Fig. 10).

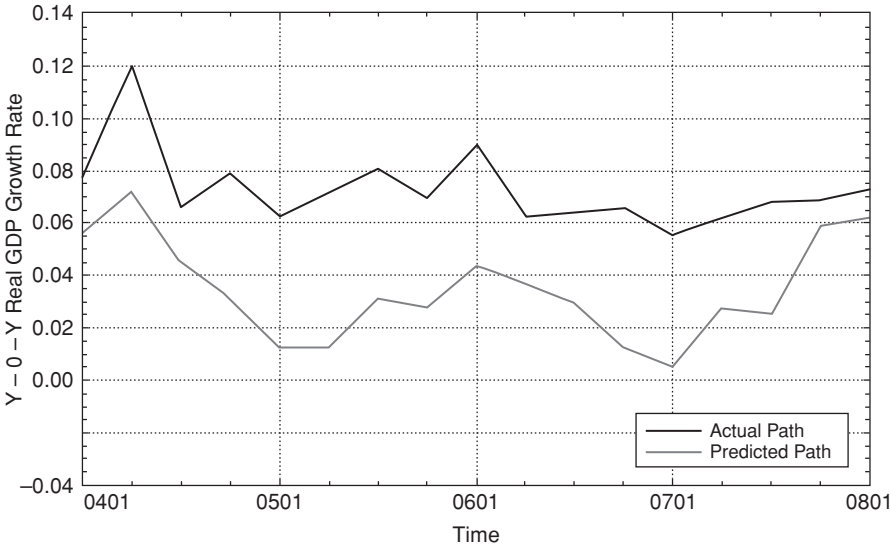


Figure 9.4. AIC – Actual and counterfactual real GDP growth rate from 2004Q1 to 2008Q1. *Source:* Hsiao et al. (2012, Fig. 11).

Using the AIC criterion, the selected group consists of Austria, Germany, Italy, Korea, Mexico, Norway, Philippines, Singapore, and Switzerland. The OLS estimates of the weights are in Table 9.4 and the estimated quarterly treatment effects are in Table 9.5. The pre- and post-intervention actual and predicted outcomes are plotted in Figures 9.3 and 9.4. It is notable that even though the two models use different combinations of countries, both groups of countries trace closely the actual Hong Kong path before the implementation of CEPA (with $R^2$ above .93). It is also quite remarkable that the post-sample predictions closely matched the actual turning points at a lower level for the treatment period even though no Hong Kong data were used. The CEPA effect at each quarter was all positive and appeared to be serially uncorrelated. The average actual growth rate from 2004Q1 to 2008Q1 is 7.26 percent. The average projected growth rate without CEPA is 3.23 percent using the group of countries selected by AICC and 3.47 percent using the group selected by AIC. The estimated average treatment effect is 4.03 percent with a standard error of 0.016 based on the AICC group and 3.79% with a standard error of 0.0151 based on the AIC group. The $t$-statistic is 3.5134 for the former group and 3.5122 for the latter group. Either set of countries yields similar predictions and highly significant CEPA effects. In other words, through liberalization and increased openness with Mainland China, the real GDP growth rate of Hong Kong is raised by more than 4 percent compared to the growth rate had there been no CEPA agreement with Mainland China.

CHAPTER 10

# Dynamic System

One of the prominent features of econometric analysis is the incorporation of economic theory into the analysis of numerical and institutional data. Economists, from León Walras onwards, perceive the economy as a coherent system. The interdependence of sectors of an economy is represented by a set of functional relations, each representing an aspect of the economy by a group of individuals, firms, or authorities. The variables entering into these relations consist of a set of *endogenous* (or *joint dependent*) variables, whose formations are conditional on a set of exogenous variables that economic theory regards as given. Two approaches have been proposed to model a system of economic behaviors – the structural equation approach and the reduced form approach. The structural approach constructs the system of behavioral equations from a priori assumed "theory," based on behavioral hypotheses and institutional and technological knowledge. However, different theoretical models may generate the same observed phenomena. To ensure the one-to-one relationship between the specified model and the observed phenomena, a priori restrictions need to be imposed to exclude other "observationally equivalent" models[1] (e.g., Dufour and Hsiao 2008, Hsiao 1983). The resulting statistical inference is conditioning on the hypothesized theoretical model. The statistical inference could be grossly misleading if the hypothesized model is not compatible with the data-generating process of the observed sample. Sims (1980) has criticized that many models are identified because of the "incredible" prior restrictions. Liu (1960), Sims (1980), etc. have therefore favored the reduced form approach. Because economic behavior is inherently dynamic due to institutional, technological, and behavioral rigidities, vector autoregressive models (VAR) have been proposed as a reduced-form formulation to take account of both the joint dependence of state variables and their dynamic dependence. We discuss panel vector autoregressive modeling when the time series dimension $T$ is fixed and cross-sectional dimension $N$ is large in Section 10.1. However, because the time series properties of a variable behave very differently if the variable

[1] By observationally equivalent structures we mean all structures that could generate the same observed sample characteristics (e.g., Hsiao 1983).

is stationary or nonstationary as time series dimension $T$ increases, Section 10.2 discusses the estimation of cointegrated system when both $N$ and $T$ are large. Section 10.3 discusses unit root and cointegration tests. Section 10.4 discusses the single-equation approach to estimating an equation in a dynamic simultaneous equations model.

## 10.1 PANEL VECTOR AUTOREGRESSIVE MODELS

### 10.1.1 "Homogeneous" Panel VAR Models

#### *10.1.1.1 Model Formulation*

Vector autoregressive models have become a widely used modeling tool in economics (e.g., Hsiao 1979a,b, 1982; Sims 1980). By "homogeneous" Panel VAR (PVAR) models we mean conditional on the unobserved time-invariant individual heterogeneity, the slope coefficients are identical over $i$ and $t$ (e.g., Holtz-Eakin, Newey, and Rosen 1988),

$$\Phi(L)\mathbf{w}_{it} = \mathbf{w}_{it} - \Phi_1\mathbf{w}_{i,t-1} \ldots - \Phi_p\mathbf{w}_{i,t-p} = \boldsymbol{\alpha}_i^* + \boldsymbol{\delta}^* t + \boldsymbol{\epsilon}_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T, \tag{10.1.1}$$

where $\mathbf{w}_{it}$ denotes an $m \times 1$ vector of observed random variables, $\boldsymbol{\alpha}_i^*$ is an $m \times 1$ vector of individual specific constants that vary with $i$, $\boldsymbol{\delta}^*$ is an $m \times 1$ vector of constants, $\boldsymbol{\epsilon}_{it}$ is an $m \times 1$ vector of random variables that is independently, identically distributed over $t$ with mean 0 and covariance matrix $\Omega$, and $\Phi(L) = I_m - \Phi_1 L - \cdots - \Phi_p L^p$ is a $p$th order polynomial of the lag operator $L$, $L^s\mathbf{w}_t = \mathbf{w}_{t-s}$.

With unrestricted intercepts or time trends, the time series property of $\mathbf{w}_{it}$ can be different whether $\mathbf{w}_{it}$ contains unit roots or not, or if $\mathbf{w}_{it}$ contains a unit root, if elements of $w_{it}$ are cointegrated.[2] (e.g., Johansen 1995; Pesaran, Shin, and Smith 2000; Phillips 1991; Sims, Stock, and Watson 1990). To make sure that the time series property of $\mathbf{w}$ remain the same whether $\mathbf{w}$ contains unit roots or not, instead of considering (10.1.1) directly, we consider

$$\Phi(L)(\mathbf{w}_{it} - \boldsymbol{\eta}_i - \boldsymbol{\delta} t) = \boldsymbol{\epsilon}_{it}, \tag{10.1.2}$$

where the roots of the determinant equation

$$| \Phi(\rho) | = 0 \tag{10.1.3}$$

[2] We say that $\mathbf{y}_t$ is stationary if $E\mathbf{y}_t = \boldsymbol{\mu}$, $E[(\mathbf{y}_t - \boldsymbol{\mu})(\mathbf{y}_{t-s} - \boldsymbol{\mu})'] = E[(\mathbf{y}_{t+q} - \boldsymbol{\mu})(\mathbf{y}_{t+q-s} - \boldsymbol{\mu})']$. We say that $\mathbf{y}_t$ is integrated of order $d$, $I(d)$, if $(1-L)^d\mathbf{y}_t$ is stationary, $I(0)$. If $\mathbf{y}_t \sim I(d)$ but $\boldsymbol{\beta}'\mathbf{y}_t \sim I(d-c)$, say $d = 1, c = 1$, then $y_t$ is cointegrated of order $c$. The maximum number of linearly independent vector $\boldsymbol{\beta}$ is called the rank of cointegration. For any $m \times 1$ $I(d)$ process, the cointegration rank can vary between 0 and $m - 1$ (e.g., Engle and Granger 1987; Intriligator, Bodkin, and Hsiao 1996).

are either equal to unity or fall outside the unit circle. Under the assumption that $E\boldsymbol{\epsilon}_{it} = \mathbf{0}$, it follows that

$$E(\mathbf{w}_{it} - \boldsymbol{\eta}_i - \boldsymbol{\delta}t) = \mathbf{0}. \tag{10.1.4}$$

To allow for the possibility of the presence of unit roots, we assume that

$$E(\mathbf{w}_{it} - \boldsymbol{\eta}_i - \boldsymbol{\delta}t)(\mathbf{w}_{it} - \boldsymbol{\eta}_i - \boldsymbol{\delta}t)' = \Psi_t. \tag{10.1.5}$$

Model (10.1.2)–(10.1.5) encompasses many well known panel VAR models (PVAR) as special cases. For instance,

(1) Stationary PVAR with individual-specific effects.

Let $\boldsymbol{\delta} = \mathbf{0}_{m\times 1}$. If all roots of (10.1.3) fall outside the unit circle, (10.1.2) becomes (10.1.1) with $\boldsymbol{\alpha}_i^* = -\Pi\boldsymbol{\eta}_i$ and

$$\Pi = -\left(I_m - \sum_{j=1}^{p} \Phi_j\right). \tag{10.1.6}$$

(2) Trend-stationary PVAR with individual-specific effects.

If all roots of (10.1.3) fall outside the unit circle and $\boldsymbol{\delta} \neq \mathbf{0}$, we have

$$\Phi(L)\mathbf{w}_{it} = \boldsymbol{\alpha}_i^* + \boldsymbol{\delta}^* t + \epsilon_{it}, \tag{10.1.7}$$

where $\boldsymbol{\alpha}_i^* = -\Pi\boldsymbol{\eta}_i + (\Gamma + \Pi)\boldsymbol{\delta}$,

$$\Gamma = -\Pi + \sum_{j=1}^{p} j\Phi_j, \tag{10.1.8}$$

and $\boldsymbol{\delta}^* = -\Pi\boldsymbol{\delta}$.

(3) PVAR with unit roots (but non-cointegrated) and individual-specific effects.

$$\Phi^*(L)\Delta\mathbf{w}_{it} = -\Pi^*\boldsymbol{\delta} + \boldsymbol{\epsilon}_{it} \tag{10.1.9}$$

where $\Delta = (1 - L)$,

$$\Phi^*(L) = I_m - \sum_{j=1}^{p-1} \Phi_j^* L^j, \tag{10.1.10}$$

$\Phi_j^* = -(I_m - \sum_{\ell=1}^{j} \Phi_\ell)$, $j = 1, 2\ldots, p-1$, and $\Pi^* = -(I_m - \sum_{j=1}^{p-1} \Phi_j^*)$.

(4) Cointegrated PVAR with individual-specific effects.

If some roots of (10.1.3) are equal to unity and rank $(\Pi) = r$, $0 < r < m$, (10.1.2) may be rewritten in the form of a panel vector error corrections model:

$$\Delta\mathbf{w}_{it} = \boldsymbol{\alpha}_i^* + (\Gamma + \Pi)\boldsymbol{\delta} + \boldsymbol{\delta}^* t + \Pi\mathbf{y}_{i,t-1} + \sum_{j=1}^{p-1} \Gamma_j \Delta\mathbf{w}_{i,t-j} + \boldsymbol{\epsilon}_{it}, \tag{10.1.11}$$

where $\Gamma_j = -\sum_{s=j+1}^{p} \Phi_s$, $j = 1, \ldots, p-1$, and $\Pi$ can be decomposed as the product of two $m \times r$ matrices $J$ and $\boldsymbol{\beta}$, with rank $r$, $\Pi = J\boldsymbol{\beta}'$ and $J_\perp'\boldsymbol{\beta}_\perp$ is of

rank $m - r$, where $J_\perp$ and $\boldsymbol{\beta}_\perp$ are $m \times (m - r)$ matrices of full column rank such that $J'J_\perp = \mathbf{0}$ and $\boldsymbol{\beta}'\boldsymbol{\beta}_\perp = \mathbf{0}$ (Johansen 1995).

The reason for formulating the PVAR model in terms of (10.1.2)–(10.1.5) rather than (10.1.1) is that it puts restrictions on the model intercepts and trend term so that the time series properties of $\mathbf{w}_{it}$ remain the same with the presence of unit roots and cointegration. For instance, when $\boldsymbol{\delta} = \mathbf{0}$ and whether the roots of (10.1.3) all fall outside the unit circle, or one or more roots of (10.1.3) are equal to unity, $\mathbf{w}_{it}$ exhibit no trend growth. However, if $\boldsymbol{\alpha}_i^*$ is unrestricted, then $\mathbf{w}_{it}$ will exhibit differential trend growth if unit roots are present. If $\boldsymbol{\delta} \neq \mathbf{0}$, (10.1.2) ensures that the trend growth of $\mathbf{w}_{it}$ is linear whether the roots of (10.1.3) are all outside the unit circle or some or all are unity. But if the trend term is unrestricted, then $\mathbf{w}_{it}$ exhibit a linear trend if the roots of (10.1.3) all fall outside the unit circle and would exhibit quadratic trends if one or more roots of (10.1.3) are equal to unity (e.g., Pesaran, Shin, and Smith 2000).

If $\boldsymbol{\alpha}_i^*$ are assumed randomly distributed with a common mean and constant covariance matrix, (10.1.1) is a random effects PVAR. The random-effects PVAR has the advantages that the number of unknown parameters stay constant as sample size increases and the efficient inference of $\Phi(L)$ can be derived by considering the marginal distribution of $(\mathbf{w}_{io}, \ldots, \mathbf{w}_{it})$,

$$
\begin{aligned}
f(\mathbf{w}_{io}, \ldots, \mathbf{w}_{iT}) &= \int f(\mathbf{w}_{io}, \ldots, \mathbf{w}_{iT} \mid \boldsymbol{\alpha}_i^*) dG(\alpha_i^*) \\
&= \int \prod_{t=p}^{T} f(\mathbf{w}_{it} \mid \mathbf{w}_{i,t-1}, \ldots, \mathbf{w}_{i,t-p}, \boldsymbol{\alpha}_i^*) \\
&\quad \cdot f(\mathbf{w}_{ip}, \ldots, \mathbf{w}_{io} \mid \alpha_i^*) dG(\alpha_i^*). \qquad (10.1.12)
\end{aligned}
$$

However, besides the difficulties of postulating the probability distribution of unobserved effects, $\boldsymbol{\alpha}_i^*$, the derivation of (10.1.12) appears computationally complicated because (10.1.12) involves multiple integration of $m \times (T + 1)$ dimensions. On the other hand, treating $\alpha_i^*$ as fixed constant, $f(\mathbf{w}_{it} \mid \mathbf{w}_{i,t-1}, \ldots, \mathbf{w}_{t-p}; \boldsymbol{\alpha}_i^*)$ is independently distributed over $t$. Moreover, even if $\boldsymbol{\alpha}_i^*$ are random, the conditional inference of $f(\mathbf{w}_{io}, \ldots, \mathbf{w}_{iT} \mid \boldsymbol{\alpha}_i^*)$ remains valid, although it is not efficient. We therefore focus on inference with $\boldsymbol{\alpha}_i^*$ fixed and discuss conditional inference procedures.

When the time dimension of the panel is short, just as in the single-equation fixed-effects dynamic panel data model (Chapter 4, Section 4.5), (10.1.2) raises the classical incidental parameters problem and the issue of modeling initial observations. For ease of exposition, we shall illustrate the estimation and inference by considering $p = 1$, namely, the model of

$$
(I - \Phi L)(\mathbf{w}_{it} - \boldsymbol{\eta}_i - \boldsymbol{\delta} t) = \boldsymbol{\epsilon}_{it}, \qquad \begin{matrix} i = 1. \ldots, N, \\ t = 1, \ldots, T, \end{matrix} \qquad (10.1.13)
$$

We also assume that $\mathbf{w}_{i0}$ are available.

*10.1.1.2 GMM Estimation*

Just as in the single-equation case, the individual effects $\boldsymbol{\eta}_i$ can be eliminated by first differencing (10.1.13):

$$\Delta \mathbf{w}_{it} - \boldsymbol{\delta} = \Phi(\Delta \mathbf{w}_{i,t-1} - \boldsymbol{\delta}) + \Delta \boldsymbol{\epsilon}_{it}, t = 2, \ldots, T. \tag{10.1.14}$$

Thus, we have the orthogonality conditions,

$$E\left\{[(\Delta \mathbf{w}_{it} - \boldsymbol{\delta}) - \Phi(\Delta \mathbf{w}_{i,t-1} - \boldsymbol{\delta})]\mathbf{q}'_{it}\right\} = \mathbf{0}, \quad t = 2, \ldots, T. \tag{10.1.15}$$

where

$$\mathbf{q}_{it} = (1, \mathbf{w}'_{io}, \ldots, \mathbf{w}'_{i,t-2})'. \tag{10.1.16}$$

Stacking the $(T-1)$ (10.1.15) together yields

$$S_i = R_i \Lambda' + E_i, \quad i = 1, 2, \ldots, N, \tag{10.1.17}$$

where

$$S_i = (\Delta \mathbf{w}_{i2}, \Delta \mathbf{w}_{i3} \ldots, \Delta \mathbf{w}_{iT})', \quad E_i = (\Delta \boldsymbol{\epsilon}_{i2}, \ldots, \Delta \boldsymbol{\epsilon}_{iT})'$$

$$\begin{aligned} R_i &= (S_{i,-1}, \mathbf{e}_{T-1}), \quad S_{i,-1} = (\Delta \mathbf{w}_{i1}, \ldots, \Delta \mathbf{w}_{i,T-1})', \\ \Lambda &= (\Phi, \mathbf{a}_1), \quad \mathbf{a}_1 = (I_m - \Phi)\boldsymbol{\delta}, \end{aligned} \tag{10.1.18}$$

and $\mathbf{e}_{T-1}$ denotes a $(T-1) \times 1$ vector of 1's. Premultiplying (10.1.17) by the $(mT/2+1)(T-1) \times (T-1)$ block-diagonal instrumental variable matrix $Q_i$,

$$Q_i = \begin{pmatrix} \mathbf{q}_{i2} & \mathbf{0} & . & . & \mathbf{0} \\ \mathbf{0} & \mathbf{q}_{is} & . & . & \\ . & . & . & . & \\ . & . & & . & \\ \mathbf{0} & . & & & \mathbf{q}_{iT} \end{pmatrix}, \tag{10.1.19}$$

one obtains

$$Q_i S_i = Q_i R_i \Lambda' + Q_i E_i, \tag{10.1.20}$$

the transpose of which in vectorized form becomes[3]

$$(Q_i \otimes I_m)\,\mathrm{vec}\,(S'_i) = (Q_i R_i \otimes I_m)\boldsymbol{\lambda} + (Q_i \otimes I_m)\,\mathrm{vec}\,(E'_i), \tag{10.1.21}$$

where $\lambda = \mathrm{vec}\,(\Lambda)$ and vec (.) is the operator transforms a matrix into a vector by stacking the columns of the matrix one underneath the other. Thus,

[3] Vec $(ABC) = (C' \otimes A)$ vec $(B)$, see Magnus and Neudecker (1999).

the GMM estimator of $\boldsymbol{\lambda}$ can be obtained by minimizing (Binder, Hsiao, and Pesaran 2005)

$$\begin{aligned}&\left[\sum_{i=1}^{N}\Big((Q_i\otimes I_m)\,\text{vec}\,(S_i')-(Q_iR_i\otimes I_m)\boldsymbol{\lambda}\Big)\right]'\\&\cdot\left[\sum_{i=1}^{N}(Q_i\otimes I_m)\tilde{\Omega}(Q_i\otimes I_m)'\right]^{-1}\\&\cdot\left[\sum_{i=1}^{N}\Big((Q_i\otimes I_m)\,\text{vec}\,(S_i')-(Q_iR_i\otimes I_m)\boldsymbol{\lambda}\Big)\right],\end{aligned}\tag{10.1.22}$$

where

$$\tilde{\Omega}=\begin{bmatrix}2\Omega & -\Omega & \mathbf{0} & \dots & \mathbf{0}\\ -\Omega & 2\Omega & -\Omega & & \\ \mathbf{0} & -\Omega & 2\Omega & & \\ \vdots & & & \ddots & \\ \mathbf{0} & & & & 2\Omega\end{bmatrix}.\tag{10.1.23}$$

The moment conditions relevant to the estimation of $\Omega$ are given by

$$\begin{aligned}E\{[\Delta\mathbf{w}_{it}-\boldsymbol{\delta}-\Phi(\Delta\mathbf{w}_{i,t-1}-\boldsymbol{\delta})][\Delta\mathbf{w}_{it}-\boldsymbol{\delta}\\-\Phi(\Delta\mathbf{w}_{i,t-1}-\boldsymbol{\delta})]'-2\Omega\}=\mathbf{0},\ t=2,3,\dots,T.\end{aligned}\tag{10.1.24}$$

Also, in the trend-stationary case, on estimation of $\mathbf{a}_1$, $\boldsymbol{\delta}$ may be obtained as

$$\boldsymbol{\delta}=(I_m-\hat{\Phi})^{-1}\hat{\mathbf{a}}_1.\tag{10.1.25}$$

The generalized method of moments (GMM) estimator is consistent and asymptotically normally distributed as $N\to\infty$ if all the roots of (10.1.3) fall outside the unit circle, but breaks down if some roots are equal to unity. To see this, note that a necessary condition for the GMM estimator (10.1.22) to exist is that rank $(N^{-1}\sum_{i=1}^{N}Q_iR_i)=m+1$ as $N\to\infty$. In the case where $\Phi=I_m$, $\Delta\mathbf{w}_{it}=\boldsymbol{\delta}+\boldsymbol{\epsilon}_{it}$, and $\mathbf{w}_{it}=\mathbf{w}_{io}+\boldsymbol{\delta}t+\sum_{\ell=1}^{t}\boldsymbol{\epsilon}_{i\ell}$. Thus it follows that for $t=2,3,\dots,T$, $j=0,1,\dots,t-2$, as $N\to\infty$,

$$\frac{1}{N}\sum_{i=1}^{N}\Delta\mathbf{w}_{i,t-1}\mathbf{w}_{ij}'\longrightarrow\boldsymbol{\delta}(\mathbf{w}_{io}+\boldsymbol{\delta}j)',\tag{10.1.26}$$

which is of rank 1. In other words, when $\Phi=I_m$, the elements of $\mathbf{q}_{it}$ are not legitimate instruments.

### *10.1.1.3 (Transformed) Maximum-Likelihood Estimator*

We note that given on $\Delta\mathbf{w}_{i1}$ (10.1.14) is well defined for $t=2,\dots,T$. However, $\Delta\mathbf{w}_{i1}$ is random. Equation (10.1.13) implies that $\Delta\mathbf{w}_{i1}$ equals

$$\Delta\mathbf{w}_{i1}-\boldsymbol{\delta}=-(I-\Phi)(\mathbf{w}_{io}-\boldsymbol{\eta}_i)+\boldsymbol{\epsilon}_{i1}.\tag{10.1.27}$$

We note that by (10.1.4) and (10.1.5), $E(\Delta \mathbf{w}_{i1} - \boldsymbol{\delta}) = -(I - \Phi)E(\mathbf{w}_{io} - \boldsymbol{\eta}_i) + E\boldsymbol{\epsilon}_{i1} = \mathbf{0}$ and $E(\Delta \mathbf{w}_{i1} - \boldsymbol{\delta})(\Delta \mathbf{w}_{i1} - \boldsymbol{\delta})' = (I - \Phi)\Psi_0(I - \Phi)' + \Omega = \Psi_1$ where $\Psi_0 = E(\mathbf{w}_{io} - \boldsymbol{\eta}_i)(\mathbf{w}_{io} - \boldsymbol{\eta}_i)'$. Therefore, the joint likelihood of $\Delta \mathbf{w}_i' = (\Delta \mathbf{w}_{i1}', \ldots, \Delta \mathbf{w}_{iT}')$ is well defined and does not involve incidental parameters. Under the assumption that $\boldsymbol{\epsilon}_{it}$ is normally distributed, the likelihood function is given by

$$\prod_{i=1}^{N} (2\pi)^{-\frac{T}{2}} \mid \Omega^* \mid^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{r}_i - H_i\boldsymbol{\phi})'\Omega^{*-1}(\mathbf{r}_i - H_i\boldsymbol{\phi})\right], \tag{10.1.28}$$

where

$$\begin{aligned}
\mathbf{r}_i &= (\Delta \mathbf{w}_i - \mathbf{e}_T \otimes \boldsymbol{\delta}), \\
H_i &= G_i' \otimes I_m, \\
G_i &= (\mathbf{0}, \Delta \mathbf{w}_{i1} - \boldsymbol{\delta}, \ldots, \Delta \mathbf{w}_{i,T-1} - \boldsymbol{\delta}), \\
\boldsymbol{\phi} &= \text{vec}\,(\Phi),
\end{aligned}$$

$$\Omega^* = \begin{pmatrix}
\Psi_1 & -\Omega & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\
-\Omega & 2\Omega & -\Omega & \mathbf{0} & & \\
\mathbf{0} & -\Omega & 2\Omega & -\Omega & & \\
\vdots & & & \ddots & & \\
\mathbf{0} & & & & & 2\Omega
\end{pmatrix}, \tag{10.1.29}$$

and $\mathbf{e}_T$ is a $T \times 1$ vector of $(1, \ldots, 1)'$. Maximizing the logarithm of (10.1.29), $\ell(\boldsymbol{\theta})$, with respect to $\boldsymbol{\theta}' = (\boldsymbol{\delta}', \boldsymbol{\phi}', \boldsymbol{\sigma}')'$, where $\boldsymbol{\sigma}$ denotes the unknown element of $\Omega^*$, yields the (transformed) maximum-likelihood estimator (MLE) that is consistent and asymptotically normally distributed with asymptotic covariance matrix given by $-E\left(\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right)^{-1}$ as $N \to \infty$ independent of whether $\mathbf{w}_{it}$ contains unit roots or cointegrated.

### *10.1.1.4 Minimum-Distance Estimator*

We note that conditional on $\Omega^*$, the MLE of $\Phi$ and $\boldsymbol{\delta}$ is equivalent to the minimum-distance estimator (MDE) that minimizes

$$\sum_{i=1}^{N} (\mathbf{r}_i - H_i\boldsymbol{\phi})'\Omega^{*-1}(\mathbf{r}_i - H_i\boldsymbol{\phi}). \tag{10.1.30}$$

Furthermore, conditional on $\boldsymbol{\delta}$ and $\Omega^*$, the MDE of $\Phi$ is given by

$$\hat{\boldsymbol{\phi}} = \left(\sum_{i=1}^{N} H_i'\Omega^{*-1}H_i\right)^{-1} \left(\sum_{i=1}^{N} H_i'\Omega^{*-1}\mathbf{r}_i\right). \tag{10.1.31}$$

Conditional on $\Phi$ and $\Omega^*$, the MDE of $\boldsymbol{\delta}$ is equal to

$$\hat{\boldsymbol{\delta}} = (NP\Omega^{*-1}P')^{-1}\left[\sum_{i=1}^{N} P\Omega^{*-1}(\Delta\mathbf{w}_i - L_i\boldsymbol{\phi})\right], \tag{10.1.32}$$

where

$$P = (I_m, I_m - \Phi', I_m - \Phi', \ldots, I_m - \Phi'), \tag{10.1.33}$$

and

$$L_i = K_i' \otimes I_m, \quad \text{and } K_i = (\mathbf{0}, \Delta\mathbf{w}_{i1}, \ldots, \Delta\mathbf{w}_{i,T-1}).$$

Conditional on $\boldsymbol{\delta}$,

$$\hat{\Psi}_1 = \frac{1}{N}\sum_{i=1}^{N}(\Delta\mathbf{w}_{i1} - \boldsymbol{\delta})(\Delta\mathbf{w}_{i1} - \boldsymbol{\delta})', \tag{10.1.34}$$

and conditional on $\boldsymbol{\delta}$, $\Phi$,

$$\begin{aligned}\hat{\Omega} = \frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=2}^{T}&[\Delta\mathbf{w}_{it} - \boldsymbol{\delta} - \Phi(\Delta\mathbf{w}_{i,t-1} - \boldsymbol{\delta})] \\ &[\Delta\mathbf{w}_{it} - \boldsymbol{\delta} - \Phi(\Delta\mathbf{w}_{i,t-1} - \boldsymbol{\delta})]'.\end{aligned} \tag{10.1.35}$$

We may iterate between (10.1.31) and (10.1.35) to obtain the feasible MDE using

$$\hat{\boldsymbol{\delta}}^{(0)} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\Delta\mathbf{w}_{it}, \tag{10.1.36}$$

and

$$\begin{aligned}\Phi^{(0)} = &\left[\sum_{i=1}^{N}\sum_{t=3}^{T}(\Delta\mathbf{w}_{it} - \boldsymbol{\delta}^{(0)})(\Delta\mathbf{w}_{i,t-2} - \boldsymbol{\delta}^{(0)})'\right] \\ &\cdot\left[\sum_{i=1}^{N}\sum_{t=3}^{T}(\Delta\mathbf{w}_{i,t-1} - \boldsymbol{\delta}^{(0)})(\Delta\mathbf{w}_{i,t-2} - \boldsymbol{\delta}^{(0)})'\right]^{-1}\end{aligned} \tag{10.1.37}$$

to start the iteration.

Conditional on $\Omega^*$, the MDE of $\boldsymbol{\phi}$ and $\boldsymbol{\delta}$ is identical to the MLE. When $\boldsymbol{\delta} = \mathbf{0}$ (no trend term), conditional on $\Omega^*$, the asymptotic covariance matrix of the MLE or MDE of $\boldsymbol{\phi}$ is equal to

$$\left[\sum_{i=1}^{N}(K_i \otimes I_m)\Omega^{*-1}(K_i' \otimes I_m)\right]^{-1}. \tag{10.1.38}$$

When $\Omega^*$ is unknown, the asymptotic variance–covariance matrices of MLE and MDE of $\boldsymbol{\phi}$ do not converge to (10.1.38) because when lagged dependent variables appear as regressors, the estimation of $\Phi$ and $\Omega^*$ is not asymptotically

independent. The asymptotic variance covariance matrix of the feasible MDE is equal to the sum of (10.1.38) and a positive semidefinite matrix attributable to the estimation error of $\Omega^*$ (Hsiao, Pesaran, and Tahmiscioglu 2002).

Both the MLE and MDE always exist whether $\mathbf{w}_{it}$ contains unit roots or not. The MLE and MDE are asymptotically normally distributed independent of whether $\mathbf{w}_{it}$ is (trend) stationary, integrated or cointegrated as $T$ is fixed and $N \to \infty$. Therefore, a conventional likelihood ratio test statistic or Wald type test statistic of unit root or the rank of cointegration can be approximated by chi-square statistics. Moreover, the limited Monte Carlo studies conducted by Binder, Hsiao, and Pesaran (2005) show that both the MLE and MDE perform very well in finite sample and dominate the conventional GMM, in particular if the roots of (10.1.3) are near unity.

### 10.1.2 Heterogeneous Vector Autoregressive Models

We shall say a VAR model heterogeneous if the slope coefficients also vary across individuals,

$$\Phi_i(L)\mathbf{w}_{it} = \boldsymbol{\alpha}_i^* + \epsilon_{it}, \quad i = 1, \ldots, N, \tag{10.1.39}$$

where

$$\Phi_i(L) = I_m - \Phi_{i1}L - \ldots - \Phi_{ip_i}L^{p_i} \text{ for } i = 1, \ldots, N. \tag{10.1.40}$$

When $\Phi_i(L) \neq \Phi_j(L)$ for $i \neq j$, there is no way one can get a consistent estimator of $\Phi_i(L)$ if $T$ is fixed. When the roots of the determinant equation (10.1.3) fall outside the unit circle, the least-squares estimator is consistent (at the speed of $\sqrt{T}$) and is asymptotically normally distributed as $T \to \infty$ (e.g., Anderson 1971). When the roots of (10.1.3) contain unit roots, the least-squares estimator remains consistent but it converges to the unit root at the speed of $T$ and its limiting distribution is nonstandard (e.g., Phillips and Durlauf 1986). Therefore, in this subsection, we restrict the discussion to the stationary case. We defer the discussion of nonstationary case to Section 10.2.

#### *10.1.2.1 Cross-Sectionally Independent Processes*

If $\boldsymbol{\epsilon}_{it}$ is independent over $i$ with mean $\mathbf{0}$ and unrestricted covariance matrix $\Omega_i$ and $T$ is large, applying the least-squares method to (10.1.39) equation by equation yields consistent and efficient estimates of $\Phi_i(L)$ and $\boldsymbol{\alpha}_i^*$.

#### *10.1.2.2 Cross-Sectionally Dependent Processes*

When $\boldsymbol{\epsilon}_{it}$ are cross-sectionally dependent, (10.1.39) can be put in Zellner's (1962) seemingly unrelated regression framework if $N$ is fixed and $T$ is large. We can first use $i$th individual's time series observation to estimate $\Phi_i(L)$ and $\boldsymbol{\alpha}_i^*$. Then we use the estimated $\hat{\Phi}_i(L)$ and $\hat{\boldsymbol{\alpha}}_i^*$ to obtain estimated $\hat{\boldsymbol{\epsilon}}_{it}$,

$i = 1, \ldots, N, t = 1, \ldots, T$ and estimate $\Omega_{ij} = E(\boldsymbol{\epsilon}_{it}\boldsymbol{\epsilon}'_{jt})$ by

$$\hat{\Omega}_{ij} = \frac{1}{T}\sum_{t=1}^{T} \hat{\boldsymbol{\epsilon}}_{it}\hat{\boldsymbol{\epsilon}}_{jt}. \tag{10.1.41}$$

Given $\hat{\Omega}_{ij}, i, j = 1, \ldots, N$, one can stack all $N$ cross-sectionally equations (10.1.39) together and apply the feasible generalized least-squares estimator to obtain efficient estimate of $\Phi_i(L)$ and $\boldsymbol{\alpha}_i^*, i = 1, \ldots, N$.

### *10.1.2.3 Global VAR*

To obtain efficient estimates of $\Phi_i(L)$ using the Zellner (1962) seemingly unrelated regression approach requires $T$ to be considerably larger than $N$. In many macroeconomic applications, the number of time series observations, $T$, could be of the same magnitude as the number of cross-sectional dimensions, $N$. When $N$ is large, it is not feasible to stack all $Nm$ equations together as a system. Pesaran, Schuermann, and Weiner (2004) propose a global VAR (GVAR) to accommodate dynamic cross-dependence by considering

$$\Phi_i(L)(\mathbf{w}_{it} - \Gamma_i \mathbf{w}_{it}^*) = \boldsymbol{\epsilon}_{it}, \quad i = 1, 2, \ldots, N, \tag{10.1.42}$$

where

$$\mathbf{w}_{it}^* = \sum_{j=1}^{N} r_{ij}\mathbf{w}_{jt}, \tag{10.1.43}$$

$$r_{ii} = 0, \sum_{j=1}^{N} r_{ij} = 1, \text{ and } \sum_{j=1}^{N} r_{ij}^2 \to 0 \text{ as } N \to \infty. \tag{10.1.44}$$

The weight $r_{ij}$ could be $\frac{1}{N-1}$ for $i \neq j$, or constructed from trade value or other measures of some economic distance and could be time-varying. Just like the cross-sectionally mean augment regression approach discussed in Chapter 9, Section 9.4, the global average $\mathbf{w}_{it}^*$ is inserted into (10.1.39) to take account of the cross-sectional dependence. When $\mathbf{w}_{i,t-s}^*$ can be treated as weakly exogenous (predetermined), the estimation of (10.1.42) for each $i$ can proceed using standard time series estimation techniques (e.g., Pesaran, Shin, and Smith 2000). Pesaran et al. (2004) show that the weak exogeneity assumption of $\mathbf{w}_{it}^*$ holds for all countries except for the United States because of the United States' dominant position in the world. They also show that (10.1.42) yields better results than (10.1.39) when cross-sectional units are correlated.[4]

[4] The computer program for GVAR and some data sources can be downloaded from http://www-cfab.jbs.cam.ac.uk/research/gvartoolbox/index.html.

## 10.2 COINTEGRATED PANEL MODELS AND VECTOR ERROR CORRECTION

### 10.2.1 Properties of Cointegrated Processes

Many macro and financial data are nonstationary. The nonstationarity of a time series is usually represented by an integrated process of order $d$, $d \geq 1$, $I(d)$. A $d$th integrated process can be transformed into a stationary process by differencing the variable $d$-times, $(1-L)^d$ (e.g., Box–Jenkins 1970). For instance, suppose all the elements of $\mathbf{w}_{it}$ are $I(1)$; then $(1-L)\mathbf{w}_{it}$ become stationary ($I(0)$ processes). However, differencing $\mathbf{w}_{it}$ also removes the underlying long-run relations among the elements of $\mathbf{w}_{it}$, which can have important economic implications. If $\mathbf{w}_{it}$ are driven by some common nonstationary variables, one notable feature is that linear combinations of $\mathbf{w}_{it}$ can remove these *common trends* and become stationary ($I(0)$). Such linear combinations capture the long-run relations among $\mathbf{w}_{it}$ and are called "cointegrating" relations.

Let $\mathbf{w}_{it}$ be an $m \times 1$ vector of random variables. We assume that each element of $\mathbf{w}_{it}$, $w_{jit}$, is integrated of order 1, $I(1)$,

$$w_{jit} \sim I(1), \quad j = 1, \ldots, m. \tag{10.2.1}$$

Following Engle and Granger (1987), we say that the elements of $\mathbf{w}_{it}$ forming $r(\geq 1)$ cointegrating relations if there exist $r$ linearly independent combinations of $\mathbf{w}_{it}$ that are stationary $I(0)$,

$$C_i \mathbf{w}_{it} = \mathbf{u}_{it} \sim I(0), \tag{10.2.2}$$

where $C_i$ denotes the $r \times m$ constant matrix with rank $(C_i) = r$, and $\mathbf{u}_{it}$ denotes the $r \times 1$ random vectors with $E(\mathbf{u}_{it}) = 0$, $E(\mathbf{u}_{it}\mathbf{u}_{it}') = \tilde{\Omega}_{io}$, $E(\mathbf{u}_{it}\mathbf{u}_{i,t-s}') = \tilde{\Omega}_{is}$, and rank $(\tilde{\Omega}_{io}) = r$.

Rewriting $\mathbf{w}_{it}$ as the sum of the impact of $(m-r)$ $I(1)$ common trends $\mathbf{z}_{it}$ and stationary components, $\xi_{it}$,

$$\mathbf{w}_{it} = A_i \mathbf{z}_{it} + \boldsymbol{\xi}_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T. \tag{10.2.3}$$

where

$$\begin{matrix} \mathbf{z}_{it} \\ (m-r) \times 1 \end{matrix} \quad \sim I(1), \tag{10.2.4}$$

and $A_i = (\mathbf{a}_i')$ is an $m \times (m-r)$ constant matrix. The cointegrating relations (10.2.2) imply that

$$C_i A_i = \mathbf{0}, \quad i = 1, \ldots, N. \tag{10.2.5}$$

If $A_i \neq \mathbf{0}$ with rank $(A_i) = m - r$, $0 < r < m$ and $\mathbf{z}_{it} \neq \mathbf{z}_{jt}$, the $m \times 1$ $I(1)$ random variables $\mathbf{w}_{it}$ are cointegrated. There exists an error correction representation (VEC) of $\mathbf{w}_{it}$ (Engle and Granger 1987),

$$\Delta \mathbf{w}_{it} = \Pi_i^* \mathbf{w}_{i,t-1} + \sum_{s=1}^{p_i} \Phi_{is}^* \Delta \mathbf{w}_{i,t-s} + \boldsymbol{\epsilon}_{it}, i = 1, \ldots, N, \tag{10.2.6}$$

where $\Delta = (1 - L)$ denotes the first difference operation, $\Delta \mathbf{w}_{it} = \mathbf{w}_{it} - \mathbf{w}_{i,t-1}$, $\boldsymbol{\epsilon}_{it}$ is independent, identically distributed with mean $\mathbf{0}$ and covariance matrix $\Omega_{ii}$, $\Pi_i^* = -(I_m - \sum_{j=1}^{p_i} \Phi_{ij})$, $\Phi_{ij}^* = (-\sum_{s=j+1}^{p_i} \Phi_{is})$, $j = 1, 2, \ldots, p_i$, and

$$\text{rank}\,(\Pi_i^*) = r. \tag{10.2.7}$$

Decompose the $m \times m$ matrix $\Pi_i^*$ into the product of two rank $r (m \times r)$ matrices, $J_i$ and $\Lambda_i$,

$$\Pi_i^* = J_i \Lambda_i', \quad i = 1, \ldots, m. \tag{10.2.8}$$

The decomposition is not unique, $\Pi_i^* = J_i^* \Lambda_i^{*'}$, where $J_i^* = \Gamma_i F_i$, $\Lambda_i^* = \Lambda_i F_i^{'-1}$, for any $r \times r$ nonsingular matrix $F_i$. To uniquely define the cointegrating relations, we choose the normalization

$$\Lambda_i' = [I_r, \tilde{\Lambda}_i'], \tag{10.2.9}$$

where $\tilde{\Lambda}_i$ is an $(m - r) \times r$ constant matrix. The advantage of considering an error correction representation (10.2.6) rather than PVAR is that one can simultaneously consider the long-run (equilibrium) relations and short-run dynamics of $\mathbf{w}_{it}$.

If the stationary components $\boldsymbol{\xi}_{it}$ are cross-sectionally dependent, we can decompose $\boldsymbol{\xi}_{it}$ into the impact of $q$ common factors $\mathbf{f}_t$ that affect all cross-sectional units $\boldsymbol{\xi}_{it}$ and the idiosyncratic components $\boldsymbol{\epsilon}_{it}$,

$$\boldsymbol{\xi}_{it} = B_i \mathbf{f}_t + \boldsymbol{\epsilon}_{it}, \quad i = 1, \ldots, N, t = 1, \ldots T, \tag{10.2.10}$$

where $B_i$ is an $m \times q$ constant matrix, $E(\mathbf{f}_t) = \mathbf{0}$, $E(\mathbf{f}_t \mathbf{f}_t')$ is normalized to be a $q$-rowed identity matrix, $E(\boldsymbol{\epsilon}_{it}) = \mathbf{0}$, $E(\boldsymbol{\epsilon}_{it} \boldsymbol{\epsilon}_{it}') = D_i$ is a diagonal matrix.

If the nonstationarity of $\mathbf{w}_{it}, i = 1, \ldots, N$, is driven by the same common trends across $i$, that is, if $\mathbf{z}_{it} = \mathbf{z}_{jt} = \mathbf{z}_t$, it also implies that each element of $\mathbf{w}_{it}, w_{kt}$ is cointegrated across cross-sectional units. Let $\mathbf{w}_{kt}$ denote the $N \times 1$ vector of the $k$th element of $\mathbf{w}_{it}, w_{kit}, (w_{k1t}, \ldots, w_{kNt})' = \mathbf{w}_{kt}$, and $\boldsymbol{\xi}_{kt} = (\xi_{k1t}, \ldots, \xi_{kNt})', \boldsymbol{\epsilon}_{kt} = (\epsilon_{k1t}, \ldots, \epsilon_{kNt})'$. Then

$$\mathbf{w}_{kt} = A_k \mathbf{z}_t + \boldsymbol{\xi}_{kt}, k = 1, \ldots, m, \tag{10.2.11}$$

where $A_k = (\mathbf{a}_{ki}')$ denotes the $N \times (m - r)$ constant matrix of the cross-sectional stacked $k$th row of $A_i$. Suppose rank $(A_k) = d_k (\leq m - r)$, there also exists an $(m - d_k) \times m$ matrix $C_k$ with rank $(m - d_k)$ such that

$$C_k A_k = \mathbf{0}, \quad k = 1, \ldots, m. \tag{10.2.12}$$

Then

$$C_k \mathbf{w}_{kt} = C_k \boldsymbol{\xi}_{kt} \sim I(0). \tag{10.2.13}$$

In other words, with nonstationary panel data, there could be cointegration relations both over time and across individuals.

### 10.2.2 Estimation

*10.2.2.1 "Homogenous" Cointegrating Relations*

In the case when there is at most one cointegration relation among $I(1)$ $\mathbf{w}_{it}$, $i = 1, \ldots, N$, the "homogenous" cointegration vector between $(w_{1it}, \tilde{\mathbf{w}}'_{it})$, $(1, \boldsymbol{\beta}')$ with individual-specific effects $\alpha_{1i}$ yields

$$w_{1it} = \tilde{\mathbf{w}}'_{it}\boldsymbol{\beta} + \alpha_{1i} + u_{1it}, \tag{10.2.14}$$

where $u_{1it}$ is stationary but independently distributed across $i$. The cointegrating vector $\boldsymbol{\beta}$ can be estimated by the within estimator,

$$\begin{aligned}\hat{\boldsymbol{\beta}} = &\left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{\mathbf{w}}_{it} - \bar{\tilde{\mathbf{w}}}_i)(\tilde{\mathbf{w}}_{it} - \bar{\tilde{\mathbf{w}}}_i)'\right]^{-1} \\ &\cdot \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{\mathbf{w}}_{it} - \bar{\tilde{\mathbf{w}}}_i)(w_{1it} - \bar{w}_{1i})\right],\end{aligned} \tag{10.2.15}$$

where $\bar{\mathbf{w}}'_i = (\bar{w}_{1i}, \bar{\tilde{\mathbf{w}}}'_i) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}'_{it}$. The least-squares estimator converges to $\boldsymbol{\beta}$ in the speed of $T\sqrt{N}$ and is asymptotically normally distributed. However, the endogeneity and unit-root of $\tilde{\mathbf{w}}_{it}$ leads to a nonzero asymptotic bias term of order $\frac{1}{T}$ when $\hat{\boldsymbol{\beta}}$ is multiplied by the scale factor $T\sqrt{N}$ (Kao and Chiang 2000). The idea of Phillips and Hansen (1990) fully modified estimator can be applied to correct the endogeneity effect,

$$w^{+}_{1it} = \tilde{\mathbf{w}}'_{it}\boldsymbol{\beta} + \alpha_{1i} + u^{+}_{1it} \tag{10.2.16}$$

where

$$w^{+}_{1it} = w_{1it} - \Omega_{u\Delta\tilde{w}}\Omega^{-1}_{\Delta\tilde{w}}\Delta\tilde{\mathbf{w}}_{it}, \tag{10.2.17}$$

$$u^{+}_{1it} = u_{1it} - \Omega_{u\Delta\tilde{w}}\Omega^{-1}_{\Delta\tilde{w}}\Delta\tilde{\mathbf{w}}_{it}, \tag{10.2.18}$$

and $\Omega_{u\Delta\tilde{w}} = \sum_{j=-\infty}^{\infty}E(u_{it}\Delta\tilde{\mathbf{w}}'_{i,t-j})$, $\Omega_{\Delta\tilde{w}} = \sum_{j=-\infty}^{\infty}E(\Delta\tilde{\mathbf{w}}_{it}\Delta\tilde{\mathbf{w}}'_{i,t-j})$, are the long-run covariance matrices of $u_{1it}$ and $\Delta\tilde{\mathbf{w}}_{it}$, and $\Delta\tilde{\mathbf{w}}_{it}$, respectively. The panel fully modified within estimator takes the form,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{FM} = &\left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{\mathbf{w}}_{it} - \bar{\tilde{\mathbf{w}}}_i)(\tilde{\mathbf{w}}_{it} - \bar{\tilde{\mathbf{w}}}_i)'\right]^{-1} \\ &\cdot \left[\sum_{i=1}^{N}\left(\sum_{t=1}^{T}(\tilde{\mathbf{w}}_{it} - \bar{\tilde{\mathbf{w}}}_i)w^{+}_{1it} - T\Delta_{\Delta\tilde{w}u}\right)\right],\end{aligned} \tag{10.2.19}$$

where $\Delta_{\Delta\tilde{w}u} = \sum_{j=0}^{\infty} E(\Delta\tilde{\mathbf{w}}_{i,t-j}u_{1it})$. The correction terms $\Omega_{u\Delta\tilde{w}}, \Omega_{\Delta\tilde{w}}$, $\Delta_{\Delta\tilde{w}u}$ can be replaced by their consistent estimator.[5] Kao and Chiang (2000) show that $T\sqrt{N}(\hat{\boldsymbol{\beta}}_{FM} - \boldsymbol{\beta})$ is asymptotically normally distributed with mean $\mathbf{0}$ and covariance matrix $2\sigma^2_{u|\Delta\tilde{w}}\Omega^{-1}_{\Delta\tilde{w}}$, where $\sigma^2_{u|\Delta\tilde{w}} = \Omega_u - \Omega_{u\Delta\tilde{w}}\Omega^{-1}_{\Delta\tilde{w}}\Omega_{\Delta\tilde{w}u}$, $\Omega_u = \sum_{j=-\infty}^{\infty} E(u_{1it}u_{1i,t-j})$.

An alternative approach is to apply the dyanmic within estimator to the lead-lag adjusted regression model (Saikkonen 1991),

$$w_{1it} = \tilde{\mathbf{w}}'_{it}\boldsymbol{\beta} + \sum_{j=-q}^{q} \Delta\tilde{\mathbf{w}}'_{i,t-j}\boldsymbol{\gamma}_j + \alpha_{1i} + \tilde{u}_{1it} \tag{10.2.20}$$

where Westerlund (2005) suggests a data-based choice of truncation lag order $q$. Kao and Chiang (2000) show that the within estimator of (10.2.20) has the same asymptotic distribution as the panel fully modified within estimator. The Monte Carlo studies conduced by Kao and Chiang (2000) show that the (lead-lag adjusted) within estimator of (10.2.20) performs better than the panel fully modified estimator, probably because of the failure of obtaining good esimates of $\Omega_{u\Delta\tilde{\mathbf{w}}}$, $\Omega_{\Delta\tilde{w}}$, etc.

When there are more than one linearly independent cointegration relations among the elements of $\mathbf{w}_{it}$, in principle, one can normalize the $r$ linearly independent cointegration relation in the form of (10.2.9), $\Lambda' = [I_r \ \tilde{\Lambda}']$, then apply the fully modified within estimator or the lead-lag adjusted within estimator equation by equation.

Alternatively, one can use the methods of estimating the fixed effects PVAR model discussed in Section10.1 to obtain the consistent estimator of $\Phi_j$'s, then solving $\hat{\Lambda}$ from the estimated $\hat{\Phi}_j$'s. The error-correction representation of cointegrated system (10.1.1) gives $\Pi = -(I_m - \sum_{j=1}^{P} \Phi_j)$ (10.1.11). If rank $(\Pi) = r(> 1)$, then we can write $\Pi = J\Lambda'$ where $J$ and $\Lambda$ are $m \times r$ with rank $r$. Then the cointegrating matrix $\Lambda'$ is equal to

$$\Lambda' = (J'J)^{-1}J'\Pi. \tag{10.2.21}$$

If we choose the normalization $\Lambda' = [I_r \ \tilde{\Lambda}']$, then $J = \Pi_1$ and

$$\tilde{\Lambda}' = (\Pi'_1\Pi_1)^{-1}\Pi'_1\Pi \tag{10.2.22}$$

where $\Pi_1$ is the $m \times r$ matrix consisting of the first $r$ columns of $\Pi$.

Although the above procedure is consistent, it is not efficient because the first-stage estimator of $\Phi$'s has not taken into account the reduced rank restrictions on $\Pi = J\Lambda'$. One way to obtain efficient estimator of $\Phi_j$'s is to apply constrained GMM or constrained MDE or constrained (transformed) MLE by

[5] If cross-sectional units have heteroscedastic long-run variance, Kao and Chiang (2000) suggest using the cross-sectional average in the long-run correction terms.

minimizing the quadratic form of the moment conditions (10.1.15) or (10.1.30) or maximizing (10.1.28) subject to

$$\Lambda = H\boldsymbol{\delta}_r + B_r, \tag{10.2.23}$$

where $H$ and $B_r$ are, respectively, $m \times (m-r)$ and $m \times r$ matrices with known elements, and $\boldsymbol{\delta}$ is a $(m-r) \times r$ matrix with unknown coefficients. In the case when $\Lambda' = [I_r \ \tilde{\Lambda}']$, $H = [0, I_{m-r}]$, $\boldsymbol{\delta}_r = \tilde{\Lambda}$, $B_r = [I_r, \mathbf{0}]'$.

*10.2.2.2 Heterogeneous Cointegrating System*

If individual units are independent across $i$ ($B_i = \mathbf{0}$ in (10.2.10)), the Johansen (1991) method can be applied to the time series data for each $i$ to obtain the maximum likelihood estimator (MLE) of (10.2.6). There is no need for pooling.

When $B_i \neq \mathbf{0}$, individual units are correlated. If $N$ is fixed and $T$ is large the covariance between $\boldsymbol{\epsilon}_{it}$ and $\boldsymbol{\epsilon}_{jt}$, $\Omega_{ij}$, can simply be estimated by

$$\hat{\Omega}_{ij} = \frac{1}{T}\sum_{t=1}^{T} \hat{\boldsymbol{\epsilon}}_{it}\hat{\boldsymbol{\epsilon}}'_{jt}, \tag{10.2.24}$$

where $\hat{\boldsymbol{\epsilon}}_{it}$ can be constructed from (10.2.6) using the Johansen MLE estimates of $\boldsymbol{\theta}'_i = \text{vec}\,(\Pi^*_i, \Phi^*_{i1}, \ldots, \Phi^*_{ip_i})'$, where vec $(\cdot)$ denotes the operator that stacks the columns of a matrix successively into a vector. Therefore, we shall not be concerned with the estimation of $B_i$, but only with the inference of $\boldsymbol{\theta}_i$.

For ease of exposition, we shall first assume that all $\Phi_{is} \equiv \mathbf{0}$, for $s \geq 2$; then

$$\Delta\mathbf{w}_{it} = \Pi^*_i\mathbf{w}_{i,t-1} + \boldsymbol{\epsilon}_{it}, i = 1, \ldots, N. \tag{10.2.25}$$

Let $\mathbf{w}_t = (\mathbf{w}'_{1t}, \ldots, \mathbf{w}'_{Nt})'$ be the $Nm \times 1$ vector that stacks the $N$ cross-sectionally observed $\mathbf{w}_{it}$ one after another. Then (10.2.25) can be written as

$$\Delta\mathbf{w}_t = \tilde{\Pi}\mathbf{w}_{t-1} + \boldsymbol{\epsilon}_t, \tag{10.2.26}$$

where

$$\tilde{\Pi} = \begin{pmatrix} \Pi^*_1 & \mathbf{0} & . & . & \mathbf{0} \\ \mathbf{0} & \Pi^*_2 & . & . & \mathbf{0} \\ . & & . & & \\ . & & & . & \\ \mathbf{0} & & & & \Pi^*_N \end{pmatrix}, \tag{10.2.27}$$

and

$$\boldsymbol{\epsilon}_t = (\boldsymbol{\epsilon}'_{1t}, \ldots, \boldsymbol{\epsilon}'_{Nt})'.$$

Under the assumption that $\boldsymbol{\epsilon}_t$ is independently normally distributed with mean $\mathbf{0}$ and covariance matrix

$$\underset{Nm \times Nm}{\tilde{\Omega}} = \begin{pmatrix} \Omega_{11} & \Omega_{12} & . & . & \Omega_{1N} \\ . & \Omega_{22} & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ \Omega_{N1} & . & . & . & \Omega_{NN} \end{pmatrix}, \tag{10.2.28}$$

the log-likelihood function of $\Delta \mathbf{w}_t$ is proportional to

$$-\frac{T}{2} \log \mid \tilde{\Omega} \mid -\frac{1}{2} tr \left[ \Omega^{-1} \sum_{t=1}^{T} (\Delta \mathbf{w}_t - \tilde{\Pi} \mathbf{w}_{t-1})(\Delta \mathbf{w}_t - \tilde{\Pi} \mathbf{w}_{t-1})' \right]. \tag{10.2.29}$$

The MLE of $\Omega$ and $\Pi$ are the solutions that simultaneously satisfy

$$\hat{\tilde{\Omega}} = \frac{1}{T} \sum_{t=1}^{T} (\Delta \mathbf{w}_t - \hat{\tilde{\Pi}} \mathbf{w}_{t-1})(\Delta \mathbf{w}_t - \hat{\tilde{\Pi}} \mathbf{w}_{t-1})', \tag{10.2.30}$$

and

$$\begin{pmatrix} \text{vec } (\hat{\Pi}_1^{*\prime}) \\ \text{vec } (\hat{\Pi}_2^{*\prime}) \\ . \\ . \\ \text{vec } (\hat{\Pi}_N^{*\prime}) \end{pmatrix} = \left( \sum_{t=1}^{T} \hat{\tilde{\Omega}}^{-1} \otimes \mathbf{w}_{t-1} \mathbf{w}_{t-1}' \right)^{-1} \left[ \sum_{t=1}^{T} \left( \hat{\tilde{\Omega}}^{-1} \otimes \mathbf{w}_{t-1} \right) \text{ vec } (\Delta \mathbf{w}_t) \right]. \tag{10.2.31}$$

Conditional on $\hat{\tilde{\Omega}}$, (10.2.31) is in the form of Zellner's (1962) seemingly unrelated regression estimator. Sequentially iterating between (10.2.30) and (10.2.31) until convergence will result in the MLE of $\Pi_i^*$ and $\tilde{\Omega}$.

However, if $\mathbf{w}_{it}$ are cointegrated, then $\Pi_i$ is subject to the restrictions of the form (10.2.8). Substituting (10.2.8) into (10.2.9) and making use of the relations (Magnus and Neudecker 1999, p. 31)

$$\begin{aligned} \text{vec } (J_i \Lambda_i') &= (I_r \otimes J_i) \text{ vec } (\Lambda_i') \\ &= (\Lambda_i \otimes I_m) \text{ vec } (J_i), \end{aligned} \tag{10.2.32}$$

we obtain the MLE of $(\text{vec } (\Lambda_1')', \ldots, \text{ vec } (\Lambda_N')')'$ conditional on $\hat{\tilde{\Omega}}$ and $\hat{J}_i, i = 1, \ldots, N$,

$$\begin{aligned} \begin{pmatrix} \text{vec } (\hat{\Lambda}_1) \\ . \\ . \\ \text{vec } (\hat{\Lambda}_N) \end{pmatrix} &= \left\{ \left[ \hat{Q}' \left[ \hat{\tilde{\Omega}}^{-1} \otimes \left( \sum_{t=1}^{T} \mathbf{w}_{t-1} \mathbf{w}_{t-1}' \right) \right] \hat{\tilde{Q}} \right]^{-1} \right\}^{-1} \\ &\quad \cdot \left\{ \hat{Q}' \left[ \hat{\tilde{\Omega}}^{-1} \otimes I_{Nm} \right] \text{ vec } \left( \sum_{t=1}^{T} \mathbf{w}_{t-1} \Delta \mathbf{w}_t' \right) \right\}, \end{aligned} \tag{10.2.33}$$

and the MLE of (vec $(\hat{J}_1')', \ldots,$ vec $(\hat{J}_N')')'$ conditional on $\hat{\Omega}$ and $\hat{\Lambda}_i, i = 1, \ldots, N$,

$$\begin{pmatrix} \text{vec}\,(\hat{J}_1') \\ \cdot \\ \cdot \\ \text{vec}\,(\hat{J}_N)' \end{pmatrix} = \left\{ \hat{P}' \left( \hat{\tilde{\Omega}}^{-1} \otimes \sum_{t=1}^{T} \mathbf{w}_{t-1}\mathbf{w}_{t-1}' \right) \hat{P} \right\}^{-1} \cdot \left\{ \hat{P}'(\hat{\tilde{\Omega}}^{-1} \otimes I_{Nm}) \text{ vec } \left( \sum_{t=1}^{T} \mathbf{w}_{t-1}\Delta\mathbf{w}_t' \right) \right\}, \tag{10.2.34}$$

where $I_{Nm}$ denotes the $Nm \times Nm$ identity matrix,

$$\hat{Q} = \left[ \left( \mathbf{d}_1 \otimes \hat{J}_1 \right) \otimes (\mathbf{d}_1 \otimes I_r), \ldots, \left( \mathbf{d}_N \otimes \hat{J}_N \right) \otimes (\mathbf{d}_N \otimes I_r) \right], \tag{10.2.35}$$

$$\hat{P} = \left[ (\mathbf{d}_1 \otimes I_m) \otimes \left( \mathbf{d}_1 \otimes \hat{\Lambda}_1 \right), \ldots, (\mathbf{d}_N \otimes I_m) \otimes \left( d_N \otimes \hat{\Lambda}_N \right) \right], \tag{10.2.36}$$

$d_j$ denotes the $j$th column of $I_N$, and $\otimes$ denotes the Kronecker product. Therefore Groen and Kleibergen (2003) suggest the following iterative scheme to obtain the panel MLE:

1. Construct initial estimates $\hat{\Omega}^{(0)}$ and $\hat{J}_i^{(0)}, i = 1, \ldots, N$ from (10.2.30) and (10.2.31), where the initial estimates of $\hat{J}_i^{(0)}$ are simply the first $r$ columns of $\hat{\Pi}_i^{*(0)}$ under the normalization (10.2.9).
2. Construct estimates of $\Lambda_i, i = 1, \ldots, N$ from (10.2.33).
3. Revise the estimate of $\Pi_i^*$ from (10.2.8) and revise the estimate of $\Omega$ from the revised estimate of $\hat{\Pi}_i^*$ using the formula (10.2.30).
4. Revise the estimate of $J_i$ given the revise estimator of $\Lambda_i$ and $\Omega$ using (10.2.34).
5. Iterate steps 1 to 4 until the solution converges.

The MLE of $\Lambda_i, i = 1, \ldots, N$ converges to their true values at rate $T$ and their limiting distributions are mixed normal. Therefore, conventional Wald-type test statistics on the null hypothesis of $\Lambda_i$ is asymptotically $\chi^2$ distributed.

When the cointegrating matrix $\Lambda_i = \Lambda_j = \Lambda$, the common cointegrating matrix can be estimated by

$$\text{vec}\,(\Lambda) = \left( \hat{\tilde{Q}}' \left( \hat{\Omega}^{-1} \otimes \sum_{t=1}^{T} \mathbf{w}_{t-1}\mathbf{w}_{t-1}' \right) \hat{\tilde{Q}} \right)^{-1} \cdot \left( \hat{\tilde{Q}}' \left( \hat{\Omega} \otimes I_{Nm} \right) \text{ vec } \left( \sum_{t=1}^{T} \mathbf{w}_{t-1}\Delta\mathbf{w}_t' \right) \right). \tag{10.2.37}$$

Because $\Lambda_i$ or $\Lambda$ is asymptotically mixed normally distributed, one can use the likelihood ratio test statistic to test the null hypothesis

$$\begin{aligned} & H_0 \ : \ \Lambda_1 = \Lambda_2 = \ldots = \Lambda_N = \Lambda \\ \text{versus} \quad & H_1 \ : \ \Lambda_i \neq \Lambda_j. \end{aligned} \tag{10.2.38}$$

The likelihood ratio statistic is asymptotically $\chi^2$ distributed with $(N-1)r(m-r)$ degrees of freedom when $T \longrightarrow \infty$.

When individual units contain an individual specific effects $\boldsymbol{\alpha}_i^*$ and $p_i$ in (10.2.6) different from 0, one can follow Johansen (1991) to concentrate out constants and $\Phi_{is}^*$ by regressing $\Delta \mathbf{w}_{it}$ and $\mathbf{w}_{i,t-1}$ on the constants and $\Delta \mathbf{w}_{i,t-s}$, $s = 1, \ldots, p_i$ respectively, to obtain $\Delta \tilde{\mathbf{w}}_{it}$ and $\tilde{\mathbf{w}}_{i,t-1}$, then proceed to estimate $\Gamma_i$ and $\Lambda_i$. Once the MLE of $\Gamma_i$, $\Lambda_i$ and $\Omega$ are obtained, the individual-specific constants and $\Phi_{ij}^*$ can then be obtained by regressing $\left(\Delta \tilde{\mathbf{w}} - \hat{J}_i \hat{\Lambda}_i' \tilde{\mathbf{w}}_{i,t-1}\right)$ on constants, $\Delta \tilde{\mathbf{w}}_{i,t-p_i-1}, \ldots, \Delta \tilde{\mathbf{w}}_{i,t-1}$ using Zellner's (1962) seemingly unrelated regression method.

When $N$ is large, it is not feasible to use either the likelihood approach or the Zellner (1962) unrelated regression framework to estimate either the homogeneous or the heterogeneous cointegration relations. Pesaran et al. (2004) suggest using a global VAR (GVAR) ((10.1.42)–(10.1.44)) to filter out cross-sectional dependence. When $w_{it}^*$ can be treated as weakly exogenous the system (10.1.42) subject to the rank condition (10.2.8) for each $i$ can be estimated using standard time series estimator techniques (e.g., Pesaran et al. 2000).

## 10.3 UNIT ROOT AND COINTEGRATION TESTS

### 10.3.1 Unit Root Tests

Panels have been used to analyze regional growth convergence (e.g., Bernard and Jones 1996, exchange rate determination (e.g., testing of purchasing power parity hypothesis; Frankel and Rose 1996), business cycle synchronization, etc. In such analysis the time series property of a variable is of significant interests to economists. The statistical properties of time series estimators also depend on whether the data is stationary or nonstationary. In the case of inference based on time series (i.e., $N = 1$), the limiting distributions of most estimators will be approximately normal when $T \longrightarrow \infty$ if the variables are stationary. Standard normal, $\chi^2$ tables can be used to construct confidence intervals or test hypotheses. If the data are nonstationary, or contain unit roots, standard estimators will have nonstandard distributions as $T \longrightarrow \infty$. The conventional Wald type test statistics cannot be approximated well by $t$- or $\chi^2$ distributions (e.g., Dickey and Fuller 1979, 1981; Phillips and Durlauf 1986). Computer simulations will have to be used to find the critical values under the null. In cases where $N$ is fixed and $T$ is large, standard time series techniques can be used and the panel aspect of the data does not pose new techniques. If $N$ is large, panel data provides the possibility of invoking a version of central limit

theorem along the cross-sectional dimension. Hence, contrary to the time series literature, the null hypothesis of panel unit root tests can be either unit root or stationarity. However, because most applications of panel unit root tests still follow the convention that the null of a time series of $y_{it}$ contain a unit root, we shall therefore discuss only tests based on the null of unit root. For reference of tests under the null of stationarity, see, for example, Kwiatkowski et al. (1992), Hadri (2000), and Hadri and Larsson (2005).

*10.3.1.1 Cross-Sectional Independent Data*

Since Quah (1994), many people have suggested panel unit root test statistics when $N$ and $T$ are large. When cross-sectional units are independent, following Dickey and Fuller (1979, 1981); Levin and Lin (1993) (LL); Levin, Lin, and Chu (LLC) (2002), consider a panel extension of the null hypothesis that each individual time series in the panel contains a unit root against the alternative hypothesis that all individual series are stationary by considering the model

$$\Delta y_{it} = \alpha_i + \delta_i t + \gamma_i y_{i,t-1} + \sum_{\ell=1}^{p_i} \phi_{i\ell} \Delta y_{i,t-\ell} + \epsilon_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{matrix} \tag{10.3.1}$$

where $\epsilon_{it}$ is assumed to be independently distributed across $i$ and $\Delta$ denotes the first difference operator, $1 - L$, with $L$ being the lag operator that shifts the observation by one period, $Ly_{it} = y_{i,t-1}$. If $\gamma_i = 0$, then $y_{it}$ contains a unit root. If $\gamma_i < 0$, $y_{it}$ is stationary. Levin and Lin (1993) specify the null hypothesis as

$$H_0 : \gamma_1 = \gamma_2 = \ldots = \gamma_N = 0, \tag{10.3.2}$$

and the alternative hypothesis as

$$H_1 : \gamma_1 = \gamma_2 = \ldots = \gamma_N = \gamma < 0. \tag{10.3.3}$$

To test $H_0$ against $H_1$, Levin and Lin (1993) suggest taking out the impact of variables in (10.3.1) that are not directly relevant to the estimation of $\gamma$ by first regressing $\Delta y_{it}$ and $y_{i,t-1}$ on the remaining variables in (10.3.1) for each $i$, providing the residuals $\hat{e}_{it}$ and $\hat{v}_{i,t-1}$, respectively. Then estimate $\gamma$ by running the regression of the following model

$$\hat{e}_{it} = \gamma \hat{v}_{i,t-1} + \epsilon_{it}. \tag{10.3.4}$$

To adjust for heteroscedasticity across $i$ in (10.3.4), they suggest using the least-squares estimate of $\gamma$, $\hat{\gamma}$, to compute the variance of $\hat{e}_{it}$,

$$\hat{\sigma}_{ei}^2 = (T - p_i - 1)^{-1} \sum_{t=p_i+2}^{T} (\hat{e}_{it} - \hat{\gamma} \hat{v}_{i,t-1})^2. \tag{10.3.5}$$

Then divide (10.3.4) by $\hat{\sigma}_{ei}$ for each $i$, to obtain the heteroscedasticity adjusted model

$$\tilde{e}_{it} = \gamma \tilde{v}_{i,t-1} + \tilde{\epsilon}_{it}, \tag{10.3.6}$$

where $\tilde{e}_{it} = \frac{\hat{e}_{it}}{\hat{\sigma}_{ei}}$, $\tilde{\epsilon}_{it} = \frac{\epsilon_{it}}{\hat{\sigma}_{ei}}$, $\tilde{v}_{i,t-1} = \frac{\hat{v}_{i,t-1}}{\hat{\sigma}_{ei}}$.

The $t$-statistic for testing $\gamma = 0$ is

$$t_{\tilde{\gamma}} = \frac{\tilde{\gamma}}{sd_{\tilde{\gamma}}}, \tag{10.3.7}$$

where $\tilde{\gamma}$ is the least squares estimates of (10.3.6),

$$sd_{\tilde{\gamma}} = \hat{\sigma}_{\epsilon} \left[ \sum_{i=1}^{N} \sum_{t=p_i+2}^{T} \tilde{v}_{i,t-1}^2 \right]^{-1/2}$$

$$\hat{\sigma}_{\epsilon}^2 = (N\tilde{T})^{-1} \sum_{i=1}^{N} \sum_{t=p_i+2}^{T} (\tilde{e}_{it} - \tilde{\gamma} \tilde{v}_{i,t-1})^2,$$

$$\bar{p} = \frac{1}{N} \sum_{i=1}^{N} p_i, \tilde{T} = (T - \bar{p} - 1).$$

Although regressing $e_{it}$ on $v_{it}$ over $t$ for given $i$ leads to a random variable with non-standard distributions as $T \longrightarrow \infty$ (e.g., Phillips and Durlauf 1986), averaging over $i$ allows the invocation of central limit theorem across cross-sectional dimension when $N$ is large. However, $t_{\tilde{\gamma}}$ is not centered at 0. To correct the asymptotic bias, Levin and Lin (1993) suggest adjusting (10.3.7) by

$$t^* = \frac{t_{\tilde{\gamma}} - N\tilde{T} S_{N,T} \hat{\sigma}_{\epsilon}^{-2} \cdot sd_{\tilde{\gamma}} \cdot \mu_{\tilde{T}}}{\sigma_{\tilde{T}}}, \tag{10.3.8}$$

where

$$S_{NT} = N^{-1} \sum_{i=1}^{N} \frac{\hat{\omega}_{yi}}{\hat{\sigma}_{ei}}, \tag{10.3.9}$$

and $\hat{\omega}_{yi}^2$ is an estimate of the long-run variance of $y_i$, say,

$$\hat{\omega}_{yi}^2 = (T-1)^{-1} \sum_{t=2}^{T} \Delta y_{it}^2 + 2 \sum_{L=1}^{\bar{K}} W_{\bar{K}}(L) \left( (T-1)^{-1} \sum_{t=L+2}^{T} \Delta y_{it} \Delta y_{i,t-L} \right), \tag{10.3.10}$$

where $W_{\bar{K}}(L)$ is the lag kernel to ensure the positivity of $\hat{\omega}_{yi}^2$; for instance, Newey and West (1987) suggest that

$$W_{\bar{K}}(L) = \begin{cases} 1 - \frac{L}{T} & \text{if } L < \bar{K}, \\ 0 & \text{if } L \geq \bar{K}. \end{cases} \tag{10.3.11}$$

The $\mu_{\bar{T}}$ and $\sigma_{\bar{T}}$ are mean and standard deviation adjustment terms, which are computed by Monte Carlo simulation and tabulated in their paper. They show that provided the augment Dickey–Fuller (1981) lag order $p$ increases at some rate $T^p$ where $0 \leq p \leq 1/4$ and the lag truncation parameter $\bar{K}$ increases at rate $T^q$ where $0 < q < 1$, the panel test statistic $t_{\bar{\gamma}}$ under the null of $\gamma = 0$ converges to a standard normal distribution as $T, N \to \infty$.

In the special case that $\alpha_i = \delta_i = \phi_{i\ell} = 0$, and $\epsilon_{it}$ is independently, identically distributed with mean 0 and variance $\sigma_\epsilon^2$, Levin and Lin (1993), Levin et al. (2002) show that under the null of $\gamma = 0$, $T\sqrt{N}\hat{\gamma}$ of the pooled least-squares estimator, $\hat{\gamma}$, converges to a normal distribution with mean 0 and variance 2 and the $t$ statistic of $\hat{\gamma}$ converges to a standard normal as $\sqrt{N}/T \to 0$ as $N, T \to \infty$ (i.e., the time dimension can expand more slowly than the cross section). When $T$ is fixed and $N \to \infty$, $\sqrt{N}\hat{\gamma}$ is asymptotically biased. For the correction of asymptotic bias of the $t$-statistics, see Harris and Tzaralis (1999).

Im, Pesaran, and Shin (2003) (IPS) relax the LL, LLC strong assumption of homogeneity for (10.3.1) under the alternative (i.e., allowing $\gamma_i \neq \gamma_j$) by postulating the alternative hypothesis as

$$H_A^* : \gamma_i < 0 \text{ for some } i, \text{ and } \frac{N_0}{N} = c > 0, \tag{10.3.12}$$

where $N_0$ denotes the number of cross-sectional units with $\gamma_i < 0$.[6] Thus, instead of pooling the data, Im et al. suggest taking the average of separate unit root tests for $N$ individual cross-sectional units of the augmented Dickey–Fuller (ADF) (Dickey and Fuller 1981) $t$-ratios $\tau_i, \bar{\tau}$. Because $\tau_i$ are independent across $i$, $\bar{\tau}$ converges to a normal distribution under the null with mean $E(\bar{\tau})$ and variance, Var $(\bar{\tau}_N)$, as $T \longrightarrow \infty$ and $N \longrightarrow \infty$. However, the statistic $\bar{\tau}/\sqrt{\text{Var}\,(\bar{\tau})}$ is equivalent to multiplying each $\tau_i$ by $\sqrt{N}$. Although the limiting distribution of each ADF statistic as $T \longrightarrow \infty$ is well defined, multiplying $\tau_i$ by $\sqrt{N}$ introduces a nonnegligble bias term as $N$ also goes to $\infty$. Therefore, IPS suggest using the statistics

$$Z = \frac{\sqrt{N}(\bar{\tau} - E(\tau_i))}{\sqrt{\text{Var}\,(\tau_i)}}. \tag{10.3.13}$$

Because $E(\tau_i)$ and Var$(\tau_i)$ will vary as the lag length in the ADF regression varies, Im et al. (2003) tabulate $E(\tau_i)$ and var $(\tau_i)$ for different lag lengths. They show in their Monte Carlos studies that their test is more powerful than

[6] Technically speaking, the alternative can be formulated for at least with one $i$, $\gamma_i < 0$. However, the test will have no power if $c \to$ as $N \to \infty$.

the Levin et al. test under certain cases. However, if the null is rejected, all we can say is that a fraction of cross-sectional units is stationary. It does not provide explicit guidance as to the size of this fraction or the identity of cross-sectional units that are stationary.

Alternatively, Maddala and Wu (1999) (MW) and Choi (2001) suggest using Fisher (1932) $P_\lambda$ to test the null (10.3.2) against (10.3.12) by combining the evidence from several independent tests. The idea is as follows: Suppose there are $N$ unit root tests as in Im et al. (2003). Let $P_i$ be the observed significance level ($P$-value) for the $i$th test. Then

$$P_\lambda = -2 \log P_i, \tag{10.3.14}$$

has a $\chi^2$ distribution with $2N$ degrees of freedom as $T_i \to \infty$ (Rao (1952, p. 44)). When $N$ is large, Choi (2001) proposes a modified $P_\lambda$ test,

$$P_m = \frac{\frac{1}{\sqrt{N}} \sum_{i=1}^{N} (-2 \log P_i - 2)}{2}, \tag{10.3.15}$$

because $E(-2 \log P_i) = 2$ and Var $(-2 \log P_i) = 4$. Using a sequential limit argument ($T_i \to \infty$ followed by $N \to \infty$), Choi (2001) shows that the $P_m$ test is asymptotically normally distributed with mean 0 and variance 1.

The LL test is based on homogeneity of the autoregressive parameter (although it allows heterogeneity in the error variances and the serial correlation structure of the errors). Thus the test is based on pooled regressions. On the other hand, both the MW (or Choi) test or the IPS test are based on the heterogeneity of the autoregressive parameter under the alternative. The tests amount to a combination of different independent tests. However, they are all consistent tests against either alternative. The advantage of the MW (or Choi) test is that it does not require a balanced panel, nor the identical lag length in the individual ADF regressions. In fact, it can be carried out for any unit root test derived. It is nonparametric. Whatever test statistic we use for testing for a unit root for each individual unit, we can get the $P$-values, $P_i$. The disadvantage is that the $P$-values have to be derived by Monte Carlo simulation. On the other hand, the LL and the IPS tests are parametric. Although the distribution of the $t_{\tilde{\gamma}}$ or $\bar{\tau}$ statistic involves the adjustment of the mean and variance, they are easy to use because ready tables are available from their papers. However, these tables are valid only for the ADF test.

The heterogeneity of panel data introduces an asymmetry in the way the null and the alternative hypotheses are treated, which is not present in the univariate time series models. This is because the same null hypothesis is imposed across $i$ but the alternative could be either homogeneous across $i$ or be allowed to vary with $i$. The IPS and MW tests allow the alternative to be heterogeneous. The drawback of their approach is that the model is overly parameterized, in the sense that the information contained in the unit-specific $t$-statistics is not used in an efficient way. The tests will have power only if $\frac{\sqrt{N}}{T} = c < \infty$. Because the interest is only in whether the null holds or not, Westerlund and Larsson

(2012) consider a random specification of $\rho_i = 1 + \gamma_i$ where $\rho_i$ is assumed independently identically distributed with mean $\mu_\rho$ and variance $\sigma_\rho^2$. Under the null of unit root,

$$H_0 : \mu_\rho = 1 \text{ and } \sigma_\rho^2 = 0 \tag{10.3.16}$$

The alternative is

$$\begin{aligned} H_1 &: \mu_\rho \neq 1 \\ &\text{or } \sigma_\rho^2 > 0 \\ &\text{or both.} \end{aligned} \tag{10.3.17}$$

The Lagrangian multiplier (LM) test statistics of the null (10.3.16) takes the form

$$LM_{\mu_\rho,\sigma_\rho^2} = LM_{\mu_\rho|\sigma_\rho^2} + LM_{\sigma_\rho^2|\mu_\rho}, \tag{10.3.18}$$

where

$$LM_{\mu_\rho|\sigma_\rho} = \frac{(A_{NT})^2}{B_{NT}} + \frac{1}{2}\frac{(C_{NT})^2}{D_{NT}}, \tag{10.3.19}$$

$$A_{NT} = \frac{1}{\sqrt{N}T}\sum_{i=1}^{N}\sum_{t=p+1}^{T}\Delta e_{it} v_{i,t-1}, \tag{10.3.20}$$

$$B_{NT} = \frac{1}{NT^2}\sum_{i=1}^{N}\sum_{t=p+1}^{T} v_{i,t-1}^2, \tag{10.3.21}$$

$$C_{NT} = \frac{1}{\sqrt{N}T^{3/2}}\sum_{i=1}^{N}\sum_{t=p+1}^{T}\left[(\Delta e_{it})^2 - 1\right] v_{i,t-1}^2, \tag{10.3.22}$$

$$D_{NT} = \frac{1}{NT^3}\sum_{i=1}^{N}\sum_{t=p+1}^{T}(\Delta e_{it})^2 v_{i,t-1}^4. \tag{10.3.23}$$

The $LM_{\mu_\rho|\sigma_\rho^2}$ can be viewed as LM test statistic for testing $\mu_\rho = 1$ versus $\mu_\rho \neq 1$ given $\sigma_\rho^2 = 0$. The $LM_{\sigma_\rho^2|\mu_\rho}$ is the LM test statistic for testing $\sigma_\rho^2 = 0$ versus $\sigma_\rho^2 > 0$ given $\mu_e = 1$. Westerlund and Larsson (2012) show that when both $N$ and $T \to \infty$ and $\frac{T}{N} = N^{\theta-1}$ for $\theta > 1$, under the null (10.3.16), $LM_{\sigma_\rho^2|\mu_\rho}$ is $\chi^2$ distributed with one degree of freedom. The distribution of $LM_{\sigma_\rho^2|\mu_\rho}$ is asymptotically independent of the limiting distribution of $LM_{\mu_\rho|\sigma_\rho^2}$ and is asymptotically equal to $\frac{5(\kappa-1)}{24}$ times a chi-square degree 1 variable, where $\kappa = 1 + \frac{1}{T}\sum_{t=p+1}^{T}[E(\epsilon_{it}^4) - 1]$. If $\epsilon_{it} \sim N(0,1)$, then $\kappa = 3$ and $\frac{12}{5}LM_{\sigma_\rho^2|\mu_e}$ is asymptotically $\chi^2$ distributed with one degree of freedom.

When $e_{it}$ and $v_{i,t-1}$ are unknown one can construct the feasible LM test using $\hat{e}_{it}$ and $\hat{v}_{i,t-1}$ as in LL test. The feasible LM is asymptotically equal to LM test. (i.e., $\delta_i = 0$ in (10.3.1)).

Assuming the $\rho_i$ random has the advantage that the number of parameters needed to be estimated is reduced. Neither does it rule out that under the alternative, some of the units may be explosive. Moreover, by considering not only the mean of $\rho_i$, but also the variance, the random coefficient formulation takes account more information, hence is more powerful in detecting the alternative than LL or IPS test.

In large $N$ or small $T$ case, it only makes sense to discuss powerful unit root tests that are informative if either all cross-sectional units have the same dynamic response or a significant fraction of cross-sectional units reject the null under heterogeneity. However, to identify the exact proportion of the sample for which the null hypothesis is rejected requires $T$ being very large. (For further discussion, see Pesaran 2012).

### *10.3.1.2 Cross-Sectionally Correlated Data*

When $v_{it}$ are cross-sectionally dependent, the preceding tests are subject to severe size distortion (e.g., Banerjee, Marcellino, and Osbat 2005; Breitung and Das (2008). Chang (2002) suggests using some nonlinear transformation of the lagged level variable, $y_{i,t-1}$, $F(y_{i,t-1})$ as an instrument (IV) for $y_{i,t-1}$ for the usual Dickey–Fuller type regression (10.3.1). As long as $F(\cdot)$ is regularly integreable, say $F(y_{i,t-1}) = y_{i,t-1}e^{-c_i|y_{i,t-1}|}$, where $c_i$ is a positive constant, the IV $t$-ratio

$$Z_i = \frac{\hat{\gamma}_i - 1}{s(\hat{\gamma}_i)}, \tag{10.3.24}$$

will converge to a standard normal when $T_i \longrightarrow \infty$ where $s(\hat{\gamma}_i)$ is the standard error of the IV estimator $\hat{\gamma}_i$ and $T_i$ denotes the time series observation of the $i$th unit because $F(y_{i,t-1})$ tends to 0 and $y_{i,t-1}$ tends to $\pm\infty$, the nonstationarity of $y_{i,t-1}$ is eliminated. Moreover, the product of $F(y_{i,t-1})$ and $F(y_{j,t-1})$ from different cross-sectional units $i$ and $j$ are asymptotically uncorrelated even though $y_{i,t-1}$ and $y_{j,t-1}$ are correlated; therefore the average IV $t$-ratio statistic

$$S_N = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} Z_i \tag{10.3.25}$$

possesses a standard normally limiting distribution as $N$ also tends to infinity.

When $y_{it}$ are correlated across cross-sectional units, Bai and Ng (2004, 2010) and Moon and Perron (2004) consider that the cross-correlations are driven by some common factors, $\mathbf{f}_t' = (f_{1t}, \ldots, f_{kt})$ that vary over time. Rewrite $y_{it}$ as the sum of the impact of the common factors, $\mathbf{b}_i' \mathbf{f}_t$ and the idiosyncratic component, $u_{it}$,

$$y_{it} = \sum_{j=1}^{k} b_{ij} f_{jt} + u_{it}, \tag{10.3.26}$$

where $\mathbf{b}_i' = (b_{i1}, \ldots, b_{ik})$ is a $k \times 1$ vector of constants, $u_{it}$ is independent across $i$ with mean 0. If $\mathbf{b}_i = \mathbf{b}_j = \mathbf{b}$ then $\lambda_t = \mathbf{b}' \mathbf{f}_t$ is the common time-specific effects. The impact of common factors, $\mathbf{f}_t$, can be eliminated by subtracting the cross-sectional mean $y_t = \frac{1}{N} \sum_{i=1}^{N} y_{it}$ from $y_{it}$. If $N$ is large, the demeaned series can be treated as if they are uncorrelated. We shall therefore consider the case that $\mathbf{b}_i \neq \mathbf{b}_j$.

Bai and Ng (2004, 2010) and Moon and Perron (2004) propose testing unit roots in both $\mathbf{f}_t$ and $u_{it}$. The factors are estimated from $k$ principal components of $\Delta y_{it}$ (e.g., Chapter 9, Section 9.3). Augmented Dickey–Fuller tests (ADF) are then applied to test if $\mathbf{f}_t$ is integrated or not (I(1) versus I(0)). If it is found that the estimated factors contain unit roots and are not cointegrated, the $N$ series are integrated of order 1. If the presence of a unit root in the factors is rejected, panel unit root tests such as the Maddala and Wu (1999) $P_\lambda$ or Choi (2001) $P_m$ test is then applied to the defactored observations, $u_{it}$, $(y_{it} - \hat{\mathbf{b}}_i' \hat{\mathbf{f}}_t)$.[7]

If $\mathbf{b}_i \neq \mathbf{b}_j$ and $k = 1$, Pesaran (2007) suggests augmenting (10.3.1) by the cross-sectional averaged values of $\bar{y}_{t-1} = \frac{1}{N} \sum_{i=1}^{N} y_{i,t-1}$ and $\Delta \bar{y}_{t-j} = \frac{1}{N} \sum_{i=1}^{N} \Delta y_{i,t-j}$,

$$\begin{aligned} \Delta y_{it} = \alpha_i + \delta_i t + \gamma_i y_{i,t-1} + \sum_{\ell=1}^{p_i} \phi_{i\ell} \Delta y_{i,t-\ell} \\ + c_i \bar{y}_{t-1} + \sum_{\ell=1}^{p_i} d_{i\ell} \Delta \bar{y}_{t-\ell} + e_{it}, \end{aligned} \tag{10.3.27}$$

to filter out the cross-sectional dependency, then taking the average of separate unit root tests for $N$ individual cross-sectional units of the augmented Dickey–Fuller $t$-ratios, $\tau_i$, $\bar{\tau}$, as in Im et al. (2003).

When $k > 1$, if there exist $K (\geq k - 1)$ observed time series $\mathbf{w}_{it}$, Kapetanios, Pesaran, and Yamagata (2011) and Pesaran, Smith, and Yamagata (2013) suggest further augmenting (10.3.1) by cross-sectional mean of $\mathbf{w}_t$

$$\begin{aligned} \Delta y_{it} = \alpha_i + \delta_i t + \gamma_i y_{i,t-1} + \sum_{\ell=1}^{p_i} \Delta y_{i,t-\ell} \\ + \sum_{\ell=1}^{p_i} d_{i\ell} \Delta \bar{y}_{t-\ell} \\ + W_t \boldsymbol{\theta} + \epsilon_{it}, \quad i = 1, \ldots, N \end{aligned} \tag{10.3.28}$$

where $W_t = (\Delta \mathbf{w}_t', \Delta \mathbf{w}_{t-1}', \ldots, \Delta \mathbf{w}_{t-p_i-1}', \mathbf{w}_{t-1}')$ and $\boldsymbol{\theta}$ is an $mp_i \times 1$ vector of constants. Pesaran, Smith, and Yamagata (2013) suggest either constructing an

[7] This approach is called PANIC (panel analysis of nonstationarity in idiosyncratic and common components) by Bai and Ng (2004, 2010).

IPS type test statistic by taking the average of $t$-statistics for $\gamma_i, i = 1, \ldots, N$ or taking the average of the Sargan and Bhargava test statistic (SB) (1983),

$$SB_i = T^{-2} \sum_{t=1}^{T} S_{it}^2 / \hat{\sigma}_i^2, \quad i = 1, \ldots, \tag{10.3.29}$$

where

$$S_{it} = \sum_{s=1}^{t} \hat{\epsilon}_{is}, \hat{\sigma}_i^2 = \frac{\sum_{t=1}^{T} \hat{\epsilon}_{it}^2}{T - K^*},$$

$\hat{\epsilon}_{it}$ is the least-squares residual of (10.3.28), and $K^*$ is the number of unknown constants in (10.3.28). Each of the $\mathrm{SB_i}$ converges to a functional of Brownian motion which is independent of the factors as well as their loadings. Pesaran et al. (2013) have provided the critical values of the augmented IPS and SB tests for $K = 0, 1, 2, 3$ and $p_i = 0, 1, \ldots, 4$ for $N, T = 20, 30, 50, 70, 100, 200$.

When the process of driving cross-sectional dependence is unknown, Choi and Chue (2007) suggest using a subsample testing procedure. They suggest to group the sample into a number of overlapping blocks of $b$ time periods. Using all $(T - b + 1)$ possible overlapping blocks, the critical value of a test statistic is estimated by the empirical distribution of the $(T - b + 1)$ test statistics computed. Although the null distribution of the test statistic may depend on the unknown nuisance parameters when $T \longrightarrow \infty$ and $N$ fixed, the subsample critical values will converge in probability to the true critical values. The Monte Carlo simulation conducted by Choi and Chue (2007) show that the size of the subsample test is indeed robust against various forms of cross-sectional dependence.

All the unit root tests discussed above assume there is no structural breaks. Structural changes are likely to happen over long horizons. Failing to consider the presence of structural breaks may lead to misleading conclusions about the order of integration of a time series. For instance, it is well known that a stationary time series that evolves around a broken trend might be regarded as a nonstationary process (e.g., Perron 1989). For panel unit root tests allowing structural break with cross-sectional independent data, see, e.g. Im, Lee, and Tieslau (2005), with cross-sectional dependent data, see, e.g. Bai and Carrion-i-Silvestre (2009).

### 10.3.2 Tests of Cointegration

#### *10.3.2.1 Residual-Based Tests*

One can test the existence of cointegration relationships by testing if the residuals of regressing one element of $\mathbf{w}_{it}$ on the rest elements of $\mathbf{w}_{it}$ have a unit root or not. For instance, if the residuals of (10.2.14) contain a unit root, $\mathbf{w}_{it}$ are not cointegrated. If the residuals are stationary, $\mathbf{w}_{it}$ are cointegrated. In a time series

framework, the null of time series unit root tests is unit root. In other words, the null is no cointegration. Under the null, time series regression of $\boldsymbol{\beta}$ fails to converge (Phillips (1986)). On the other hand, panels with large $N$ and large $T$ can yield convergent estimate of $\boldsymbol{\beta}$. If there is no cointegration, Kao (1999), using a sequential limit argument (first $T \to \infty$, followed by $N \to \infty$), has shown that the least-squares estimator of (10.2.14) converges to $\Omega_{\Delta\tilde{w}}^{-1}\Omega_{\Delta\tilde{w}u}$. If there is cointegration, then the panel fully modified within estimator (10.2.19) or the within estimator for the lead-lag adjusted regression model (10.2.20) is consistent and asymptotically normally distributed. Moreover, because $u_{it}$ is independently distributed across $i$, a version of the central limit theorem can be invoked on the cross-sectional average of a statistic. That is, even though a statistic could have different asymptotic properties depending on whether $u_{it}$ contains a unit root or not along time series dimension, the properly scaled cross-sectional average of such a statistic is asymptotically normally distributed as $N \to \infty$. What this implies is that the panel null for the distribution of $u_{it}$ could either be unit root (no cointegration) or stationary (cointegration).

McCoskey and Kao (1998) propose to test the null of cointegration using the statistic

$$\overline{\mathrm{LM}} = \frac{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T^2}\sum_{t=1}^{T} S_{it}^{+2}}{s^{+2}}, \tag{10.3.30}$$

where

$$S_{it}^{+} = \sum_{s=1}^{t} \hat{u}_{1is},\ s^{+2} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \hat{u}_{1it}^{+2},$$

$\hat{u}_{1it}$ are the estimated residuals of (10.2.14) with $\boldsymbol{\beta}$ estimated by either (10.2.19) or applying the within estimator to (10.2.20) and $\hat{u}_{1it}^{+}$ is the estimated residual of (10.2.18). McCoskey and Kao (1998) show that $\sqrt{N}(\overline{\mathrm{LM}})$ is asymptotically normally distributed with mean $\mu$ and variance $\sigma_v^2$, if individual units are independent of each other where $\mu$ and $\sigma_v^2$ can be found through Monte Carlo simulations.

Kao (1999) and Pedroni (2004) propose test cointegration under the null of no cointegration (residuals contain unit root). Kao proposes modified Dickey–Fuller type test statistics to take account of correlations between regressors and error or serial correlations. Pedroni (2004) considers the unit root test statistic against "homogeneous" alternatives or "heterogeneous" alternative. For surveys of panel unit root and cointegration tests, see Baltagi and Kao (2000), Banerjee (1999), Breitung and Pesaran (2008), Choi (2006), etc.

### *10.3.2.2 Likelihood Approach*

The residual-based tests can only tell if there exist cointegration relations among elements of $\mathbf{w}_{it}$, they cannot tell the cointegration rank, nor account the correlations across cross-sectional units. For the "homogeneous" PVAR

model, a test of the cointegration rank can be conducted by maximizing the (transformed) likelihood function (10.1.28) subject to either

$$\Lambda_r = H_r \boldsymbol{\delta}_r + B_r, \tag{10.3.31}$$

or

$$\Lambda_{r+1} = H_{r+1} \boldsymbol{\delta}_{r+1} + B_{r+1}, \tag{10.3.32}$$

where $H_r$ and $H_{r+1}$ are, respectively, $m \times (m-r)$ and $m \times (m-r-1)$ matrices with known elements, $B_r$ and $B_{r+1}$ are, respectively, $m \times r$ and $m \times (r+1)$ matrices with known elements, $\boldsymbol{\delta}_r$ and $\boldsymbol{\delta}_{r+1}$ are, respectively, $r \times m$ and $(r+1) \times m$ matrices with unknown elements. When $T$ is fixed and $N \to \infty$, the likelihood ratio test statistic of cointegration rank $r$ versus rank $(r+1)$ is asymptotically $\chi^2$ distributed with $(m-r)^2 - (m-r-1)^2 = 2(m-r) - 1$ degrees of freedom (Binder et al. 2005 (imposing $\Pi$ to be of rank $r$ leaves $m^2 - (m-r)^2$ unrestricted coefficients in $\Pi$).

For the case of "heterogeneous" PVAR, the panel VEC approach discussed in the previous section can provide a basis to test for a common rank of cointegration on all individual units while allowing different dynamics across the individuals and interdependencies between the different individuals.

When rank $(\Pi_i^*) = m$, (10.3.1) implies $\mathbf{w}_{it}$ is stationary. For given $i$ Johansen (1991, 1995) provided the likelihood ratio tests for each of the restrictions $r = 0, \ldots, m-1$ versus stationarity (rank $(\Pi_i^*) = m$). He showed that as $T \longrightarrow \infty$, the likelihood ratio statistic of testing rank $(\Pi_i^*) = r$ versus rank $(\Pi_i^*) = m$ converges in distribution to

$$tr\left(\int dB_{m-r} B'_{m-r} \left[\int B_{m-r} B'_{m-r}\right]^{-1} \int B_{m-r} dB'_{m-r}\right), \tag{10.3.33}$$

where $B_{m-r}$ is an $(m-r)$-dimensional vector Brownian motion with an identity covariance matrix. Larsson et al. (1998) presented a likelihood-based panel test of cointegrating rank for heterogeneous panels based on the average of the individual (Johansen) rank trace statistics. When $N$ is fixed and $T$ is large, Groen and Kleibergen (2003) show that the likelihood ratio test of common cointegration rank $r (r = 0, 1, \ldots, m-1)$ across cross-sectional units versus $r = m$ (stationarity) converges in distribution to

$$\sum_{i=1}^{N} LR_i(r \mid m) = \sum_{i=1}^{N} tr\left(\int dB_{m-r,i} B'_{m-r,i} \left[\int B_{m-r,i} B'_{m-r,i}\right]^{-1} \cdot \int B_{m-r,i} dB'_{m-r,i}\right), \tag{10.3.34}$$

where $B_{m-r,i}$ is an $(m-r)$-dimensional Brownian motion for individual $i$. They show that (10.3.34) is robust with respect to cross-sectional dependence. Using a sequential limit argument ($T \to \infty$ followed by $N \to \infty$),

Groen and Kleibergen (2003) show that

$$\frac{\overline{LR}(r \mid m) - E(\overline{LR}(r \mid m))}{\sqrt{\text{Var}\,(\overline{LR}(r \mid m))}} \tag{10.3.35}$$

converges to a standard normal, where $\overline{LR}(r \mid m) = \frac{1}{N}\sum_{i=1}^{N} LR_i(r \mid m)$.

## 10.4 DYNAMIC SIMULTANEOUS EQUATIONS MODELS

### 10.4.1 The Model

The discussions in previous sections, in particular, the panel VAR model (10.1.1), can be considered as a reduced form specification of the panel dynamic simultaneous equations model of the form (assuming $\boldsymbol{\delta} = \mathbf{0}$ for ease of exposition),

$$\Gamma(L)\mathbf{w}_{it} + C\mathbf{x}_{it} = \boldsymbol{\alpha}_i^* + \boldsymbol{\epsilon}_{it}, \tag{10.4.1}$$

where $\mathbf{x}_{it}$ is a $K \times 1$ vector satisfying $E(\mathbf{x}_{it}\boldsymbol{\epsilon}_{is}') = \mathbf{0}$,

$$\Gamma(L) = \Gamma_0 + \Gamma_1 L + \ldots, +\Gamma_p L^p, \tag{10.4.2}$$

and

$$\Gamma_0 \neq I_m, \tag{10.4.3}$$

with $C = \mathbf{0}$. For simplicity, we maintain the assumption that $\boldsymbol{\epsilon}_{it}$ is independently, identically distributed (i.i.d) across $i$ and over $t$ with covariance matrix $\Omega$. Model (10.4.1), in addition to the issues of (1) the presence of time-invariant individual effects and (2) the assumption about initial observations, also (3) contains contemporaneous dependence among elements of $\mathbf{w}_{it}$. Because statistical inference can be made only in terms of observed data, the joint dependence of observed variables raises the possibility that many observational equivalent structures could generate the same observed phenomena (e.g. Hood and Koopmans 1953). To uniquely identify (10.4.1) from observed data, prior restrictions are needed. However, the presence of $\boldsymbol{\alpha}_i^*$ creates correlations between $\mathbf{w}_{it}$ and all future or past $\mathbf{w}_{is}$, $\boldsymbol{\alpha}_i^*$ can be removed from the system through a linear transformation. For instance, taking first difference of (10.4.1) yields a system of

$$\Gamma(L)\Delta\mathbf{w}_{it} + C\Delta\mathbf{x}_{it} = \Delta\boldsymbol{\epsilon}_{it}. \tag{10.4.4}$$

System (10.4.4) is in the form of Coweles Commission dynamic simultaneous equations model with first-order moving average error terms; hence the usual rank condition is necessary and sufficient to identify an equation in the system (e.g., Hsiao 1983) if we assume that the prior information takes the form of exclusion restrictions, that is, some variables are excluded from the $g$th equation. Let $[\Gamma_{0g}, \ldots, \Gamma_{pg}, C_g]$ be the matrix formed from the columns of $[\Gamma_0, \ldots, \Gamma_p, C]$ that are zero on the $g$th row. Then the necessary and sufficient

condition for the identification of the $g$th equation of (10.4.1) is (e.g., Hsiao 1983),[8]

$$\text{rank } \left[\Gamma_{og}, \ldots, \Gamma_{pg}, C_g\right] = m - 1. \tag{10.4.5}$$

For ease of exposition, we assume $p = 1$; then (10.4.1) becomes

$$\Gamma_0 \mathbf{w}_{it} + \Gamma_1 \mathbf{w}_{i,t-1} + C\mathbf{x}_{it} = \boldsymbol{\alpha}_i^* + \boldsymbol{\epsilon}_{it}. \tag{10.4.6}$$

When model (10.4.6) is identified, and $\boldsymbol{\alpha}_i^*$ is treated random and uncorrelated with $\mathbf{x}_{is}$ with known distribution, one can estimate the model by the maximum likelihood method. The resulting estimator is consistent and asymptotically normally distributed irrespective of how $N$ and $T \to \infty$. However, the MLE involves multiple integration of the joint likelihood function over time, which can be computationally complicated. On the other hand, conditional on $\boldsymbol{\alpha}_i^*$, $\boldsymbol{\epsilon}_{it}$ is independently distributed across $i$ and over $t$. Furthermore, one can ignore the issue of whether $\alpha_i^*$ is correlated with $\mathbf{x}_{is}$ or not. So we shall concentrate on the estimation methods assuming $\boldsymbol{\alpha}_i^*$ fixed.

We consider the Anderson–Rubin (1949) type limited information approach in which only the prior information on the $g$th equation is utilized. Let $g = 1$. Because only the prior restrictions of the first equation are utilized in inference there is no loss of to write (10.4.6) in the form,

$$\begin{aligned} \Gamma_0 &= \begin{bmatrix} 1 & -\boldsymbol{\gamma}_{10}' \\ \mathbf{0} & I_{m-1} \end{bmatrix}, \Gamma_1 = -\begin{bmatrix} \gamma_{11} & \boldsymbol{\gamma}_{12}' \\ \Pi_{21} & \Pi_{22} \end{bmatrix}, \\ C &= -\begin{bmatrix} \mathbf{c}_1' \\ \Pi_{23} \end{bmatrix}. \end{aligned} \tag{10.4.7}$$

We assume that the prior restrictions are in the form of exclusion restrictions and there are at least $(m - 1)$ elements in the vector $(\boldsymbol{\gamma}_{10}', \gamma_{11}, \boldsymbol{\gamma}_{12}', \mathbf{c}_1')$ which are 0 and the rank condition for the identification of the first equation is satisfied.

Premultiplying $\Gamma_0^{-1}$ to (10.4.6) yields the reduced form

$$\mathbf{w}_{it} = H_1 \mathbf{w}_{i,t-1} + H_2 \mathbf{x}_{it} + \boldsymbol{\eta}_i + \mathbf{v}_{it}, \tag{10.4.8}$$

where $H_1 = -\Gamma_0^{-1}\Gamma_1$, $H_2 = -\Gamma_0^{-1}C$, $\boldsymbol{\eta}_i = \Gamma_0^{-1}\boldsymbol{\alpha}_i^*$, $\mathbf{v}_{it} = \Gamma_0^{-1}\boldsymbol{\epsilon}_{it}$.

Taking the first difference of (10.4.6) yields

$$\begin{aligned} \Gamma_0 \Delta\mathbf{w}_{it} + \Gamma_1 \Delta\mathbf{w}_{i,t-1} + C\Delta\mathbf{x}_{it} = \Delta\boldsymbol{\epsilon}_{it}, \quad & t = 2, \ldots, T, \\ & i = 1, \ldots, N. \end{aligned} \tag{10.4.9}$$

### 10.4.2 Likelihood Approach

For ease of notations, we assume all $m$ elements of $\mathbf{w}_{it}$ appear in the first equation. Let $\boldsymbol{\theta}'$ denote the unknown elements of the vector $(\boldsymbol{\gamma}_{10}', \gamma_{11}, \boldsymbol{\gamma}_{12}', \mathbf{c}_1')$.

[8] The vector error correction representation (10.2.6) subject to (10.2.7) is a reduced form specification with the prior knowledge of the rank of cointegrations among $\mathbf{w}_{it}$. For the identification for a structural VAR and dichotomization between long-run equilibrium and short-run dynamics, see Hsiao (2001).

The likelihood approach assumes that the data-generating process of $\mathbf{w}_{i0}$ is no different from the data-generating process of any $\mathbf{w}_{it}$; hence

$$\mathbf{w}_{i0} = [I - H_1 L]^{-1}[H_2 \mathbf{x}_{i0} + \boldsymbol{\eta}_i + \mathbf{v}_{i0}], \tag{10.4.10}$$

where $L$ denotes the lag operator, $L\mathbf{w}_{it} = \mathbf{w}_{i,t-1}$. Under the assumption that the data-generating process of $\mathbf{x}_{it}$ is homogeneous (e.g., (4.3.21) or (4.5.3)), we can write

$$[I - H_1 L]^{-1} H_2 \mathbf{x}_{i0} = A\bar{\mathbf{x}}_i + \boldsymbol{\omega}_i, \tag{10.4.11}$$

where $\bar{\mathbf{x}}_i = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{it}$, and $\boldsymbol{\omega}_i$ is i.i.d. across $i$. Substituting (10.4.11) into (10.4.10) yields

$$\mathbf{w}_{i0} = A\bar{\mathbf{x}}_i + [I - H_1]^{-1}\boldsymbol{\eta}_i + \boldsymbol{\omega}_i + [I - H_1 L]^{-1}\mathbf{v}_{i0}. \tag{10.4.12}$$

Premultiplying $\Gamma_0$ to the $\Delta\mathbf{w}_{i1}$ equation yields[9]

$$\Gamma_0 \Delta\mathbf{w}_{i1} + C\mathbf{x}_{i1} + A^*\bar{\mathbf{x}}_i = \boldsymbol{\xi}_i + \boldsymbol{\epsilon}_{i1}, i = 1, \ldots, N, \tag{10.4.13}$$

where $A^* = \Gamma_0 A$, $\boldsymbol{\xi}_i = \Gamma_0\boldsymbol{\omega}_i - (\Gamma_0 + \Gamma_1)(I - H_1 L)^{-1}\mathbf{v}_{i0}$.

Maximizing the log-likelihood of the system (10.4.9) and (10.4.13) involves the complicated computation of the determinant and inversion of the covariance matrix of $(\xi_i + \epsilon_{i1}, \Delta\epsilon_{i2}, \ldots, \Delta\epsilon_{iT})$, which takes the form of a first-order moving average process. Therefore, instead of using the first difference, Hsiao and Zhou (2013), following the suggestion of Grassetti (2011), use the long difference, $\mathbf{w}^*_{it} = \mathbf{w}_{it} - \mathbf{w}_{io}$ to eliminate the individual specific effects, $\boldsymbol{\alpha}^*_i$. The alternative transformed system, making use of (10.4.12), becomes

$$\begin{aligned} \Gamma_0\mathbf{w}^*_{it} + \Gamma_1\mathbf{w}^*_{i,t-1} + C\mathbf{x}_{it} + A^*\bar{\mathbf{x}}_i = \boldsymbol{\xi}_i + \boldsymbol{\epsilon}_{it}, \\ i = 1, \ldots, N \\ t = 1, \ldots, T. \end{aligned} \tag{10.4.14}$$

The $(mT \times 1)$ error term $\mathbf{v}_i = [(\boldsymbol{\xi}_i + \boldsymbol{\epsilon}_{i1})', \ldots, (\boldsymbol{\xi}_i + \boldsymbol{\epsilon}_{iT})']'$ has mean $\mathbf{0}$ and covariance matrix of the form analogous to the single equation random effects covariance matrix,

$$\Omega = E\mathbf{v}_i\mathbf{v}_i' = \Omega_\epsilon \otimes I_T + \Omega_\xi \otimes \mathbf{e}_T\mathbf{e}_T', \tag{10.4.15}$$

where $\Omega_\epsilon = E(\boldsymbol{\epsilon}_{it}\boldsymbol{\epsilon}_{it}')$, $\Omega_\xi = E(\boldsymbol{\xi}_i\boldsymbol{\xi}_i')$ and $\mathbf{e}_T$ is a $T \times 1$ vector $(1, \ldots, 1)'$. Rewrite the covariance matrix in terms of the eigenvalues $\Omega_\epsilon$ and $\Omega_{\xi^*} = \Omega_\epsilon + T\Omega_\xi$ (Chapter 5, Section 5.2 or Avery 1977),

$$\Omega = \Omega_\epsilon \otimes Q + \Omega_{\xi^*} \otimes J, \tag{10.4.16}$$

where $Q = I_T - \frac{1}{T}\mathbf{e}_T\mathbf{e}_T'$ and $J = \frac{1}{T}\mathbf{e}_T\mathbf{e}_T'$. It follows that

$$\Omega^{-1} = \Omega_\epsilon^{-1} \otimes Q + \Omega_{\xi^*}^{-1} \otimes J. \tag{10.4.17}$$

[9] Should (4.5.3) be assumed as a data-generating process for $\mathbf{x}_{it}$, then (10.4.12) should be specified as $\Gamma_0\Delta\mathbf{w}_{i,t-1} + C\mathbf{x}_{it} + A^*\Delta\mathbf{x}_i = \boldsymbol{\xi}_i + \boldsymbol{\epsilon}_{it}$, where $\Delta\mathbf{x}_i = \frac{1}{T}\sum_{t=1}^{T}\Delta\mathbf{x}_{it}$.

Thus, the log-likelihood function of the transformed system (10.4.14) takes the multivariate analogue of the single equation log-likelihood function (3.3.19),

$$\begin{aligned}\log = &-\frac{N(T-1)}{2}\log\mid\Omega_{\epsilon}\mid-\frac{N}{2}\log\mid\Omega_{\xi^*}\mid\\ &-\frac{1}{2}\sum_{i=1}^{N}\Big\{[\mathbf{w}_{1i}^{*'}-\tilde{\boldsymbol{\theta}}'Z_i',\ \text{vec}\,(W_{2i}^{*'})'-\ \text{vec}\,(\Pi')'(I_{m-1}\otimes\tilde{X}_i')]\Omega^{-1}\\ &[\mathbf{w}_{1i}^{*'}-\tilde{\boldsymbol{\theta}}'Z_i',\ \text{vec}\,(W_{2i}^{*'})'-\ \text{vec}\,(\Pi')'(I_{m-1}\otimes\tilde{X}_i')]'\Big\}.\end{aligned}\tag{10.4.18}$$

where $W_i^*=(\mathbf{w}_{1i}^*,W_{2i}^*)$, $W_{2i}^*=(\mathbf{w}_{2i}^*,\ldots,\mathbf{w}_{mi}^*)$, $\mathbf{w}_{ji}^{*'}=(\mathbf{w}_{ji1}^*,\ldots,w_{jiT}^*)$, $j=1,\ldots,m$, $Z_i=(W_{2i}^*,\tilde{W}_{i,-1}^*,X_{1i},\mathbf{e}_T\bar{\mathbf{x}}_i')$, $\tilde{X}_i=(W_{i,-1}^*,X_i,\mathbf{e}_T\bar{\mathbf{x}}_i')$, $W_{i,-1}^*=(\tilde{W}_{i,-1}^*,\tilde{\tilde{W}}_{i,-1}^*)$, $\tilde{W}_{i,-1}^*$ and $\tilde{\tilde{\mathbf{W}}}_{i,-1}$ denote the $T\times\tilde{m}$ and $T\times(m-\tilde{m})$ matrix of $\mathbf{w}_{ji,-1}^*=(0,w_{ji1}^*,\ldots,w_{ji,T-1}^*)'$ that appear or excluded from equation 1, respectively, $X_i=(X_{1i},X_{2i})$, $X_{1i}$ and $X_{2i}$ denote the $T\times k$ and $T\times(K-k)$ included and excluded $\mathbf{x}_{it}$ in the first equation, respectively, $\tilde{\boldsymbol{\theta}}=(\boldsymbol{\theta}',\mathbf{a}_1^{*'})$, $\Pi=(\Pi_{21},\Pi_{22},\Pi_{23},A_2^*)$, where $\mathbf{a}_1^{*'}$ and $A_2^*$ denote the $1\times K$ and $(m-1)\times K$ of $A^*=\begin{pmatrix}\mathbf{a}_1^{*'}\\A_2^*\end{pmatrix}$, respectively.

The panel limited information maximum likelihood estimator (PLIML) can be obtained by iterating between

$$\begin{aligned}\Omega_{\epsilon}=\frac{1}{N(T-1)}\sum_{i=1}^{N}\Big\{&\big[\mathbf{w}_{1i}^{*'}-\hat{\tilde{\boldsymbol{\theta}}}_1'Z_i',\ \text{vec}\,(W_{2i}^{*'})-\ \text{vec}\,(\Pi')'(I_{m-1}\otimes\tilde{X}_i')\big]\\ &\big[I_m\otimes Q\big]\big[\mathbf{w}_{1i}^{*'}-\hat{\tilde{\boldsymbol{\theta}}}'Z_i',\ \text{vec}\,(W_{2i}^{*'})'-\ \text{vec}\,(\Pi')'(I_{m-1}\otimes\tilde{X}_i')\big]'\Big\},\end{aligned}\tag{10.4.19}$$

$$\hat{\Omega}_{\xi^*}=\frac{1}{N}\sum_{i=1}^{N}\begin{bmatrix}\hat{\tilde{u}}_{i1}^2 & \hat{\tilde{u}}_{i1}\hat{\tilde{\mathbf{u}}}_{2i}'\\ \hat{\tilde{\mathbf{u}}}_{2i}\hat{\tilde{u}}_{i1} & \hat{\tilde{\mathbf{u}}}_{2i}\hat{\tilde{\mathbf{u}}}_{2i}'\end{bmatrix},\tag{10.4.20}$$

$$\begin{aligned}\begin{pmatrix}\hat{\tilde{\boldsymbol{\theta}}}_1\\ \text{vec}\,(\hat{\Pi}')\end{pmatrix}=&\left\{\sum_{i=1}^{N}\begin{pmatrix}Z_i' & \mathbf{0}\\ \mathbf{0} & I_{m-1}\otimes\tilde{X}_i'\end{pmatrix}\hat{\Omega}^{-1}\begin{pmatrix}Z_i & \mathbf{0}\\ \mathbf{0} & I_{m-1}\otimes\tilde{X}_i\end{pmatrix}\right\}^{-1}\\ &\cdot\left\{\sum_{i=1}^{N}\begin{pmatrix}Z_i' & \mathbf{0}\\ \mathbf{0} & I_{m-1}\otimes\tilde{X}_i'\end{pmatrix}\hat{\Omega}^{-1}\begin{pmatrix}\mathbf{w}_{1i}^*\\ \text{vec}\,(W_{2i}^{*'})\end{pmatrix}\right\},\end{aligned}\tag{10.4.21}$$

until convergence, where $\Omega^{-1}=\Omega_{\epsilon}^{-1}\otimes Q+\Omega_{\xi^*}^{-1}\otimes J$, $\bar{u}_{i1}=\frac{1}{T}\mathbf{e}_T'(\mathbf{w}_{1i}^*-Z_i\tilde{\boldsymbol{\theta}})$, $\bar{\mathbf{u}}_{2i}'=\frac{1}{T}\mathbf{e}_T'[W_{2i}^*-\tilde{X}_i\Pi']$. Hsiao and Zhou (2013) show that the PLIML is consistent and asymptotically normally distributed centered at the true value (i.e., no asymptotic bias) independent of the way $N$ or $T$ or both go to infinity. However, if the initial value $\mathbf{w}_{i0}$ is mistakenly treated as fixed constants, maximizing the log-likelihood function of the system (10.4.9) is consistent and asymptotically unbiased only if $N$ is fixed and $T$ tends to infinity. If both

$N$ and $T$ tend to infinity and $\frac{N}{T} \to c \neq 0 < \infty$, the (quasi) MLE mistakenly treating $\Delta \mathbf{w}_{i1}$ as fixed constants is asymptotically biased of order $\sqrt{\frac{N}{T}}$.

### 10.4.3 Method of Moments Estimator

The method of moments method derives the estimator from the orthogonality conditions:

$$E(\mathbf{q}_{it} \Delta \boldsymbol{\epsilon}'_{it}) = \mathbf{0} \tag{10.4.22}$$

for some variables, $\mathbf{q}_{it}$. However, panel data is multidimensional. How the sample moments approximate the population moments plays a pivotal role in the asymptotic distribution of a method of moments estimator. For instance, generalization of Anderson and Hsiao (1981, 1982) (Section 4.3.3.c) simple instrumental variable estimator using the moment conditions $E\big[\binom{\mathbf{w}_{i,t-2}}{\Delta \mathbf{x}_{it}} \Delta \epsilon_{1it}\big] = \mathbf{0}$, or $E\big[\binom{\Delta \mathbf{w}_{i,t-2}}{\Delta \mathbf{x}_{it}} \Delta \epsilon_{1it}\big] = \mathbf{0}$, yields consistent and asymptotically unbiased estimator of $\boldsymbol{\theta}$ independent of the way $N$ or $T$ or both tend to infinity (for details, see Hsiao and Zhang 2013). However, $\mathbf{w}_{i,t-2}$ or $\Delta \mathbf{w}_{i,t-2}$ are not the only instruments that satisfy (10.4.22). All $\mathbf{w}_{i,t-2-j}$ (or $\Delta \mathbf{w}_{i,t-2-j}$), $j = 1, \ldots, t-2$ and $\Delta \mathbf{x}'_i = (\Delta \mathbf{x}'_{i2}, \ldots, \Delta \mathbf{x}'_{iT})$ are legitimate instruments. Let $\mathbf{q}'_{it} = (\mathbf{w}'_{i,t-2}, \ldots, \mathbf{w}'_{i0}, \Delta \mathbf{x}'_i)$ and let

$$D_i = \begin{bmatrix} \mathbf{q}_{i2} & \mathbf{0} & \mathbf{0} & \cdot & 0 \\ \mathbf{0} & \mathbf{q}_{i3} & \mathbf{0} & \cdot & \mathbf{0} \\ \cdot & \cdot & \cdot & \cdot & \mathbf{0} \\ \mathbf{0} & \cdot & \cdot & \cdot & \mathbf{q}_{iT} \end{bmatrix}, \tag{10.4.23}$$

be the $R \times (T-1)$ block dimensional matrix, where $R$ denotes the dimension of $(\mathbf{q}'_{i2}, \ldots, \mathbf{q}'_{iT})'$; then

$$E(D_i \Delta \boldsymbol{\epsilon}_{1i}) = 0, \quad i = 1, \ldots, N, \tag{10.4.24}$$

where $\Delta \boldsymbol{\epsilon}'_{1i} = (\Delta \epsilon_{1i2}, \ldots, \Delta \epsilon_{1iT})$ with $E(\Delta \boldsymbol{\epsilon}_{1i}) = \mathbf{0}$ and $E(\Delta \boldsymbol{\epsilon}_{1i} \Delta \boldsymbol{\epsilon}'_{1i}) = \sigma_1^2 \tilde{A}$, $\sigma_1^2 = E(\epsilon_{1it}^2)$,

$$\tilde{A} = \begin{bmatrix} 2 & -1 & 0 & \cdot & 0 \\ -1 & 2 & -1 & \cdot & \cdot \\ 0 & -1 & 2 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & -1 & 2 \end{bmatrix}.$$

The Arellano–Bond (1991) type GMM estimator of $\boldsymbol{\theta}$ is to find $\hat{\boldsymbol{\theta}}$ that minimizes

$$\left( \sum_{i=1}^{N} \Delta \boldsymbol{\epsilon}'_{1i} D'_i \right) \left( \sum_{i=1}^{N} D_i \tilde{A} D'_i \right)^{-1} \left( \sum_{i=1}^{N} D_i \Delta \boldsymbol{\epsilon}_{1i} \right), \tag{10.4.25}$$

which yields an estimator of the form (4.3.47).

When $T$ is fixed and $N \to \infty$, $\sqrt{NT}(\hat{\boldsymbol{\theta}}_{\text{GMM}} - \boldsymbol{\theta})$ is asymptotically normally distributed with mean $\mathbf{0}$ and covariance matrix

$$\sigma_1^2 \left\{ \frac{1}{NT} \left[ \sum_{i=1}^{N} \tilde{Z}_i' D_i' \right] \left( \sum_{i=1}^{N} D_i \tilde{A} D_i' \right)^{-1} \left[ \sum_{i=1}^{N} D_i \tilde{Z}_i \right] \right\}^{-1}, \quad (10.4.26)$$

where $\tilde{Z}_i = (\Delta W_{2i}, \Delta \tilde{W}_{i,-1}, \Delta X_{1i})$, $\Delta W_{2i}$, $\Delta \tilde{W}_{i,-1}$ denote the $T$ time series stacked $(\Delta w_{2it}, \ldots, \Delta w_{mit})$ and the lagged $(\Delta \mathbf{w}_{1i,-1}, \ldots, \Delta \mathbf{w}_{mi,-1})$ that appear in the first equation respectively. However, because

$$E\left[ \begin{pmatrix} \Delta W_{2i} \\ \Delta \tilde{W}_{i,-1}' \end{pmatrix} D_i' \left( \sum_{i=1}^{N} D_i \tilde{A} D_i' \right)^{-1} D_i \Delta \boldsymbol{\epsilon}_{1i} \right] \neq \mathbf{0}, \quad (10.4.27)$$

the process of removing individual-specific effects creates a second-order bias that is of order $\log T$. If $T$ increases with $N$, it is shown by Akashi and Kunitomo (2011, 2012) that the GMM is inconsistent if $\frac{T}{N} \to c \neq 0$ as $N \to \infty$. Even when $c \to 0$, as long as $\frac{T^3}{N} < \infty$ as $N \to \infty$, $\sqrt{NT}(\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta})$ is not centered at 0. Monte Carlo studies conducted by Akashi and Kunitomo (2012a,b) show that the GMM estimator is badly biased when $N$ and $T$ are both large.

Akashi and Kunitomo (2011, 2012) propose a panel least variance ratio estimator (PLVAR), which is a panel generalization of Anderson and Rubin (1949) limited information maximum likelihood estimator formula.[10] They show that when $\frac{T}{N} \to 0$ as $N \to \infty$, the panel least variance ratio estimator is asymptotically normally distributed centered at the true value. They have also derived the asymptotic bias of PLVAR when $\frac{T}{N} \to c \neq 0$. Monte Carlo studies confirm that the PLVAR is almost median-unbiased after correcting the bias.

Whether an estimator of $\boldsymbol{\theta}$ multiplied by the scale factor $\sqrt{NT}$ (the magnitude of the inverse of the standard error) is centered at the true value or not has important implication in hypothesis testing. A consistent but asymptotically biased estimataor could lead to significant size distortion in hypothesis testing (Hsiao and Zhou 2013). The source of asymptotic bias of the Arellano–Bond type GMM is the use of using cross-sectional mean $\frac{1}{N}\sum_{i=1}^{N} D_i \Delta \boldsymbol{\epsilon}_{1i}$ to approximate the population moments $E(D_i \Delta \boldsymbol{\epsilon}_{1i}) = \mathbf{0}$. On the other hand, the MLE or the simple instrumental variable estimator uses all NT observations to approximate the moment conditions (10.4.22), and hence is asymptotically unbiased independent of the way $N$ or $T$ or both tend to infinity.

[10] Akashi and Kunitomo (2012) actually called their estimator PLIML. However, it appears that their estimator is more in the spirit of LVR estimator, see (10.4.18)–(10.4.21).

CHAPTER 11

# Incomplete Panel Data

Thus far our discussions have been concentrated on situations in which the sample of $N$ cross-sectional units over $T$ time periods is sufficient to identify a behavioral model. In this chapter we turn to issues of incomplete panel data. We first discuss issues when some individuals are dropped out of the experiment or survey. We note that when individuals are followed over time, there is a high probability that this may occur. Because the situations where individuals are missing for a variety of behavioral reasons have been discussed in Chapter 8, Section 8.3, in this chapter we consider only the situations where (1) individuals are missing randomly or are being rotated; (2) a series of independent cross-sections are observed over time; and (3) only a single set of cross-sectional data is available in conjunction with the aggregate time series observations. We then consider the problems of estimating dynamic models when the length of time series is shorter than the maximum order of the lagged variables included in the equation.

## 11.1 ROTATING OR RANDOMLY MISSING DATA

In many situations we do not have complete time series observations on cross-sectional units. Instead, individuals are selected according to a "rotating" scheme that can be briefly stated as follows: Let all individuals in the population be numbered consecutively. Suppose the sample in period 1 consists of individuals $1, 2, \ldots, N$. In period 2, individuals $1, \ldots, m_1$ $(0 \leq m_1 \leq N)$ are replaced by individuals $N+1, \ldots, N+m_1$. In period 3, individuals $m_1+1, \ldots, m_1+m_2$ $(0 \leq m_2 \leq N)$ are replaced by individuals $N+m_1+1, \ldots, N+m_1+m_2$, and so on. This procedure of dropping the first $m_{t-1}$ individuals from the sample selected in the previous period and augmenting the sample by drawing $m_{t-1}$ individuals from the population so that the sample size remains the same continues through all periods. Hence, for $T$ periods, although the total number of observations remains at $NT$, we have observed $N + \sum_{t=1}^{T-1} m_t$ individuals.

"Rotation" of a sample of micro units over time is quite common. It can be caused by deliberate policy of the data-collecting agency (e.g., the Bureau

of the Census) because of the worry that if the number of times respondents have been exposed to a survey gets large, the data may be affected and even behavioral changes may be induced. Or it can arise because of the consideration of optimal simple design so as to gain as much information as possible from a given budget (e.g., Aigner and Balestra 1988; Nijman, Verbeek, and van Soest 1991). It can also arise because the data-collecting agency can neither force nor persuade randomly selected individuals to report more than once or twice, particularly if detailed and time-consuming reporting is required. For example, the Survey of Income and Program Participation, which began field work in October 1983, has been designed as an ongoing series of national panels, each consisting of about 20,000 interviewed households and having a duration of 2.5 years. Every four months the Census Bureau will interview each individual of age 15 years or older in the panel. Information will be collected on a monthly basis for most sources of money and non-money income, participation in various governmental transfer programs, labor-force status, and household composition.

Statistical methods developed for analyzing complete panel data can be extended in a straightforward manner to analyze rotating samples if rotation is by design (i.e., randomly dropping and addition of individuals) and if a model is static and the error terms are assumed to be independently distributed across cross-sectional units. The likelihood function for the observed samples in this case is simply the product of the $N + \sum_{t=1}^{T-1} m_t$ joint density of $(y_{it_i}, y_{i,t_i+1}, \ldots, y_{iT_i})$,

$$L = \prod_{i=1}^{N+\sum_{t=1}^{T-1} m_t} f(y_{it_i}, \ldots, y_{iT_i}), \tag{11.1.1}$$

where $t_i$ and $T_i$ denote the first and the last periods during which the $i$th individual was observed. Apart from the minor modifications of $t_i$ for 1 and $T_i$ for $T$, (11.1.1) is basically of the same form as the likelihood functions for the complete panel data.

As an illustration, we consider a single-equation error-components model (Biørn 1981). Let

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + v_{it}, \tag{11.1.2}$$

where $\boldsymbol{\beta}$ and $\mathbf{x}_{it}$ are $k \times 1$ vectors of parameters and explanatory variables, respectively, and

$$v_{it} = \alpha_i + u_{it}. \tag{11.1.3}$$

The error terms $\alpha_i$ and $u_{it}$ are independent of each other and are independently distributed, with zero means and constant variances $\sigma_\alpha^2$ and $\sigma_u^2$, respectively. For ease of exposition, we assume that $\alpha_i$ and $u_{it}$ are uncorrelated with $\mathbf{x}_{it}$.[1]

[1] If $\alpha_i$ are correlated with $\mathbf{x}_{it}$, we can eliminate the linear dependence between $\alpha_i$ and $\mathbf{x}_{it}$ by assuming $\alpha_i = \Sigma_t \mathbf{a}_t' \mathbf{x}_{it} + \epsilon_i$. For details, see Chapter 3 or Mundlak (1978a).

We also assume that in each period a fixed number of individuals are dropped out of the sample and the same number of individuals from the population are added back to the sample (namely, $m_t = m$ for all $t$). Thus, the total number of individuals observed is

$$H = (T-1)m + N. \tag{11.1.4}$$

Denote the number of times the $i$th individual is observed by $q_i$, then $q_i = T_i - t_i + 1$. Stacking the time series observations for the $i$th individual in vector form, we have

$$\mathbf{y}_i = X_i\boldsymbol{\beta} + \mathbf{v}_i, \tag{11.1.5}$$

where

$$\underset{q_i\times 1}{\mathbf{y}_i} = (y_{it_i}, \ldots, y_{iT_i})', \qquad \underset{q_i\times k}{X_i} = (\mathbf{x}'_{it}),$$

$$\mathbf{v}_i = (\alpha_i + u_{it_i}, \ldots, \alpha_i + u_{iT_i})'.$$

The variance–covariance matrix of $v_i$ is

$$V_i = \sigma_u^2 + \sigma_\alpha^2 \quad \text{if } q_i = 1 \tag{11.1.6a}$$

and is

$$V_i = E\mathbf{v}_i\mathbf{v}_i' = \sigma_u^2 I_{q_i} + \sigma_\alpha^2 J_i \quad \text{if } q_i > 1, \tag{11.1.6b}$$

where $J_i$ is a $q_i \times q_i$ matrix with all elements equal to 1. Then, for $q_i = 1$,

$$V_i^{-1} = (\sigma_u^2 + \sigma_\alpha^2)^{-1}, \tag{11.1.7a}$$

and for $q_i > 1$,

$$V_i^{-1} = \frac{1}{\sigma_u^2}\left[I_{q_i} - \frac{\sigma_\alpha^2}{\sigma_u^2 + q_i\sigma_\alpha^2} J_i\right]. \tag{11.1.7b}$$

Because $\mathbf{y}_i$ and $\mathbf{y}_j$ are uncorrelated, the variance–covariance matrix of the stacked equations $(y'_1, \ldots, y'_{N+(T-1)m})'$ is block-diagonal. Therefore, the GLS estimator of $\beta$ is

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = \left[\sum_{i=1}^{N+(T-1)m} X_i'V_i^{-1}X_i\right]^{-1}\left[\sum_{i=1}^{N+(T-1)m} X_i'V_i^{-1}\mathbf{y}_i\right]. \tag{11.1.8}$$

The GLS estimator of $\boldsymbol{\beta}$ is equivalent to first premultiplying the observation matrix $[\mathbf{y}_i, X_i]$ by $P_i$, where $P_i'P_i = V_i^{-1}$, and then regressing $P_i\mathbf{y}_i$ on $P_iX_i$ (Theil 1971, Chapter 6). In other words, the least-squares method is applied to the data transformed by the following procedure: For individuals who are observed only once, multiply the corresponding $y$'s and $\mathbf{x}$'s by $(\sigma_u^2 + \sigma_\alpha^2)^{-1/2}$. For individuals who are observed $q_i$ times, subtract from the corresponding $y$'s and $\mathbf{x}$'s a fraction $1 - [\sigma_u/(\sigma_u^2 + q_i\sigma_\alpha^2)^{1/2}]$ of their group means, $\overline{y}_i$ and $\overline{\mathbf{x}}_i$, where $\overline{y}_i = (1/q_i)\sum_t y_{it}$ and $\overline{\mathbf{x}}_i = (1/q_i)\sum_t \mathbf{x}_{it}$ and then divide them by $\sigma_u$.

To obtain separate estimates $\sigma_u^2$ and $\sigma_\alpha^2$ we need at least one group for which $q_i > 1$. Let $\Theta$ denote the set of those individuals with $q_i > 1$, $\Theta = \{i \mid q_i > 1\}$, and $H^* = \sum_{i\in\Theta} q_i$. Then $\sigma_u^2$ and $\sigma_\alpha^2$ can be consistently estimated by

$$\hat{\sigma}_u^2 = \frac{1}{H^*} \sum_{i\in\Theta} \sum_{t=t_i}^{Ti} [(y_{it} - \overline{y}_i) - \hat{\boldsymbol{\beta}}'(\mathbf{x}_{it} - \overline{\mathbf{x}}_i)]^2, \tag{11.1.9}$$

and

$$\hat{\sigma}_\alpha^2 = \frac{1}{N + (T-1)m} \sum_{i=1}^{N+(T-1)m} \left[ (\overline{y}_i - \hat{\boldsymbol{\beta}}'\overline{\mathbf{x}}_i)^2 - \frac{1}{q_i}\hat{\sigma}_u^2 \right]. \tag{11.1.10}$$

Similarly, we can apply the MLE by maximizing the logarithm of the likelihood function (11.1.1):

$$\begin{aligned} \log L &= -\frac{NT}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{N+(T-1)m} \log \mid V_i \mid \\ &\quad - \frac{1}{2}\sum_{i=1}^{N+(T-1)m} (\mathbf{y}_i - X_i\boldsymbol{\beta})' V_i^{-1} (\mathbf{y}_i - X_i\boldsymbol{\beta}) \\ &= -\frac{NT}{2}\log 2\pi - \frac{1}{2}\left[\sum_{i=1}^{N+(T-1)m} (q_i - 1)\right] \log \sigma_u^2 \\ &\quad - \frac{1}{2}\sum_{i=1}^{N+(T-1)m} \log(\sigma_u^2 + q_i\sigma_\alpha^2) \\ &\quad - \frac{1}{2}\sum_{i=1}^{N+(T-1)m} (\mathbf{y}_i - X_i\boldsymbol{\beta})' V_i^{-1} (\mathbf{y}_i - X_i\boldsymbol{\beta}). \end{aligned} \tag{11.1.11}$$

Conditioning on $\sigma_u^2$ and $\sigma_\alpha^2$, the MLE is the GLS (11.1.8). Conditioning on $\boldsymbol{\beta}$, the MLEs of $\sigma_u^2$ and $\sigma_\alpha^2$ are the simultaneous solutions of the following equations:

$$\begin{aligned} \frac{\partial \log L}{\partial \sigma_u^2} &= -\frac{1}{2\sigma_u^2}\left[\sum_{i=1}^{N+(T-1)m} (q_i - 1)\right] \\ &\quad - \frac{1}{2}\left[\sum_{i=1}^{N+(T-1)m} \frac{1}{(\sigma_u^2 + q_i\sigma_\alpha^2)}\right] \\ &\quad + \frac{1}{2\sigma_u^4}\sum_{i=1}^{N+(T-1)m} (y_i - X_i\boldsymbol{\beta})' Q_i (y_i - X_i\boldsymbol{\beta}) \\ &\quad + \frac{1}{2}\sum_{i=1}^{N+(T-1)m} \frac{q_i}{(\sigma_u^2 + q_i\sigma_\alpha^2)^2} (\overline{y}_i - \overline{\mathbf{x}}_i'\boldsymbol{\beta})^2 = 0 \end{aligned} \tag{11.1.12}$$

and

$$\frac{\partial \log L}{\partial \sigma_\alpha^2} = -\frac{1}{2}\sum_{i=1}^{N+(T-1)m} \times \left[\frac{q_i}{\sigma_u^2 + q_i\sigma_\alpha^2} - \frac{q_i^2}{(\sigma_u^2 + q_i\sigma_\alpha^2)^2}(\overline{y}_i - \overline{\mathbf{x}}_i'\boldsymbol{\beta})^2\right] = 0 \tag{11.1.13}$$

where $Q_i = I_{q_i} - (1/q_i)\mathbf{e}_{q_i}\mathbf{e}_{q_i}'$, and $\mathbf{e}_{q_i}$ is a $q_i \times 1$ vector of ones. Unfortunately, because $q_i$ are different for different $i$, (11.1.12) and (11.1.13) cannot be put in the simple form of (3.3.25) and (3.3.26). Numerical methods will have to be used to obtain a solution. However, computation of the MLEs of $\boldsymbol{\beta}$, $\sigma_u^2$ and $\sigma_\alpha^2$ can be simplified by iteratively switching between (11.1.8) and (11.1.12)–(11.1.13).

If $\alpha_i$ are treated as fixed constants, $\boldsymbol{\beta}$ can be consistently estimated through the within transformation,

$$\hat{\boldsymbol{\beta}}_{cv} = \left[\sum_{i=1}^{N}\sum_{t=t_i}^{Ti}(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\right]^{-1} \cdot \left[\sum_{i=1}^{N}\sum_{t=t_i}^{Ti}(\mathbf{x}_{it} - \bar{\mathbf{x}}_i)(y_{it} - \bar{y}_i)\right]. \tag{11.1.14}$$

If the model is dynamic, similar modification of the GMM (e.g., (4.3.47)) (e.g., Collado (1997); Moffitt 1993) can be applied to obtain consistent estimators of the coefficients. The likelihood approach for dynamic models has the issue of initial conditions.[2] Different assumptions about initial conditions will suggest different ways of incorporating new observations with those already in the sample. If $\alpha_i$ are treated as random, it would appear a reasonable approximation in this case is to modify the methods based on the assumption that initial observations are correlated with individual effects and have stationary variances (Chapter 4, Case IVc or IVc′). However, the assumption imposed on the model will have to be even more restrictive. If $\alpha_i$ are treated as fixed, a similar modification can be applied to the transform MLE (e.g., (4.5.6)).

When data are randomly missing, a common procedure is to focus on the subset of individuals for which complete time series observations are available. However, the subset of incompletely observed individuals also contains some information about unknown parameters. A more efficient and computationally somewhat more complicated way is to treat randomly missing samples in the same way as rotating samples. For instance, the likelihood function (11.1.1), with the modification that $t_i = 1$ for all $i$, can also be viewed the likelihood function for this situation: In time period 1 there are $N + \sum_{t=1}^{T-1} m_t$ individuals; in period 2, $m_1$ of them randomly drop out, and so on, such that at the end of $T$ periods there are only $N$ individuals remaining in the sample. Thus, the

[2] For details, see Chapters 4 and 6.

procedure for obtaining the GLS or MLE for unknown parameters with all the observations utilized is similar to the situation of rotating samples.

To test if attrition is indeed random, we note that either the complete sample unbalanced panel estimators discussed above or the estimators based on the balanced panel subsample estimators converge to the true value under the null. Under the alternative that attrition is behaviorally related, neither estimators are consistent. However, if the individual-specific effects $\alpha_i$ and the error $u_{it}$ are independent of the regressors $\mathbf{x}_{it}$, and are independently normally distributed, a test of random attrition versus behaviorally related attrition is a student $t$-test of the significance of sample selection effect (e.g., (8.2.7)). If $\alpha_i$ are correlated with $\mathbf{x}_{it}$, one can construct a Hausman (1978) type test statistic for the significance of the difference between the Kyriazidou (1997) fixed effects sample selection estimator (e.g., (8.5.4)) and the complete sample unbalanced panel data with estimator (11.1.14). Further, if all initial samples are observed for at least two periods before attrition occurs, then the within estimator based on initial complete samples within estimator and the within estimator based on all observed samples (unbalanced panel) converge to the true value under the null and converge to different values under the alternative. A straightforward Hausman (1978) test statistic,

$$(\hat{\boldsymbol{\beta}}_{cv} - \tilde{\boldsymbol{\beta}}_{cvs})' \Big[\text{Cov}(\tilde{\boldsymbol{\beta}}_{cvs}) - \text{Cov}(\hat{\boldsymbol{\beta}}_{cv})\Big]^{-1} (\hat{\boldsymbol{\beta}}_{cv} - \tilde{\boldsymbol{\beta}}_{cvs}) \qquad (11.1.15)$$

can be used to test the null of attrition being random, where $\tilde{\boldsymbol{\beta}}_{cvs}$ and $\text{Cov}(\tilde{\boldsymbol{\beta}}_{cvs})$ denote the within estimator of $\boldsymbol{\beta}$ and its covariance matrix based on the initial sample from period 1 to $t^*$, where $t^*$ denotes the last time period before any attrition (at period $t^* + 1$) occurs.

## 11.2 PSEUDO-PANELS (OR REPEATED CROSS-SECTIONAL DATA)

In many situations there could be no genuine panel where specific individuals or firms are followed over time. However, repeated cross-sectional surveys may be available, where random samples are taken from the population at consecutive points in time. The major limitation of repeated cross-sectional data is that individual histories are not available, so it is not possible to control the impact of unobserved individual characteristics in a linear model of the form

$$y_{it} = \mathbf{x}'_{it}\boldsymbol{\beta} + \alpha_i + u_{it}, \qquad (11.2.1)$$

if $\alpha_i$ and $\mathbf{x}_{it}$ are correlated through the fixed effects estimator discussed in Chapter 3.[3] However, several authors have argued that with some additional assumptions $\boldsymbol{\beta}$ may be identifiable from a single cross-section or a series of

[3] If $\alpha_i$ and $\mathbf{x}_{it}$ are uncorrelated, there is no problem of consistently estimating (11.2.1) with repeated cross-sectional data because $E(\alpha_i + u_{it} \mid \mathbf{x}_{it}) = 0$.

independent cross-sections (e.g., Blundell, Browning, and Meghir 1994; Deaton 1985; Heckman and Robb 1985; Moffitt 1993).

Deaton (1985) suggests using a cohort approach to obtain consistent estimators of $\boldsymbol{\beta}$ of (11.2.1) if repeated cross-sectional data are available. In this approach individuals sharing common observed characteristics, such as age, sex, education, or socioeconomic background are grouped into *cohorts*. For instance, suppose that one can divide the sample into $C$ cohorts in terms of an $L \times 1$ vector of individual characteristics, $\mathbf{z}_c$, $c = 1, \ldots, C$. Let $\mathbf{z}_{it}$ be the corresponding $L$-dimensional vector of individual-specific variables for the $i$th individual of the $t$th cross-sectional data. Then $(y_{it}, \mathbf{x}_{it})$ belong to the $c$th cohort if $\mathbf{z}_{it} = \mathbf{z}_c$. Let $\psi_{ct} = \{i \mid \mathbf{z}_{it} = \mathbf{z}_c \text{ for the } t\text{th cross-sectional data}\}$ be the set of individuals that belong to the cohort $c$ at time $t$, $c = 1, \ldots, C$, $t = 1, \ldots, T$. Let $N_{ct}$ be the number of individuals in $\psi_{ct}$. Deaton (1985) assumes individuals belonging to the same cohort have the same specific effects,

$$\alpha_i = \sum_{c=1}^{C} \alpha_c d_{itc}, \tag{11.2.2}$$

where $d_{itc} = 1$ if the $i$th individual of the $t$th cross-sectional data belongs to cohort $c$ and 0 otherwise. Let $\bar{y}_{ct} = \frac{1}{N_{ct}} \sum_{i \in \psi_{ct}} y_{it}$ and $\bar{\mathbf{x}}_{ct} = \frac{1}{N_{ct}} \sum_{i \in \psi_{ct}} \mathbf{x}_{it}$, then the data $(\bar{y}_{ct}, \bar{\mathbf{x}}'_{ct})$ becomes a pseudo-panel with repeated observations on $C$ cohorts over $T$ time periods. Aggregation of observations to cohort level for the model (11.2.1) leads to

$$\bar{y}_{ct} = \bar{\mathbf{x}}'_{ct}\boldsymbol{\beta} + \alpha_c + \bar{u}_{ct}, \quad \begin{matrix} c = 1, \ldots, C, \\ t = 1, \ldots, T, \end{matrix} \tag{11.2.3}$$

where $\bar{u}_{ct} = \frac{1}{N_{ct}} \sum_{i \in \psi_{ct}} u_{it}$.

If $\mathbf{x}_{it}$ are uncorrelated with $u_{it}$, the within estimator (3.2.8) can be applied to the pseudo panel

$$\hat{\boldsymbol{\beta}}_w = \left( \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{\mathbf{x}}_{ct} - \bar{\mathbf{x}}_c)(\bar{\mathbf{x}}_{ct} - \bar{\mathbf{x}}_c)' \right)^{-1} \left( \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{\mathbf{x}}_{ct} - \bar{\mathbf{x}}_c)(\bar{y}_{ct} - \bar{y}_c) \right), \tag{11.2.4}$$

where $\bar{\boldsymbol{x}}_c = \frac{1}{T} \sum_{t=1}^{T} \bar{\boldsymbol{x}}_{ct}$, and $\bar{y}_c = \frac{1}{T} \sum_{t=1}^{T} \bar{y}_{ct}$. When $T \to \infty$ or if $T$ is fixed but $N \to \infty$, $C \to \infty$, and $\frac{C}{N} \longrightarrow 0$, (11.2.4) is consistent.

Although the cohort approach offers a useful framework to make use of independent cross-sectional information, there are problems with some of its features. First, the assertion of intra-cohort homogeneity (11.2.2) appears very strong, in particular, in view of the cohort classification is often arbitrary. Second, the practice of establishing the large sample properties of econometric estimators and test statistics by assuming that the number of cohorts, $C$, tends to infinity is not satisfactory. There is often a physical limit beyond which one cannot increase the number of cohorts. The oft-cited example

of date of birth cohorts is a case in point. Third, grouping or aggregating individuals may result in the loss of information. Moreover, in general, the number of individuals at different cohorts or different time are different, $N_{ct} \neq N_{c's}$. Even $u_{it}$ is homoscedastic and independently distributed, Var $(\bar{u}_{ct}) = \frac{\sigma_u^2}{N_{ct}} \neq$ Var $(\bar{u}_{c's}) = \frac{\sigma_u^2}{N_{c's}}$. Therefore, the $t$-statistic based on the conventional within estimator formula is not asymptotically standard normally distributed unless $N_{ct} = N_{c's}$ for all $c, c', t, s$, and var $(u_{it})$ is a constant across $i$. Hence, the resulting inference can be misleading (Inoue 2008).

Suppose (11.2.2) indeed holds and if $u_{it}$ is independently, identically distributed the problem of heterocesdasticity of $\bar{u}_{ct}$ can be corrected by applying the weighted within estimator,

$$\hat{\boldsymbol{\beta}}_{ww} = \left\{ \sum_{c=1}^{C} \sum_{t=1}^{T} \left[ N_{ct} (\bar{\mathbf{x}}_{ct} - \bar{\mathbf{x}}_c)(\bar{\mathbf{x}}_{ct} - \bar{\mathbf{x}}_c)' \right] \right\}^{-1} \cdot \left\{ \sum_{c=1}^{C} \sum_{t=1}^{T} [N_{ct} (\bar{\mathbf{x}}_{ct} - \bar{\mathbf{x}}_c)(\bar{y}_{ct} - \bar{y}_c)] \right\}. \tag{11.2.5}$$

The variance covariance matrix of $\hat{\boldsymbol{\beta}}_{ww}$ is

$$\text{Cov}\,(\hat{\boldsymbol{\beta}}_{ww}) = \sigma^2 \left\{ \sum_{c=1}^{C} \sum_{t=1}^{T} \left[ N_{ct} (\bar{\mathbf{x}}_{ct} - \bar{\mathbf{x}}_c)(\bar{x}_{ct} - \bar{\mathbf{x}}_c)' \right] \right\}^{-1}. \tag{11.2.6}$$

A cohort approach also raises a complicated issue for the estimation of a dynamic model of the form,

$$y_{it} = \gamma y_{i,t-1} + \mathbf{x}_{it}' \boldsymbol{\beta} + \alpha_i + u_{it}, \tag{11.2.7}$$

because $y_{i,t-1}$ is unavailable. The cohort approach will have to use the $y$-values of other individuals observed at $t-1$ to predict the missing $y_{i,t-1}$, $\hat{y}_{i,t-1}$. Suppose there exists a set of instruments $\mathbf{z}_{it}$ such that the orthogonal projection of $y_{it}$ on $\mathbf{z}_{it}$ are available,

$$E^*(y_{it} \mid \mathbf{z}_{it}) = \mathbf{z}_{it}' \boldsymbol{\delta}_t, \tag{11.2.8}$$

where $E^*(y \mid \mathbf{z})$ denotes the minimum mean-square-error linear predictor of $y$ by $\mathbf{z}$. Let $\hat{y}_{i,t-1} = \mathbf{z}_{i,t-1} \hat{\boldsymbol{\delta}}_{t-1}$, then (11.2.7) becomes

$$y_{it} = \gamma \hat{y}_{i,t-1} + \mathbf{x}_{it}' \boldsymbol{\beta} + v_{it}, \tag{11.2.9}$$

where

$$v_{it} = \alpha_i + u_{it} + \gamma (y_{i,t-1} - \hat{y}_{i,t-1}). \tag{11.2.10}$$

Girma (2000); Moffitt (1993); and McKenzie (2004) assume that $\mathbf{z}_{it} = \mathbf{z}_c$, are a set of cohort dummies for all $i$ belonging to cohort $c$. This is equivalent to simply using the dummy variable $d_{itc}$, as instruments for $y_{it}$ where $d_{itc} = 1$ if $y_{it}$ belongs to cohort $c$ and 0 otherwise, for $c = 1, \ldots, C$. Taking the average

of $y_{it}$ or $\hat{y}_{it}$ for $i$ belonging to cohort $c$ leads to the following pseudo panel dynamic model

$$\bar{y}_{ct} = \gamma \bar{y}_{c,t-1} + \alpha_c + \bar{\mathbf{x}}'_{ct}\boldsymbol{\beta} + v_{ct}, \quad \begin{matrix} c = 1, \ldots, C, \\ t = 1, \ldots, T, \end{matrix} \tag{11.2.11}$$

where all variables denote period-by-period averages within each cohort. The covariance estimator of (11.2.11) would be consistent estimators of $\gamma$ and $\boldsymbol{\beta}$ provided

$$\text{Cov}\,(v_{ct}, \bar{y}_{c,t-1}) = 0, \tag{11.2.12}$$

and

$$\text{Cov}\,(v_{ct}, \bar{\mathbf{x}}_{ct}) = \mathbf{0}. \tag{11.2.13}$$

However, even under the assumption (11.2.2),

$$E[(\alpha_i + u_{it})\mathbf{z}_{i,t-1} \mid i \in \psi_{c,t-1}] = \mathbf{0}, \tag{11.2.14}$$

in general,

$$E[(y_{i,t-1} - \hat{y}_{i,t-1})\hat{y}_{i,t-1}] \neq 0. \tag{11.2.15}$$

Moreover, as pointed out by Verbeek (2007); Verbeek and Vella (2005) that although under the exogeneity assumption

$$\text{Cov}\,[(\alpha_i + u_{it})\mathbf{x}_{it}] = \mathbf{0}, \tag{11.2.16}$$

(11.2.13) is unlikely to hold because $\mathbf{x}_{i,t-1}$ drives $y_{i,t-1}$ and $\mathbf{x}_{it}$ is likely to be serially correlated. To overcome the problem of correlations between the regressors and errors in (11.2.11), one will have to also find instruments for $\mathbf{x}_{it}$ as well. Unfortunately, the availability of such instruments in addition to $\mathbf{z}_i$ in many applications may be questionable (e.g., Verbeek and Vella 2005). It remains to be seen whether in empirical applications of cohort approach suitable instruments can be found that have time-varying relationships with $\mathbf{x}_{it}$ and $y_{i,t-1}$, while in the meantime they should not have any time-varying relationship with the error term (11.2.10) (e.g., Verbeek 2007; Verbeek and Vella 2005).

## 11.3 POOLING OF SINGLE CROSS-SECTIONAL AND SINGLE TIME SERIES DATA

### 11.3.1 Introduction

In this section we consider the problem of pooling when we have a single cross-sectional and a single time series data set. Empirical studies based solely on time series data often result in very inaccurate parameter estimates because of the high collinearity among the explanatory variables. For instance, income

and price time series can be highly correlated. On the other hand, a cross-sectional data set may contain good information on household income, but not on price, because the same price is likely to be faced by all households. Thus, each data set contains useful information on some of the variables, but not on all the variables to allow accurate estimates of all the parameters of interest. A classic example of this is provided in a study (Stone 1954) of aggregate-demand systems in which there was no cross-sectional variation in commodity prices and inadequate time-series variation in real incomes.

To overcome the problem of lack of information on interesting parameters from time series or cross-sectional data alone, one frequently estimates some parameters from cross-sectional data, then introduces these estimates into time series regression to estimate other parameters of the model. For instance, Tobin (1950) calculated income elasticity from cross-sectional data, then multiplied it by the time series income variable and subtracted the product from the annual time series of quantity demand to form a new dependent variable. This new dependent-variable series was then regressed against the time series of the price variable to obtain an estimate of the price elasticity of demand.

The purpose of pooling here, as in the cases analyzed earlier, is to get more efficient estimates for the parameters that are of interest. In a time series, the number of observations is usually limited, and variables are highly correlated. Moreover, an aggregate data set, or a single individual time series data set does not contain information on micro-sociodemographic variables that affect economic behavior. Neither are cross-sectional data more structurally complete. Observations on individuals at one point in time are likely to be affected by prior observations. These raise two fundamental problems: One is that the source of estimation bias in cross-sectional estimates may be different from that in time series estimates. In fact, many people have questioned the suitability and comparability of estimates from different kinds of data (micro or aggregate, cross section or time series) (e.g., Kuh 1959; Kuh and Meyer 1957). The second is that if pooling is desirable, what is the optimal way to do it? It turns out that both problems can be approached simultaneously from the framework of an analysis of the likelihood functions (Maddala 1971b) or a Bayesian approach (Hsiao et al. 1995).

The likelihood function provides a useful way to extract the information contained in the sample provided that the model is correctly specified. Yet a model is a simplification of complex real-world phenomena. To be most useful, a model must strike a reasonable balance between realism and manageability. It should be realistic in incorporating the main elements of the phenomena being represented and at the same time be manageable in eliminating extraneous influences. Thus, when specifying a regression equation, it is common to assume that the numerous factors that affect the outcome of the dependent variable, but are individually unimportant or unobservable, can be appropriately summarized by a random disturbance term. However, the covariations of these omitted variables and the included explanatory variables in a cross-sectional regression may be different from those in a time series regression. For example,

if high income is associated with high consumption levels and is also correlated with age, the regression of consumption on income cross-sectionally will yield an income coefficient that measures the joint effects of age and income on consumption, unless age is introduced as another explanatory variable. But the age composition of the population could either be constant or be subject only to gradual, slow change in aggregate time series. Hence, the time series estimate of the income elasticity, ignoring the age variable, could be smaller than the cross-sectional estimates because of the negligible age-income correlation.

Another reason that cross-sectional and time series estimates in demand analysis may differ is that cross-sectional estimates tend to measure long-run behavior and time series estimates tend to measure short-run adjustment (Kuh 1959; Kuh and Meyer 1957). The assumption is that the majority of the observed families have enjoyed their present positions for some time, and the disequilibrium among households tends to be synchronized in response to common market forces and business cycles. Hence, many disequilibrium effects wash out (or appear in the regression intercept), so that the higher cross-sectional slope estimates may be interpreted as long-run coefficients. However, this will not be true for time series observations. Specifically, changes over time usually represent temporary shifts. Recipients or losers from this change probably will not adjust immediately to their new levels. A incompletely adjusted response will typically have a lower coefficient than the fully adjusted response.

These observations on differential cross-sectional and time series behavior suggest that the impacts of omitted variables can be strikingly different in time series and cross sections. Unless the assumption that the random term (representing the omitted-variables effect) is uncorrelated with the included explanatory variables holds, the time series and cross-sectional estimates of the common coefficients can diverge. In fact, if the time series and cross-sectional estimates differ, this is an indication that either or both models are misspecified. In Chapter 3 we discussed specification tests without using extraneous information. We now discuss a likelihood approach when extraneous information in the form of cross-sectional data for the time series model, or time series data for the cross-sectional model, is available.

### 11.3.2 The Likelihood Approach to Pooling Cross-Sectional and Time Series Data

Assume that we have a single cross section consisting of $N$ units and a time series extending over $T$ time periods. Suppose that the cross-sectional model is

$$\mathbf{y}_c = Z_1\boldsymbol{\delta}_1 + Z_2\boldsymbol{\delta}_2 + \mathbf{u}_c, \tag{11.3.1}$$

where $\mathbf{y}_c$ is an $N \times 1$ vector of observations on the dependent variable, $Z_1$ and $Z_2$ are $N \times K$ and $N \times L$ matrices of independent variables, and $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$ are

$K \times 1$ and $L \times 1$ vectors of parameters, respectively. The $N \times 1$ error term $\mathbf{u}_c$ is independently distributed, with variance–covariance matrix $\sigma_u^2 I_N$.

The time series model is

$$\mathbf{y}_T = X_1 \boldsymbol{\beta}_1 + X_2 \boldsymbol{\beta}_2 + \mathbf{v}_T. \tag{11.3.2}$$

where $\mathbf{y}_T$ is a $T \times 1$ vector of observations on the dependent variable, $X_1$ and $X_2$ are $T \times K$ and $T \times M$ matrices of observations on the independent variables, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are $K \times 1$ and $M \times 1$ vectors of parameters, and $\mathbf{v}_T$ is a $T \times 1$ vector of disturbances.[4] For simplicity, we assume that $\mathbf{v}_T$ is uncorrelated with $\mathbf{u}_c$ and is serially uncorrelated, with the variance–covariance matrix $E\mathbf{v}_T\mathbf{v}_T' = \sigma_v^2 I_T$.

The null hypothesis here is that $\boldsymbol{\delta}_1 = \boldsymbol{\beta}_1$. So with regard to the question whether or not to pool, we can use a likelihood-ratio test. Let $L_1^*$ and $L_2^*$ denote the maxima of the log joint likelihood functions for (11.3.1) and (11.3.2) with and without the restriction that $\delta_1 = \beta_1$. Then, under the null hypothesis, $2(L_2^* - L_1^*)$ is asymptotically $\chi^2$ distributed, with $K$ degrees of freedom. The only question is: What is the appropriate level of significance? If the costs of mistakenly accepting the pooling hypothesis and rejecting the pooling hypothesis are the same, Maddala (1971b) suggested using something like a 25 to 30 percent level of significance, rather than the conventional 5 percent, in our preliminary test of significance.

The specifications of the maximum-likelihood estimates and their variance–covariances merely summarize the likelihood function in terms of the location of its maximum and its curvature around the maximum. It is possible that the information that the likelihood function contains is not fully expressed by these. When the compatibility of cross-sectional and time series estimates is investigated, it is useful to plot the likelihood function extensively. For this purpose, Maddala (1971b) suggested that one should also tabulate and plot the relative maximum likelihoods of each data set,

$$R_M(\delta_1) = \frac{\underset{\boldsymbol{\theta}}{\text{Max}}\ L(\boldsymbol{\delta}_1, \boldsymbol{\theta})}{\underset{\boldsymbol{\delta}_1, \boldsymbol{\theta}}{\text{Max}}\ L(\boldsymbol{\delta}_1, \boldsymbol{\theta})}, \tag{11.3.3}$$

where $\boldsymbol{\theta}$ represents the set of nuisance parameters, $\max_\theta L(\boldsymbol{\delta}_1, \boldsymbol{\theta})$ denotes the maximum of $L$ with respect to $\boldsymbol{\theta}$, given $\boldsymbol{\delta}_1$ and $\max_{\boldsymbol{\delta}_1, \boldsymbol{\theta}} L(\boldsymbol{\delta}_1, \boldsymbol{\theta})$ denotes the maximum of $L$ with respect to both $\boldsymbol{\delta}_1$ and $\boldsymbol{\theta}$. The plot of (11.3.3) summarizes almost all the information contained in the data on $\boldsymbol{\delta}_1$. Hence, the shapes and locations of the relative maximum likelihoods will reveal more information

[4] If the cross-sectional data consist of all individuals in the population, then in the year in which cross-sectional observations are collected, the sum across individual observations of a variable should be equal to the corresponding aggregate time-series variable. Because in most cases cross-sectional samples consist of a small portion of the population, we shall ignore this relation and assume that the variables are unrelated.

about the compatibility of the different bodies of data than a single test statistic can.

If the hypothesis $\boldsymbol{\delta}_1 = \boldsymbol{\beta}_1$ is acceptable, then, as Chetty (1968), Durbin (1953), and Maddala (1971b) have suggested, we can stack (11.3.1) and (11.3.2) together as

$$\begin{bmatrix} \mathbf{y}_c \\ \mathbf{y}_t \end{bmatrix} = \begin{bmatrix} Z_1 \\ X_1 \end{bmatrix} \boldsymbol{\delta}_1 + \begin{bmatrix} Z_2 \\ \mathbf{0} \end{bmatrix} \boldsymbol{\delta}_2 + \begin{bmatrix} \mathbf{0} \\ X_2 \end{bmatrix} \boldsymbol{\beta}_2 + \begin{bmatrix} \mathbf{u}_c \\ \mathbf{v}_T \end{bmatrix}. \tag{11.3.4}$$

It is clear that an efficient method of estimating of $\boldsymbol{\delta}_1$, $\boldsymbol{\delta}_2$, and $\boldsymbol{\beta}_2$ is to apply the maximum-likelihood method to (11.3.4). An asymptotically equivalent procedure is to first apply least-squares separately to (11.3.1) and (11.3.2) to obtain consistent estimates of $\sigma_u^2$ and $\sigma_v^2$, and then substitute the estimated $\sigma_u^2$ and $\sigma_v^2$ into the equation

$$\begin{aligned} \begin{bmatrix} \frac{1}{\sigma_u} y_c \\ \frac{1}{\sigma_v} y_T \end{bmatrix} &= \begin{bmatrix} \frac{1}{\sigma_u} Z_1 \\ \frac{1}{\sigma_v} X_1 \end{bmatrix} \boldsymbol{\delta}_1 + \begin{bmatrix} \frac{1}{\sigma_u} Z_2 \\ \mathbf{0} \end{bmatrix} \boldsymbol{\delta}_2 \\ &\quad + \begin{bmatrix} \mathbf{0} \\ \frac{1}{\sigma_v} X_2 \end{bmatrix} \boldsymbol{\beta}_2 + \begin{bmatrix} \frac{1}{\sigma_u} \mathbf{u}_c \\ \frac{1}{\sigma_v} \mathbf{v}_T \end{bmatrix} \end{aligned} \tag{11.3.5}$$

and apply the least-squares method to (11.3.5).

The conventional procedure of substituting the cross-sectional estimates of $\boldsymbol{\beta}_1$, $\hat{\boldsymbol{\delta}}_{1c}$, into the time series model

$$\mathbf{y}_T - X_1 \hat{\boldsymbol{\delta}}_{1c} = X_2 \boldsymbol{\beta}_2 + \mathbf{v}_T + X_1 (\boldsymbol{\beta}_1 - \hat{\boldsymbol{\delta}}_{1c}), \tag{11.3.6}$$

and then regressing $(\mathbf{y}_T - X_1 \hat{\boldsymbol{\delta}}_{1c})$ on $X_2$, yields only conditional estimates of the parameters $\boldsymbol{\beta}_2$ – conditional on the estimates obtained from the cross-sectional data.[5] However, there is also some information about $\boldsymbol{\beta}_1$ in the time series sample, and this should be utilized. Moreover, one should be careful in the use of two-step procedures. Proper evaluation of the asymptotic variance–covariance matrix of $\boldsymbol{\beta}_2$ should take account of the uncertainty (variance) in substituting $\hat{\boldsymbol{\delta}}_{1c}$ for $\boldsymbol{\beta}_1$. (For details, see Chetty 1968; Hsiao et al. 1995; Jeong 1978; and Maddala 1971b.)

[5] In the Bayesian framework this is analogous to making inferences based on the conditional distribution of $\boldsymbol{\beta}_2$, $f(\boldsymbol{\beta}_2 \mid \boldsymbol{\beta}_1 = \boldsymbol{\delta}_{1c})$, whereas it is the marginal distribution of $\boldsymbol{\beta}_2$ that should be used whenever $\boldsymbol{\beta}_1$ is not known with certainty. For details see Chetty (1968).

### 11.3.3 An Example

To illustrate application of the likelihood approach to pooling, Maddala (1971b) analyzed a simple econometric model relating to the demand for food in the United States. The model and the data were taken from Tobin (1950).

The cross-sectional demand equation is

$$y_{1i} = \delta_0 + \delta_1 z_{1i} + \delta_2 z_{2i} + u_i, \qquad i = 1, \ldots, N, \tag{11.3.7}$$

where $y_{1i}$ is the logarithm of the average food consumption of the group of families at a point in time, and $z_{1i}$ and $z_{2i}$ are the logarithms of the average income of the $i$th family and the average family size, respectively. The time series demand function is

$$y_{2t} = \beta_0 + \beta_1(x_{1t} - \beta_2 x_{2t}) + \beta_3(x_{2t} - x_{2,t-1}) + v_t, \qquad t = 1, \ldots, T. \tag{11.3.8}$$

where $y_{2t}$, $x_{1t}$, and $x_{2t}$ are the logarithms of the food price index, per capita food supply for domestic consumption, and per capita disposable income, respectively. The income elasticity of demand, $\delta_1$, was assumed common to both regressions, namely, $\delta_1 = \beta_2$. The error terms $u_i$ and $v_t$ were independent of each other and were assumed independently normally distributed, with 0 means and constant variances $\sigma_u^2$ and $\sigma_v^2$, respectively.

The results of the cross-sectional estimates are

$$\begin{matrix} \hat{y}_{1i} = 0.569 + 0.5611 z_{1i} + 0.2540 z_{2i} \\ \qquad\qquad\quad (0.0297) + (0.0367) \end{matrix}, \tag{11.3.9}$$

where standard errors are in parentheses. The results of the time series regression are

$$\begin{matrix} \hat{y}_{2t} = 7.231 + 1.144 x_{2t} - 0.1519(x_{2t} - x_{2,t-1}) - 3.644 x_{1t} \\ \qquad\qquad\quad (0.0612) \quad (0.0906) \qquad\qquad (0.4010). \end{matrix} \tag{11.3.10}$$

The implied income elasticity, $\delta_1$, is 0.314.

When the cross-sectional estimate of $\delta_1$, 0.56, is introduced into the time series regression, the estimated $\beta_1$ is reduced to $-1.863$, with a standard error of 0.1358. When $\delta_1$ and $\beta_1$ are estimated simultaneously by the maximum-likelihood method, the estimated $\delta_1$ and $\beta_1$ are 0.5355 and $-1.64$, with a covariance $\begin{bmatrix} 0.00206 & 0.00827 \\ & 0.04245 \end{bmatrix}$.

Although there is substantial improvement in the accuracy of the estimated coefficient using the combined data, the likelihood-ratio statistic turns out to be 17.2, which is significant at the 0.001 level with 1 degree of freedom. It strongly suggests that in this case we should not pool the time series and cross-sectional data.

Figure 11.1 reproduces Maddala's plot of the relative maximum likelihood $R_M(\delta_1)$ for the parameter $\delta_1$ (the income elasticity of demand) in the Tobin

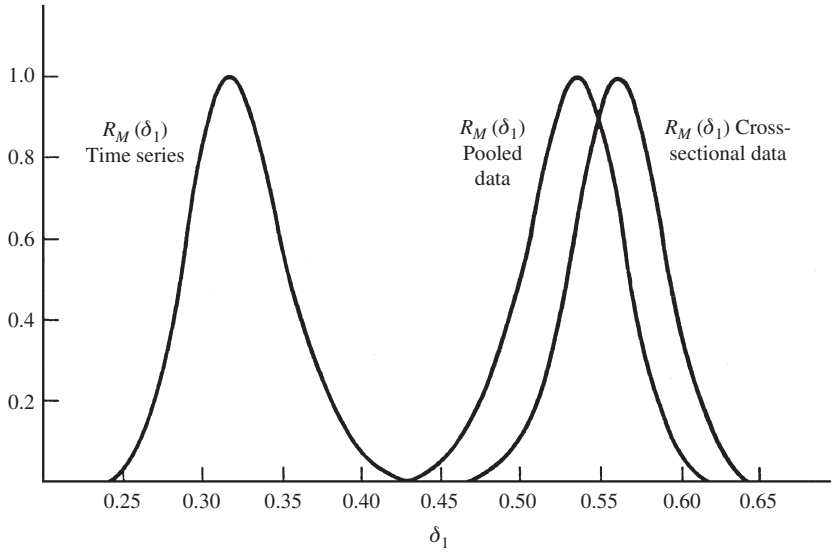


Figure 11.1. Relative maximum likelihood for the parameter $\delta_1$. *Source:* Maddala (1971b, fig. 1).

model from cross-sectional data alone, from time series data alone, and from the pooled sample. The figure reveals that the information on $\delta_1$ provided by the time series data is almost as precise as that provided by the cross-sectional data (otherwise the likelihood function would be relatively flat). Furthermore, there is very little overlap between the likelihood functions from time series and cross-sectional data. Again, this unambiguously suggests that the data should not be pooled.[6]

Given that the time series data arise by aggregating some microeconomic process, there cannot possibly be a conflict between the time series and cross-sectional inferences if individual differences conditional on explanatory variables are viewed as chance outcomes. Thus, whenever the empirical results differ systematically between the two, as in the foregoing example, this is an indication that either or both models may be misspecified. The existence of supporting extraneous information in the form of cross-sectional or time series data provides an additional check to the appropriateness of a model specification that cannot be provided by a single cross-sectional or time series data set, because there may be no internal evidence of this omitted-variable bias. However, until a great deal is learned about the cross-sectional time series relation,

[6] It should be noted that the foregoing result is based on the assumption that both $u_i$ and $v_t$ are independently normally distributed. In practice, careful diagnostic checks should be performed before exploring the pooling issue, using the likelihood-ratio test or relative maximum likelihoods. In fact, Izan (1980) redid the analysis by allowing $v_t$ to follow a first-order autoregressive process. The likelihood-ratio test after allowing for autocorrelation resulted in accepting the pooling hypothesis.

there appears no substitute for the completeness of information. Sequential observations on a number of individuals or panel data are essential for a full understanding of the systematic interrelations at different periods of time.

## 11.4 ESTIMATING DISTRIBUTED LAGS IN SHORT PANELS[7]

### 11.4.1 Introduction

Because of technical, institutional, and psychological rigidities, often behavior is not adapted immediately to changes in the variables that condition it. In most cases this adaptation is progressive. The progressive nature of adaptations in behavior can be expressed in various ways. Depending on the rationale behind it, we can set up an autoregressive model with the current value of $y$ being a function of lagged dependent variables and exogenous variables, or we can set up a distributed-lag model, with the current value of $y$ being a function of current and previous values of exogenous variables. Although usually a linear distributed-lag model can be expressed in an autoregressive form, and, similarly, as a rule any stable linear autoregressive model can be transformed into a distributed-lag model,[8] the empirical determination of time lags is very important in applied economics. The roles of many economic measures can be correctly understood only if we know when they will begin to take effect and when their effects will be fully worked out. Therefore, we would like to distinguish these two types of dynamic models when a precise specification (or reasoning) is possible. In Chapter 4 we discussed the issues of estimating autoregressive models with panel data. In this section we discuss estimation of distributed-lag models (Pakes and Griliches 1984).

A general distributed-lag model for a single time series of observations is usually written as

$$y_t = \mu + \sum_{\tau=0}^{\infty} \beta_\tau x_{t-\tau} + u_t, \quad t = 1, \ldots, T, \tag{11.4.1}$$

where, for simplicity, we assume that there is only one exogenous variable, $x$, and, conditional on $\{x_t\}$, the $u_t$ are independent draws from a common distribution function. When no restrictions are imposed on the lag coefficients, one cannot obtain consistent estimates of $\beta_\tau$ even when $T \to \infty$, because the number of unknown parameters increases with the number of observations. Moreover, the available samples often consist of fairly short time series on variables that are highly correlated over time. There is not sufficient information to obtain precise estimates of any of the lag coefficients without specifying, a

[7] The material in this section is adapted from Pakes and Griliches (1984) with permission.

[8] We must point out that the errors are also transformed when we go from one form to the other (e.g., Malinvaud 1970, Chapter 15).

priori, that all of them are functions of only a very small number of parameters (Koyck lag, Almon lag, etc.) (Dhrymes 1971; Malinvaud 1970).

On the other hand, when there are $N$ time series, we can use cross-sectional information to identify and estimate (at least some of the) lag coefficients without having to specify a priori that the sequence of lag coefficients progresses in a particular way. For instance, consider the problem of using panel data to estimate the model (11.4.1), which for a given $t$ we rewrite as

$$y_{it} = \mu + \sum_{\tau=0}^{t-1} \beta_\tau x_{i,t-\tau} + b_{it} + u_{it}, \qquad i = 1, \ldots, N, \tag{11.4.2}$$

where

$$b_{it} = \sum_{\tau=0}^{\infty} \beta_{t+\tau} x_{i,-\tau} \tag{11.4.3}$$

is the contribution of the unobserved presample $x$ values to the current values of $y$, to which we shall refer as the truncation remainder. Under certain assumptions about the relationships between the unobserved $b_{it}$ and the observed $x_{it}$, it is possible to obtain consistent estimates of $\beta_\tau$, $\tau = 0, \ldots, t-1$, by regressing (11.4.2) cross-sectionally. Furthermore, the problem of collinearity among $x_t, x_{t-1}, \ldots,$ in a single time series can be reduced or avoided by use of the cross-sectional differences in individual characteristics.

### 11.4.2 Common Assumptions

To see under what conditions the addition of a cross-sectional dimension can provide information that cannot be obtained in a single time series, first we note that if the lag coefficients vary across individuals $\{\beta_{i\tau}\}_{\tau=0}^{\infty}$, for $i = 1, \ldots, N$, and if there is no restriction on the distribution of these sequences over members of the population, each time series contains information on only a single sequence of coefficients. The problem of lack of information remains for panel data. Second, even if the lag coefficients do not vary across individuals ($\beta_{i\tau} = \beta_\tau$ for $i = 1, \ldots, N$ and $\tau = 0, 1, 2, \ldots$), the (often very significant) increase in sample size that accompanies the availability of panel data is entirely an increase in cross-sectional dimension. Panel data sets, in fact, usually track their observations over only a relatively short time interval. As a result, the contributions of the unobserved presample $x$ values to the current values of $y$ (the truncation remainder, $b_{it}$) are likely to be particularly important if we do not wish to impose the same type of restrictions on the lag coefficients as we often do when a single time-series data set is used to estimate a distributed-lag model. Regression analysis, ignoring the unobserved truncation-remainder term, will suffer from the usual omitted-variable bias.

Thus, to combine $N$ time series to estimate a distributed-lag model, we have to impose restrictions on the distribution of lag coefficients across cross-sectional units and/or on the way the unobserved presample terms affect current

behavior. Pakes and Griliches (1984) considered a distributed-lag model of the form

$$y_{it} = \alpha_i^* + \sum_{\tau=0}^{\infty} \beta_{i\tau} x_{i,t-\tau} + u_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{matrix} \tag{11.4.4}$$

where $u_{it}$ is independent of $x_{is}$ and is independently, identically distributed, with mean zero and variance $\sigma_u^2$. The coefficients of $\alpha_i^*$ and $\beta_{i\tau}$ are assumed to satisfy the following assumptions.

**Assumption 11.4.1:** $E(\beta_{i\tau}) = \beta_\tau$.

**Assumption 11.4.2:** Let $\bar{\beta}_{i\tau} = \beta_{i\tau} - \beta_\tau$, $\xi_{it} = \sum_{\tau=0}^{\infty} \bar{\beta}_{i\tau} x_{i,t-\tau}$, and $\boldsymbol{\xi}_i' = (\xi_{i1}, \ldots, \xi_{iT})$; then $E^*[\boldsymbol{\xi}_i \mid \mathbf{x}_i] = \mathbf{0}$.

**Assumption 11.4.3:** $E^*(\alpha_i^* \mid \mathbf{x}_i) = \mu + \mathbf{a}'\mathbf{x}_i$

Here $E^*(Z_1 \mid Z_2)$ refers to the minimum mean-square-error linear predictor (or the projection) of $Z_1$ onto $Z_2$; $\mathbf{x}_i$ denotes the vector of all observed $\mathbf{x}_{it}$. We assume that there are $\ell + 1$ observations on $x$ before the first observation on $y$, and the $1 \times (\ell + 1 + T)$ vector $\mathbf{x}_i' = [x_{i,-\ell}, \ldots, x_{iT}]$ is an independent draw from a common distribution with $E(\mathbf{x}_i \mathbf{x}_i') = \sum_{xx}$ positive definite.[9]

A sufficient condition for Assumption 11.4.2 to hold is that differences in lag coefficients across individuals are uncorrelated with the $\mathbf{x}_i$ [i.e., $\beta_{i\tau}$ is a random variable defined in the sense of Swamy (1970), or see Chapter 6]. However, Assumption 11.4.3 does allow for individual-specific constant terms (the $\alpha_i^*$) to be correlated with $\mathbf{x}_i$. The combination of Assumptions 11.4.1–11.4.3 is sufficient to allow us to identify the expected value of the lag-coefficient sequence $\{\beta_\tau\}$ if both $N$ and $T$ tend to infinity.

If $T$ is fixed, substituting Assumptions 11.4.1 and 11.4.2 into equation (11.4.4), we rewrite the distributed-lag model as

$$y_{it} = \alpha_i^* + \sum_{\tau=0}^{t+\ell} \beta_\tau x_{i,t-\tau} + b_{it} + \tilde{u}_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{matrix} \tag{11.4.5}$$

where $b_{it} = \sum_{\tau=\ell+1}^{\infty} \beta_{t+\tau} x_{i,-\tau}$ is the truncation remainder for individual $i$ in period $t$, and $\tilde{u}_{it} = \xi_{it} + u_{it}$ is the amalgamated error term satisfying $E^*[\tilde{\mathbf{u}}_{it} \mid \mathbf{x}_i] = \mathbf{0}$. The unobserved truncation remainders are usually correlated with the included explanatory variables. Therefore, without additional restrictions, we still cannot get consistent estimates of any of the lag coefficients $\beta_\tau$ by regressing $y_{it}$ on $x_{i,t-\tau}$, even when $N \to \infty$.

[9] Note that assuming that there exist $\ell + 1$ observations on $x$ before the first observation on $y$ is not restrictive. If $x_{it}$ does not exist before time period 0, we can always let $\ell = -1$. If $\ell$ has to be fixed, we can throw away the first $\ell + 1$ observations of $y$.

Because the values of the truncation remainders $b_{it}$ are determined by the lag coefficients and the presample $x$ values, identification requires constraints either on the lag coefficients or on the stochastic process generating these $x$ values. Because there usually are many more degrees of freedom available in panel data, this allows us to use prior restrictions of different kind than in the usual approach of constraining lag coefficients to identify truncation remainders (e.g., Dhrymes 1971). In the next two subsections we illustrate how various restrictions can be used to identify the lag coefficients.

### 11.4.3 Identification Using Prior Structure on the Process of the Exogenous Variable

In this subsection we consider the identification of a distributed-lag model using a kind of restriction different from that in the usual approach of constraining lag coefficients. Our interest is focused on estimating at least some of the population parameters $\beta_\tau = E(\beta_{i\tau})$ for $\tau = 0, 1, \ldots,$ without restricting $\beta_\tau$ to be a function of a small number of parameters. We consider a lag coefficient identified if it can be calculated from the matrix of coefficients obtained from the projection of $\mathbf{y}_i$ onto $\mathbf{x}_i$, a $T \times (T + \ell + 1)$ matrix labeled $\Pi$, where $E^*(\mathbf{y}_i \mid \mathbf{x}_i) = \boldsymbol{\mu}^* + \Pi\mathbf{x}_i$, $\boldsymbol{\mu}^* = (\mu_1^*, \ldots, \mu_T^*)'$ and $\mathbf{y}_i' = (y_{i_1}, \ldots, y_{iT})$ is a $1 \times T$ vector.

Equation (11.4.5) makes it clear that each row of $\Pi$ will contain a combination of the lag coefficients of interest and the coefficients from the projections of the two unobserved components, $\alpha_i^*$ and $b_{it}$, on $\mathbf{x}_i$. Therefore, the problem is to separate out the lag coefficients from the coefficients defining these two projections.

Using equation (11.4.5), the projection of $\mathbf{y}_i$ onto $\mathbf{x}_i$ and $\alpha_i^*$ is given by[10]

$$E^*(\mathbf{y}_i \mid \mathbf{x}_i, \alpha_i^*) = [B + W]\mathbf{x}_i + [\mathbf{e} + \mathbf{c}]\alpha_i^* \tag{11.4.6}$$

where $B$ is the $T \times (T + \ell + 1)$ matrix of the lag coefficients

$$B = \begin{bmatrix} \beta_{\ell+1} & \beta_\ell & \cdot & \beta_1 & \beta_0 & 0 & \cdot & \cdot & \cdot & 0 \\ \beta_{\ell+2} & \beta_{\ell+1} & \cdot & \beta_2 & \beta_1 & \beta_0 & 0 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \beta_{T+\ell-1} & \beta_{t+\ell-2} & \cdot & \beta_{T+1} & \beta_T & \cdot & \cdot & \cdot & \beta_0 & 0 \\ \beta_{T+\ell} & \beta_{T+\ell-1} & \cdot & \beta_T & \beta_{T-1} & \cdot & \cdot & \cdot & \beta_1 & \beta_0 \end{bmatrix}$$

$W$ and $\mathbf{c}$ are defined by the unconstrained projection of $\mathbf{b}_i = (b_{i1}, \ldots, b_{iT})'$ onto $\mathbf{x}_i$ and $\alpha_i^*$,

$$E^*[\mathbf{b}_i \mid \mathbf{x}_i, \alpha_i^*] = W\mathbf{x}_i + \mathbf{c}\alpha_i^*. \tag{11.4.7}$$

[10] Note that we allow the projection of presample $x_{i,-\tau}$ on in-sample $\mathbf{x}_i$ and $\alpha_i^*$ to depend freely on the $\alpha_i^*$ by permitting each element of the vector $\mathbf{c}$ to be different.

Equation (11.4.6) and the fact that $E^*\{E^*(\mathbf{y}_i \mid \mathbf{x}_i, \alpha_i^*) \mid \mathbf{x}_i\} = E^*[\mathbf{y}_i \mid \mathbf{x}_i] = (\mathbf{e} + \mathbf{c})\mu + \Pi\mathbf{x}_i$ imply that

$$\Pi = B + [W + (\mathbf{e} + \mathbf{c})\mathbf{a}']. \tag{11.4.8}$$

where $\mathbf{a}$ is defined by the unconstrained projection of $\alpha_i^*$ onto $\mathbf{x}_i$, $[E^*(\alpha_i^* \mid \mathbf{x}_i) = \mu + \mathbf{a}'\mathbf{x}_i]$.

Clearly, if the $T \times (T + \ell + 1)$ matrix $W$ is unrestricted, we cannot separate out the lag coefficients, $B$, and the impact of the truncation-remainder term from the $\Pi$ matrix. But given that $\mathbf{ca}'$ is a matrix of rank 1, we may be able to identify some elements of $B$ if there are restrictions on $W$. Thus, to identify some of the lag coefficients from $\Pi$, we shall have to restrict $W$. $W$ will be restricted if it is reasonable to assume that the stochastic process generating $\{x_{it}\}_{t=-\infty}^{T}$ restricts the coefficients on $\mathbf{x}_i$ in the projection of the presample $x_{i,-j}$ values onto the in-sample $\mathbf{x}_i$ and $\alpha_i^*$. The particular case analyzed by Pakes and Griliches (1984) is given by the following assumption.[11]

**Assumption 11.4.4:** For $q \geq 1$, $E^*[x_{i,-\ell-q} \mid \mathbf{x}_i,\ \alpha_i^*] = c_q\alpha_i^* + \Sigma_{j=1}^{p}\rho_j^{(q)}$ $x_{i,-\ell+j-1}$. That is, in the projection of the unseen presample $x$ values onto $\mathbf{x}_i$ and $\alpha_i^*$, only $[x_{i,-\ell}, x_{i,-\ell+1}, \ldots, x_{i,-\ell+p-1}]$ have nonzero coefficients.

If $c_q = 0$, a sufficient condition for Assumption 11.4.4 to hold is that $x$ is generated by a $p$th-order autoregressive process.[12]

Because each element of $\mathbf{b}_i$ is just a different linear combination of the same presample $x$ values, the addition of Assumption 11.4.4 implies that

$$E^*[b_{it} \mid \mathbf{x}_i, \alpha_i^*] = c_t\alpha_i^* + \sum_{j=1}^{p} w_{t,j-\ell-1}x_{i,j-\ell-1}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{matrix} \tag{11.4.9}$$

where $w_{t,j-\ell-1} = \Sigma_{q=1}^{\infty}\beta_{t+\ell+q}\rho_j^{(q)}$, $j = 1, \ldots, p$, and $c_t = \Sigma_{q=1}^{\infty}\beta_{t+l+q}c_q$. This determines the vector $\mathbf{c}$ and the matrix $W$ in (11.4.7). In particular, it implies that $W$ can be partitioned into a $T \times (T + \ell - p + 1)$ matrix of zeros and $T \times p$ matrix of free coefficients,

$$W = \left[\begin{matrix} \tilde{W} \\ T \times p \end{matrix} \vdots \begin{matrix} \mathbf{0} \\ T \times (T + \ell - p + 1). \end{matrix}\right]. \tag{11.4.10}$$

Substituting (11.4.10) into (11.4.8) and taking partial derivatives of $\Pi$ with respect to the leading $(T + \ell - p + 1)$ lag coefficients, we can show that the resulting Jacobian matrix satisfies the rank condition for identification of these coefficients (e.g., Hsiao 1983, Theorem 5.1.2). A simple way to check that

[11] One can use various model-selection criteria to determine $p$ (e.g., Amemiya 1980a).

[12] We note that $c_q = 0$ implies that $\alpha_i^*$ is uncorrelated with presample $x_i$.

the leading $(T + \ell - p + 1)$ lag coefficients are indeed identified is to show that consistent estimators for them exist. We note that by construction, cross-sectional regression of $\mathbf{y}_i$ on $\mathbf{x}_i$ yields consistent estimates of $\Pi$. For the special case in which $c_q = 0$, the projections of each period's value of $y_{it}$ on all in-sample values of $\mathbf{x}_i' = (x_{i,-\ell}, x_{i,-\ell+1}, \ldots, x_{iT})$ are[13]

$$
\begin{aligned}
E^*(y_{i1} \mid \mathbf{x}_i) &= \mu + \sum_{j=1}^{p} \phi_{1,j-\ell-1} x_{i,j-\ell-1}, \\
E^*(y_{i2} \mid \mathbf{x}_i) &= \mu + \beta_0 x_{i2} + \sum_{j=1}^{p} \phi_{2,j-\ell-1} x_{i,j-\ell-1}, \\
E^*(y_{i3} \mid \mathbf{x}_i) &= \mu + \beta_0 x_{i3} + \beta_1 x_{i2} + \sum_{j=1}^{p} \phi_{3,j-\ell-1} x_{i,j-\ell-1} \\
&\vdots \\
E^*(y_{iT} \mid \mathbf{x}_i) &= \mu + \beta_0 x_{iT} + \cdots + \beta_{T+\ell-p} x_{i,p-\ell} + \sum_{j=1}^{p} \phi_{T,j-\ell-1} x_{i,j-\ell-1},
\end{aligned}
\tag{11.4.11}
$$

where $\phi_{t,j-\ell-1} = \beta_{t+\ell+1-j} + w_{t,j-\ell-1}$ for $t = 1, \ldots, T$, and $j = 1, \ldots, p$, and for simplicity we have let $p = \ell + 2$. The first $p$ values of $\mathbf{x}_i$ in each projection have nonzero partial correlations with the truncation remainders (the $b_{it}$). Hence, their coefficients do not identify the parameters of the lag distribution. Only when $(t + \ell - p + 1) > 0$, the leading coefficients in each equation are, in fact, estimates of the leading lag coefficients. As $t$ increases, we gradually uncover the lag structure.

When $c_q \neq 0$, the finding of consistent estimators (hence identification) for the leading $(T + \ell - p + 1)$ lag coefficients is slightly more complicated. Substituting (11.4.9) into (11.4.5), we have

$$
\begin{aligned}
E^*\left(y_{it} \mid \mathbf{x}_i, \alpha_i^*\right) = (1 + c_t)\alpha_i^* &+ \sum_{\tau=0}^{t+\ell-p} \beta_\tau x_{i,t-\tau} \\
&+ \sum_{j=1}^{p} \phi_{t,j-\ell-1} x_{i,j-\ell-1}, \qquad t = 1, \ldots, T,
\end{aligned}
\tag{11.4.12}
$$

where again (for simplicity) we have assumed $p = \ell + 2$. Conditioning this equation on $\mathbf{x}_i$, and passing through the projection operator once more,

[13] The coefficient of (11.4.11) is another way of writing $\Pi$ (11.4.8).

we obtain

$$
\begin{aligned}
E^*(y_{i1} \mid \mathbf{x}_i) &= \mu(1+c_1) + (1+c_1)\sum_{t=p-\ell}^{T} a_t x_{it} \\
&\quad + \sum_{j=1}^{p}[(1+c_1)a_{j-\ell-1} + \phi_{1,j-\ell-1}]x_{i,j-\ell-1}, \\
E^*(y_{i2} \mid \mathbf{x}_i) &= \mu(1+c_2) + \beta_0 x_{i2} + (1+c_2)\sum_{t=p-\ell}^{T} a_t x_{it} \\
&\quad + \sum_{j=1}^{p}[(1+c_2)a_{j-\ell-1} + \phi_{2,j-\ell-1}]x_{i,j-\ell-1}, \qquad (11.4.13) \\
&\vdots \\
E^*(y_{iT} \mid \mathbf{x}_i) &= \mu(1+c_T) + \sum_{\tau=0}^{T+\ell-p} \beta_\tau x_{i,T-\tau} + (1+c_T)\sum_{t=p-\ell}^{T} a_t x_{it} \\
&\quad + \sum_{j=1}^{p}[(1+c_T)a_{j-\ell-1} + \phi_{T,j-\ell-1}]x_{i,j-\ell-1}.
\end{aligned}
$$

Multiplying $y_{i1}$ by $\tilde{c}_t$ and subtracting it from $y_{it}$, we produce the system of equations

$$
y_{it} = \tilde{c}_t y_{i1} + \sum_{\tau=0}^{t+\ell-p} \beta_\tau x_{i,t-\tau} + \sum_{j=1}^{p} \tilde{\phi}_{t,j-\ell-1} x_{i,j-\ell-1} + \nu_{it}, \qquad (11.4.14)
$$

for $t = 2, \ldots, T$, where

$$
\tilde{c}_t = \frac{(1+c_t)}{1+c_1}, \quad \tilde{\phi}_{t,j-\ell-1} = \phi_{t,j-\ell-1} - \tilde{c}_t \phi_{1,j-\ell-1},
$$

and

$$
\nu_{it} = y_{it} - \tilde{c}_t y_{i1} - E^*(y_{it} - \tilde{c}_t y_{i1} \mid \mathbf{x}_i).
$$

By construction, $E^*(\nu_{it} \mid \mathbf{x}_i) = 0$.

For given $t$, the only variable on the right-hand side of (11.4.14) that is correlated with $\nu_{it}$ is $y_{i1}$. If we know the values of $\{\tilde{c}_t\}_{t=2}^{T}$, the system (11.4.14) will allow us to estimate the leading $(T + \ell - p + 1)$ lag coefficients consistently by first forming $\tilde{y}_{it} = y_{it} - \tilde{c}_t y_{i1}$ (for $t = 2, \ldots, T$) and then regressing this sequence on in-sample $x_{it}$ values cross-sectionally. In the case in which all $c_t$ values are identical, we know that the sequence $\{\tilde{c}_t\}_{t=2}^{T}$ is just a sequence of 1's. In the case in which $\alpha_i^*$ have a free coefficient in each period of the sample, we have unknown $(1 + c_t)$. However, we can consistently estimate $\tilde{c}_t$, $\beta_\tau$, and $\tilde{\phi}_{t,j}$

by the instrumental-variable method, provided there is at least one $x_{is}$ that is excluded from the determinants of $y_{it} - \tilde{c}_t y_{i1}$ and that is correlated with $y_{i1}$. If $T \geq 3$, $x_{i3}, \ldots, x_{iT}$ are excluded from the equation determining $(y_{i2} - \tilde{c}_2 y_{i1})$, and provided that not all of $a_3$ to $a_T$ are 0, at least one of them will have the required correlation with $y_{i1}$.

We have shown that under Assumptions 11.4.1–11.4.4, the use of panel data allows us to identify the leading $T + \ell - p + 1$ lag coefficients without imposing any restrictions on the sequence $\{\beta_\tau\}_{\tau=0}^{\infty}$. Of course, if $T + \ell$ is small relative to $p$, we will not be able to build up much information on the tail of the lag distribution. This simply reflects the fact that short panels, by their very nature, do not contain unconstrained information on that tail. However, the early coefficients are often of significant interest in themselves. Moreover, they may provide a basis for restricting the lag structure (to be a function of a small number of parameters) in further work.

### 11.4.4 Identification Using Prior Structure on the Lag Coefficients

In many situations we may know that all $\beta_\tau$ are positive. We may also know that the first few coefficients $\beta_0$, $\beta_1$, and $\beta_2$ are the largest and that $\beta_\tau$ decreases with $\tau$ at least after a certain value of $\tau$. In this subsection we show how the conventional approach of constraining the lag coefficients to be a function of a finite number of parameters can be used and generalized for identification of a distributed-lag model in the panel data context. Therefore, we drop Assumption 11.4.4. Instead, we assume that we have prior knowledge of the structure of lag coefficients. The particular example we use here is the one assumed by Pakes and Griliches (1984), where the sequence of lag coefficients, after the first few free lags, has an autoregressive structure. This restriction is formalized as follows.

**Assumption 11.4.5:**

$$\beta_\tau = \begin{cases} \beta_\tau, & \text{for } \tau \leq k_1, \\ \sum_{j=1}^{J} \delta_j \beta_{\tau-j}, & \text{otherwise,} \end{cases}$$

where the roots of the characteristic equation $1 - \sum_{j=1}^{J} \delta_j L^j = 0$, say, $\lambda_1^{-1}, \ldots, \lambda_J^{-1}$, lie outside the unit circle.[14] For simplicity, we assume that $k_1 = \ell + 1$, and that $\lambda_1, \ldots, \lambda_J$ are real and distinct.

Assumption 11.4.5 implies that $\beta_\tau$ declines geometrically after the first $k_1$ lags. Solving the $J$th-order difference equation

$$\beta_\tau - \delta_1 \beta_{\tau-1} - \cdots - \delta_J \beta_{\tau-J} = 0, \tag{11.4.15}$$

[14] The condition for the roots of the characteristics equation to lie outside the unit circle is to ensure that $\boldsymbol{\beta}_\tau$ declines geometrically as $\tau \to \infty$ (e.g., Anderson 1971, their Chapter 5), so that the truncation remember term will stay finite for any reasonable assumption on the $x$ sequence.

we obtain the general solution (e.g., Box and Jenkins 1970, Chapter 3)

$$\beta_\tau = \sum_{j=1}^{J} A_j \lambda_j^\tau, \tag{11.4.16}$$

where $A_j$ are constants to be determined by the initial conditions of the difference equation.

Substituting (11.4.16) into (11.4.5), we write the truncation-remainder term $b_{it}$ as

$$\begin{aligned} b_{it} &= \sum_{\tau=\ell+1}^{\infty} \left( \sum_{j=1}^{J} A_j \lambda_j^{t+\tau} \right) x_{i,-\tau} \\ &= \sum_{j=1}^{J} \lambda_j^t \left( A_j \sum_{\tau=\ell+1}^{\infty} \lambda_j^\tau x_{i,-\tau} \right) \\ &= \sum_{j=1}^{J} \lambda_j^t b_{ij}, \end{aligned} \tag{11.4.17}$$

where $b_{ij} = A_j \sum_{\tau=\ell+1}^{\infty} \lambda_j^\tau x_{i,-\tau}$. That is, we can represent the truncation remainder $b_{it}$ in terms of $J$ unobserved initial conditions $(b_{i1}, \ldots, b_{iJ})$. Thus, under Assumptions 11.4.1–11.4.3 and 11.4.5, the distributed-lag model becomes a system of $T$ regressions with $J + 1$ freely correlated unobserved factors $(\alpha_i^*, b_{i1}, \ldots, b_{iJ})$ with $J$ of them decaying geometrically over time.

Because the conditions for identification of a model in which there are $J + 1$ unobserved factors is a straightforward generalization from a model with two unobserved factors, we deal first with the case $J = 1$ and then point out the extensions required for $J > 1$.

When $J = 1$, it is the familiar case of a modified Koyck (or geometric) lag model. The truncation remainder becomes an unobserved factor that follows an exact first-order autoregression (i.e., $b_{it} = \delta b_{i,t-1}$). Substituting this result into (11.4.5), we have

$$y_{it} = \alpha_i^* + \sum_{\tau=0}^{\ell+1} \beta_\tau x_{i,t-\tau} + \beta_{\ell+1} \sum_{\tau=\ell+2}^{t+\ell} \delta^{\tau-(\ell+1)} x_{i,t-\tau} + \delta^{t-1} b_i + \tilde{u}_{it}, \tag{11.4.18}$$

where, $b_i = \beta_{\ell+1} \sum_{\tau=1}^{\infty} \delta^\tau x_{i,-\tau-\ell}$.

Recall from the discussion in Section 11.4.3 that to identify the lag parameters we require a set of restrictions on the projection matrix $E^*(\mathbf{b}_i \mid \mathbf{x}_i) = [W + \mathbf{c}\mathbf{a}']\mathbf{x}_i$ [equation (11.4.7)]. The Koyck lag model implies that $b_{it} = \delta b_{i,t-1}$, which implies that $E^*(b_{it} \mid \mathbf{x}_i) = \delta E^*(b_{i,t-1} \mid \mathbf{x}_i)$; that is, $w_{tr} = \delta w_{t-1,r}$ for $r = 1, \ldots, T + \ell + 1$ and $t = 2, \ldots, T$. It follows that the $\Pi$ matrix has

the form

$$\Pi = B^* + \boldsymbol{\delta}^* \mathbf{w}^{*'} + \mathbf{e}\mathbf{a}', \tag{11.4.19}$$

where $\boldsymbol{\delta}^{*'} = [1, \delta, \ldots, \delta^{T-1}]$, $\mathbf{w}^*$ is the vector of coefficients from the projection of $b_i$ on $\mathbf{x}_i$ [i.e., $E^*(b_i \mid \mathbf{x}_i) = \sum_{t=-\ell}^{T} w_t^* x_{it}$], and

$$B^* = \begin{bmatrix} \beta_{\ell+1} & . & . & \beta_1 & \beta_0 & 0 \\ \delta\beta_{\ell+1} & . & . & \beta_2 & \beta_1 & \beta_0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \delta^{T-1}\beta_{\ell+1} & . & . & . & \delta^{T-\ell-1}\beta_{\ell+1} & \delta^{T-\ell-2}\beta_{\ell+1} \end{bmatrix}$$

$$\begin{bmatrix} . & . & . & . & 0 & 0 \\ . & . & . & . & 0 & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & \delta\beta_{\ell+1} & \beta_{\ell+1} & . & \beta_1 & \beta_0 \end{bmatrix}.$$

Taking partial derivatives of (11.4.19) with respect to unknown parameters, it can be shown that the resulting Jacobian matrix satisfies the rank condition for identification of the lag coefficients, provided $T \geq 3$ (e.g., Hsiao 1983, Theorem 5.1.2). In fact, an easy way to see that the lag coefficients are identified is to note that (11.4.18) implies that

$$\begin{aligned} (y_{it} - y_{i,t-1}) - \delta(y_{i,t-1} - y_{i,t-2}) &= \beta_0 x_{it} + [\beta_1 - \beta_0(1+\delta)]x_{i,t-1} \\ &+ \sum_{\tau=2}^{\ell} [\beta_\tau - (1+\delta)\beta_{\tau-1} + \delta\beta_{\tau-2}]x_{i,t-\tau} + v_{it}, \\ i &= 1, \ldots, N, \\ t &= 1, \ldots, T, \end{aligned} \tag{11.4.20}$$

where $v_{it} = \tilde{u}_{it} - (1+\delta)\tilde{u}_{i,t-1} + \delta\tilde{u}_{i,t-2}$ and $E^*[\boldsymbol{\nu}_i \mid \mathbf{x}_i] = \mathbf{0}$. Provided $T \geq 3$, $x_{i3}, \ldots, x_{iT}$ can serve as instruments for cross-sectional regression of the equation determining $y_{i2} - y_{i1}$.

In the more general case, with $J > 1$, $\boldsymbol{\delta}^* \mathbf{w}^{*\prime}$ in (11.4.19) will be replaced by $\sum_{j=1}^{J} \boldsymbol{\lambda}_j^* \mathbf{w}_j^{*\prime}$, where $\boldsymbol{\lambda}_j^{*\prime} = [1, \lambda_j, \ldots, \lambda_j^{T-1}]$, and $\mathbf{w}_j^*$ is the vector of coefficients from the projection of $b_{ij}$ on $\mathbf{x}_i$. Using a similar procedure, we can show that the $\Pi$ matrix will identify the lag coefficients if $T \geq J + 2$.

Of course, if in addition to Assumption 11.4.5 we also have information on the structure of $x$ process, there will be more restrictions on the $\Pi$ matrices than in the models in this subsection. Identification conditions can consequently be relaxed.

### 11.4.5 Estimation and Testing

We can estimate the unknown parameters of a distributed-lag model using short panels by first stacking all $T$ period equations as a system of reduced-form equations:

$$\underset{T\times 1}{\mathbf{y}_i} = \boldsymbol{\mu}^* + [I_T \otimes \mathbf{x}_i']\boldsymbol{\pi} + \boldsymbol{\nu}_i, \quad i = 1, \ldots, N, \tag{11.4.21}$$

where $\boldsymbol{\nu}_i = \mathbf{y}_i - E^*[\mathbf{y}_i \mid \mathbf{x}_i]$, and $\boldsymbol{\pi}' = [\boldsymbol{\pi}_1', \ldots, \boldsymbol{\pi}_T']$, where $\boldsymbol{\pi}_j'$ is the $j$th row of the matrix $\Pi$. By construction, $E(\boldsymbol{\nu}_i \otimes \mathbf{x}_i) = \mathbf{0}$. Under the assumption that the $N$ vectors $(\mathbf{y}_i', \mathbf{x}_i')$ are independent draws from a common distribution, with finite fourth-order moments and with $E\mathbf{x}_i\mathbf{x}_i' = \Sigma_{xx}$ positive definite, the least-squares estimator of $\boldsymbol{\pi}$, $\hat{\boldsymbol{\pi}}$, is consistent, and $\sqrt{N}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi})$ is asymptotically normally distributed, with mean 0 and variance–covariance matrix $\Omega$, which is given by (3.8.11).

The models of Sections 11.4.3 and 11.4.4 imply that $\boldsymbol{\pi} = \mathbf{f}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of the model's parameters of dimensions $m \leq (T + \ell + 1)$. We can impose these restrictions by a minimum-distance estimator that chooses $\hat{\boldsymbol{\theta}}$ to minimize

$$[\hat{\boldsymbol{\pi}} - \mathbf{f}(\boldsymbol{\theta})]'\hat{\Omega}^{-1}[\hat{\boldsymbol{\pi}} - \mathbf{f}(\boldsymbol{\theta})], \tag{11.4.22}$$

where $\hat{\Omega}$ is a consistent estimator of (3.8.11). Under fairly general conditions, the estimator $\hat{\boldsymbol{\theta}}$ is consistent, and $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normally distributed, with asymptotic variance–covariance matrix

$$(F'\Omega^{-1}F)^{-1}, \tag{11.4.23}$$

where $F = \partial\,\mathbf{f}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$. The identification condition ensures that $F$ has rank $m$. The quadratic form

$$N[\hat{\boldsymbol{\pi}} - \mathbf{f}(\hat{\boldsymbol{\theta}})]'\Omega^{-1}[\hat{\boldsymbol{\pi}} - \mathbf{f}(\hat{\boldsymbol{\theta}})] \tag{11.4.24}$$

is asymptotically $\chi^2$ distributed with $T(T + \ell + 1) - m$ degrees of freedom.

Equation (11.4.24) provides us with a test of the $T(T + \ell + 1) - m$ constraints $\mathbf{f}(\boldsymbol{\theta})$ placed on $\boldsymbol{\pi}$. To test nested restrictions, consider the null hypothesis $\boldsymbol{\theta} = \mathbf{g}(\boldsymbol{\omega})$, where $\boldsymbol{\omega}$ is a $k$-dimensional vector ($k \leq m$) of the parameters of the restricted model. Let $\mathbf{h}(\boldsymbol{\omega}) = \mathbf{f}[\mathbf{g}(\boldsymbol{\omega})]$; that is, $\mathbf{h}$ embodies the restrictions of the constrained model. Then, under the null hypothesis,

$$N[\hat{\boldsymbol{\pi}} - \mathbf{h}(\hat{\boldsymbol{\omega}})]'\Omega^{-1}[\hat{\boldsymbol{\pi}} - \mathbf{h}(\hat{\boldsymbol{\omega}})] \tag{11.4.25}$$

is asymptotically $\chi^2$ distributed with $T(T + \ell + 1) - k$ degrees of freedom, where $\hat{\boldsymbol{\omega}}$ minimizes (11.4.25). Hence, to test the null hypothesis, we can use

the statistic[15]

$$N[\hat{\pi} - \mathbf{h}(\hat{\boldsymbol{\omega}})]'\hat{\Omega}^{-1}[\hat{\boldsymbol{\pi}} - \mathbf{h}(\hat{\boldsymbol{\omega}})] - N[\hat{\boldsymbol{\pi}} - \mathbf{f}(\hat{\boldsymbol{\theta}})]'\hat{\Omega}^{-1}[\hat{\boldsymbol{\pi}} - \mathbf{f}(\hat{\boldsymbol{\theta}})], \tag{11.4.26}$$

which is asymptotically $\chi^2$ distributed, with $m - k$ degrees of freedom.

To illustrate the method of estimating unconstrained distributed-lag models using panel data, Pakes and Griliches (1984) investigated empirically the issues of how to construct the "stock of capital ($G$)" for analysis of rates of return. The basic assumption of their model is that there exists a stable relationship between earnings (gross or net profits) ($y$) and past investments ($x$), and firms or industries differ only in terms of the level of the yield on their past investments, with the time shapes of these yields being the same across firms and implicit in the assumed depreciation formula. Namely,

$$E^*[y_{it} \mid G_{it}, \alpha_i^*] = \alpha_i^* + \gamma G_{it}, \tag{11.4.27}$$

and

$$G_{it} = \sum_{\tau=1}^{\infty} \beta_{i\tau}\, x_{it-\tau}. \tag{11.4.28}$$

Substituting (11.4.28) into (11.4.27), we have a model that consists in regressing the operating profits of firms on a distributed lag of their past investment expenditures.

Using a sample of 258 manufacturing firms' annual profit data for the years 1964–72 and investment data for the years 1961–71, and assuming that $p$ in Assumption 11.4.4 equals three,[16] they found that the estimated lag coefficients rose over the first three periods and remained fairly constant over the next four or five. This pattern implies that the contribution of past investment to the capital stock first "appreciates" in the early years as investments are completed, shaken down, or adjusted to. This is distinctly different from the pattern implied by the commonly used straight-line or declining-balance depreciation formula to construct the "stock of capital." Both formulas imply that the lag coefficients decline monotonically in $\tau$, with the decline being the greatest in earlier periods for the second case.

[15] See Neyman (1949) or Hsiao (1984).

[16] Thus, they assume that this year's investment does not affect this year's profits and that there are two presample observations ($\ell = 1$) on investment.

CHAPTER 12

# Miscellaneous Topics

In this chapter we give brief introduction to some miscellaneous topics. We shall first consider panel duration and count data models in Section 12.1 and 12.2, respectively. Section 12.3 introduces the quantile regression model. Section 12.4 considers statistical inference using simulation methods. Section 12.5 discusses the conventional error components formulation for panels with more than two dimensions. Section 12.6 considers issues of measurement errors and indicates how one can take advantage of the panel structure to identify and estimate an otherwise unidentified model. Section 12.7 discusses nonparametric approaches for panel data modeling.

## 12.1 DURATION MODEL

Duration models study the length of time spent in a given state before transition to another state, such as the length of time unemployed. The length of the interval between two successive events is called a *duration*. A *duration* is a nonnegative random variable, denoted by $D$, representing the length of a time period spent by an individual or a firm in a given state. The cumulative distribution function of $D$ is defined as

$$\begin{aligned} F(t) &= \text{Prob}(D < t) \\ &= \textstyle\int_0^t f(s)\,ds, \end{aligned} \tag{12.1.1}$$

where

$$f(t) = \frac{dF(t)}{dt}. \tag{12.1.2}$$

Let $A_{its}$ denote the event that nothing happens between time period $t$ and $t+s$ for individual $i$; then an event of interest (say doctor visit) occurs at $t+s$. Suppose the probability of event $A_{it,t+\Delta}$ occurs where the time distance between two adjacent time periods approaches 0 is given by

$$P(A_{it,t+\Delta}) = \mu_{it}\Delta t. \tag{12.1.3}$$

Under the assumption that $\mu_{it}$ stays constant between period $t$ and $t+s$,

$$P(A_{its}) = (1-\mu_{it}\Delta t)^{\frac{s}{\Delta t}}\mu_{it}\Delta t. \tag{12.1.4}$$

Let $n = \frac{s}{\Delta t}$, then $\Delta t \longrightarrow 0, n \longrightarrow \infty$. Using the identity $\lim_{n\longrightarrow\infty}(1-n^{-1})^n = e^{-1}$, we obtain that for small $\Delta t$,

$$P(A_{its}) = \exp(-\mu_{it}s)\mu_{it}\Delta t. \tag{12.1.5}$$

Suppose $\mu_{it}$ stays constant between time period 0 and $t+\Delta t$, the probability that an individual stayed in a state (say unemployment) from 0 to $t$ and moved out at $t+\Delta t$, $f_i(t)\Delta t$, is given by (12.1.5). Then

$$\text{Prob}(D_i \geq t) = \exp(-\mu_{it}t). \tag{12.1.6}$$

The cumulative distribution function of $D$,

$$\begin{aligned} F_i(t) &= \text{Prob}(D_i < t) \\ &= 1 - \text{Prob}(D_i \geq t) \\ &= 1 - \exp(-\mu_{it}t). \end{aligned} \tag{12.1.7}$$

Hence

$$\begin{aligned} \mu_{it} &= \lim_{\Delta t\to 0}\frac{\text{Prob}[t \leq D_i < t+\Delta t \mid D_i \geq t]}{\Delta t} \\ &= \frac{f_i(t)}{1-F_i(t)}, \end{aligned} \tag{12.1.8}$$

where $\mu_{it}$ is called the *hazard function* of the duration variable $D_i$ and $f_i(t) = \exp(-\mu_{it}t)\mu_{it}$. The hazard function, $\mu_{it}$, gives the instantaneous conditional probability of leaving the current state (*the death of a process*) and

$$S_i(t) = 1 - F_i(t) = \text{Prob}[D_i \geq t], \tag{12.1.9}$$

is called the *survival function*. Let $\mu_{is}$ be the instant hazard rate at time $s$, the probability that the $i$th individual survives from time $s$ to $s+\Delta s$ is equal to $e^{-\mu_{is}\Delta s}$ ((12.1.6)). Breaking up the interval $(0,t)$ to $n$ subintervals, $n = \frac{t}{\Delta s}$, and using $S(0) = 1$ yields the probability that the $i$th individual survives at time $t$ equals to

$$\prod_{s=1}^{n}\exp(-\mu_{is}\Delta s) = \exp\left\{-\sum_{s=1}^{n}\mu_{is}\Delta s\right\}. \tag{12.1.10}$$

As $\Delta s \to 0$, it yields the survival function as

$$S_i(t) = \exp\left(-\int_0^t \mu_{is}ds\right). \tag{12.1.11}$$

It follows that

$$\mu_{it} = \text{Prob}(D_i = t \mid D_{it} \geq t) = -\frac{d\, ln S_i(t)}{dt}. \tag{12.1.12}$$

When $\mu_{it} = \mu_i$, the expected duration for the $i$th individual is

$$E(D_i) = \int_0^\infty t\mu_i \exp(-\mu_i t)dt = \frac{1}{\mu_i}. \tag{12.1.13}$$

Suppose the data on $N$ individuals take the form that each individual either experiences one complete spell at time $t_i$, that is, $D_i = t_i$, or right-censored at time $t_i^*$, that is, $D_i \geq t_i^*$. Suppose $\mu_{it} = \mu_i$ and $i = 1, \ldots, n$ complete their spells of duration $t_i$; then

$$f_i(t_i) = \mu_i \exp(-\mu_i t_i), \quad i = 1, \ldots, n, \tag{12.1.14}$$

Suppose $i = n + 1, \ldots, N$ are right-censored at $t_i^*$, then

$$S_i(t_i^*) = 1 - F_i(t_i^*) = \exp(-\mu_i t_i^*), \quad i = n + 1, \ldots, N. \tag{12.1.15}$$

Under the assumption that cross-sectional units are independently distributed, the likelihood function for the $N$ units takes the form,

$$L = \prod_{i=1}^{n} f_i(t_i) \cdot \prod_{i=n+1}^{N} [1 - F_i(t_i^*)]. \tag{12.1.16}$$

The hazard rate $\mu_{it}$ or $\mu_i$ is often assumed to be a function of socio-demographic variables, $\mathbf{x}_i$. Because duration is a nonnegative random variable, $\mu_i$ (or $\mu_{it}$) should clear to be nonnegative. A simple way to ensure this is to let

$$\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}). \tag{12.1.17}$$

Substituting (12.1.17) into (12.1.16), one can obtain the maximum-likelihood estimator of $\boldsymbol{\beta}$ by maximizing the logarithm of (12.1.16).

Alternatively, noting that

$$\begin{aligned} E \log t_i &= \mu_i \int_0^\infty (\log t) \exp(-\mu_i t)\, dt \\ &= -c - \log \mu_i, \end{aligned} \tag{12.1.18}$$

where $c \simeq 0.577$ is Euler's constant, and

$$\begin{aligned} \text{Var}(\log t_i) &= E(\log t_i)^2 - [E \log t_i]^2 \\ &= \frac{\pi^2}{6}, \end{aligned} \tag{12.1.19}$$

one can put the duration model in a regression framework,

$$\begin{aligned} \log t_i + 0.557 &= -\mathbf{x}_i'\boldsymbol{\beta} + u_i, \\ i &= 1, \ldots, n, \end{aligned} \tag{12.1.20}$$

where $E(u_i) = 0$ and $\text{var}(u_i) = \frac{\pi^2}{6}$. Consistent estimate of $\boldsymbol{\beta}$ can be obtained by the least-squares method using the $n$ subsample of individuals who experience one complete spell. However, the least-squares estimator has covariance matrix

$\frac{\pi^2}{6}\left(\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i'\right)^{-1}$, which is greater than the covariance matrix of the MLE of (12.1.16).

Cox (1972) proposes a proportional hazard model to take account the heterogeneity across individuals (and over time) by letting

$$\mu_{it} = \mu(t)g(\mathbf{x}_i), \tag{12.1.21}$$

where $\mu(t)$ is the so-called *baseline hazard* function and $g(\cdot)$ is a known function of observable exogenous variables $\mathbf{x}_i$. To ensure nonnegativity of $\mu_{it}$, a common formulation for $g(\mathbf{x}_i)$ is to let

$$g(\mathbf{x}_i) = \exp(\mathbf{x}_i'\boldsymbol{\beta}). \tag{12.1.22}$$

Then

$$\frac{\partial \mu_{it}}{\partial x_{ki}} = \beta_k \cdot \exp(\mathbf{x}_i'\boldsymbol{\beta})\mu(t) = \beta_k \cdot \mu_{it} \tag{12.1.23}$$

has a constant proportional effect on the instantaneous conditional probability of leaving the state. However, $\mu_{it} = \mu(t)cc^{-1}g(\mathbf{x}_i)$ for any $c > 0$. We need to define a reference individual. A common approach is to choose a particular value of $x = x^*$ such that $g(\mathbf{x}^*) = 1$.

One can simultaneously estimate the *baseline hazard* function $\mu(t)$ and $\boldsymbol{\beta}$ by maximizing the logarithm of the likelihood function (12.1.16). However, Cox's proportional hazard model allows the separation of the estimation of $g(\mathbf{x}_i)$ from the estimation of the baseline hazard $\mu(t)$. Let $t_1 < t_2 < \cdots < t_j < \cdots < t_n$ denote the observed ordered discrete exit times of the spell (it is referred as *failure* time when the date of change is interpreted as a breakdown or a failure) for $i = 1, \ldots, n$, in a sample consisting of $N$ individuals, $N \geq n$, and let $t_i^*, i = n+1, \ldots, N$ be the censored time for those with censored durations. Substituting (12.1.21) and (12.1.22) into the likelihood function (12.1.16) yields Cox's proportional hazard model likelihood function,

$$\begin{aligned} L &= \prod_{i=1}^{n} \exp(\mathbf{x}_i'\boldsymbol{\beta})\mu(t_i) \cdot \exp[-\exp(\mathbf{x}_i'\boldsymbol{\beta})\int_0^{t_i}\mu(s)ds] \\ &\quad \cdot \prod_{i=n+1}^{N} \exp[-\exp(\mathbf{x}_i'\boldsymbol{\beta})\int_0^{t_i^*}\mu(s)ds] \\ &= \prod_{i=1}^{n} \exp(\mathbf{x}_i'\boldsymbol{\beta})\mu(t_i) \\ &\quad \cdot \exp\left\{-\int_0^{\infty}\left[\sum_{h\in R(t_\ell)} \exp(\mathbf{x}_h'\boldsymbol{\beta})\right]\mu(t)dt\right\} \\ &= L_1 \cdot L_2, \end{aligned} \tag{12.1.24}$$

where

$$R(t_\ell) = \{i \mid S_i(t) \geq t_\ell\}$$

denotes the set of individuals who are at risk of exiting just before the $\ell$th ordered exiting,

$$L_1 = \prod_{i=1}^{n} \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{\sum_{h \in R(t_i)} \exp(\mathbf{x}_h'\boldsymbol{\beta})}, \tag{12.1.25}$$

$$\begin{aligned} L_2 = \sum_{i=1}^{n} \left[ \sum_{h \in R(t_i)} \exp(\mathbf{x}_h'\boldsymbol{\beta}) \mu(t_i) \right] \\ \cdot \exp\left\{ -\int_0^{\infty} \left[ \sum_{h \in R(t_j)} \exp(\mathbf{x}_h'\boldsymbol{\beta}) \right] \mu(s) ds \right\}. \end{aligned} \tag{12.1.26}$$

Because $L_1$ does not depend on $\mu(t)$, Cox (1975) suggests maximizing the partial likelihood function $L_1$ to obtain the PMLE estimator $\hat{\boldsymbol{\beta}}_p$. It was shown by Tsiatis (1981) that the partial MLE of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_p$ is consistent and asymptotically normally distributed with the asymptotic covariance

$$\text{Cov}(\hat{\boldsymbol{\beta}}_p) = -\left[ E \frac{\partial^2 \log L_1}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \right]^{-1}. \tag{12.1.27}$$

Once $\hat{\boldsymbol{\beta}}_p$ is obtained, one can estimate $\mu(t)$ parametrically by substituting $\hat{\boldsymbol{\beta}}_p$ for $\boldsymbol{\beta}$ in the likelihood function (12.1.16) or semiparametrically through the relation

$$-\log S_i(t_i) = \mathbf{x}_i'\boldsymbol{\beta} + \int_0^{t_i} \mu(s) ds + \epsilon_i \tag{12.1.28}$$

(For details, see Han and Hausman 1990 or Florens, Fougére, and Mouchart 2008).

To allow for the presence of unobserved heterogeneity, mixture models have been proposed for the hazard rate,

$$\mu_{it} = \mu(t) \exp(\mathbf{x}_i'\boldsymbol{\beta}) \alpha_i, \alpha_i > 0 \tag{12.1.29}$$

where $\alpha_i$ denotes an unobserved heterogeneity term for the $i$th individual, independent of $\mathbf{x}_i$ and normalized with $E(\alpha_i) = 1$. Common assumptions for the heterogeneity are gamma, inverse gamma, or log-normal. Once the heterogeneity distribution is specified, one can integrate out $\alpha_i$ to derive the marginal survivor function or expected duration conditional on $\mathbf{x}_i$.

The idea of studying the duration of an event based on the hazard rate (instant rate of exit) has wide applications in economics and finance, for example, duration of a strike, unemployment, or a mortgage. It can also be applied to predict exit of an event in the future based on current state variables. For

instance, Duan et al. (2012) define the average exit intensity for the period $[t, t+\tau]$ as:

$$\mu_{it}(\tau) = -\frac{ln[1 - F_{it}(\tau)]}{\tau}, \tag{12.1.30}$$

where $F_{it}(\tau)$ is the conditional distribution function of the exit at time $t+\tau$ evaluated at time $t$ for the $i$th individual. (When $\tau = 0$, $\mu_{it}(0)$ is the hazard function defined at (12.1.8) or (12.1.12), which they call the forward intensity rate.) Then as in (12.1.9) and (12.1.12), the average exit intensity for the period $[t, t+\tau]$ becomes

$$\begin{aligned} \mu_{it}(\tau) &= -\frac{ln[1 - F_{it}(\tau)]}{\tau} \\ &= -\frac{ln[\exp\,(-\int_t^{t+\tau} \mu_{is} ds)]}{\tau}. \end{aligned} \tag{12.1.31}$$

Hence the survival probability over $[t, t+\tau]$ becomes $\exp\,[-\mu_{it}(\tau)\tau]$.

Assume $\mu_{it}(\tau)$ is differentiable with $\tau$, it follows from (12.1.8) that the instantaneous forward exit intensity at time $t+\tau$ is:

$$\psi_{it}(\tau) = \frac{F'_{it}(\tau)}{1 - F_{it}(\tau)} = \mu_{it}(\tau) + \mu'_{it}(\tau)\tau. \tag{12.1.32}$$

Then $\mu_{it}(\tau) \cdot \tau = \int_0^\tau \psi_{it}(s) ds$. The forward exit probability at time $t$ for the period $[t+\tau, t+\tau+1]$ is then equal to

$$\int_0^1 e^{-\mu_{it}(\tau+s)s} \psi_{it}(\tau+s)\, ds. \tag{12.1.33}$$

A firm can exit either due to bankruptcy or other reasons such as mergers or acquisitions. In other words, $\psi_{it}(s)$ is a combined exit intensity of default and other exit. Let $\phi_{it}(\tau)$ denote the forward default intensity. Then the default probability over $[t+\tau, t+\tau+1]$ at time $t$ is

$$\int_0^1 e^{-\mu_{it}(\tau+s)s} \phi_{it}(\tau+s)\, ds. \tag{12.1.34}$$

The actual exit is recorded at discrete time, say in a month or year. Discretizing the continuous version by $\Delta t$ for empirical implementation yields the forward (combined) exit probability and forward default probability at time $t$ for the period $[t+\tau, t+\tau+1]$ as

$$e^{-\mu_{it}(\tau)\tau\Delta t}\left[1 - e^{-\psi_{it}(\tau)\Delta t}\right], \tag{12.1.35}$$

and

$$e^{-\mu_{it}(\tau)\tau\Delta t}\left[1 - e^{-\phi_{it}(\tau)\Delta t}\right], \tag{12.1.36}$$

Table 12.1. *Total number of active firms, defaults/bankruptcies, and other exits for each year over the sample period 1991–2011*

| Year | Active firms | Defaults/bankruptcies | (%) | Other exit | (%) |
|---|---|---|---|---|---|
| 1991 | 4012 | 32 | 0.80 | 257 | 6.41 |
| 1992 | 4009 | 28 | 0.70 | 325 | 8.11 |
| 1993 | 4195 | 25 | 0.6 | 206 | 4.91 |
| 1994 | 4433 | 24 | 0.54 | 273 | 6.16 |
| 1995 | 5069 | 19 | 0.37 | 393 | 7.75 |
| 1996 | 5462 | 20 | 0.37 | 463 | 8.48 |
| 1997 | 5649 | 44 | 0.78 | 560 | 9.91 |
| 1998 | 5703 | 64 | 1.12 | 753 | 13.20 |
| 1999 | 5422 | 77 | 1.42 | 738 | 13.61 |
| 2000 | 5082 | 104 | 2.05 | 616 | 12.12 |
| 2001 | 4902 | 160 | 3.26 | 577 | 11.77 |
| 2002 | 4666 | 81 | 1.74 | 397 | 8.51 |
| 2003 | 4330 | 61 | 1.41 | 368 | 8.50 |
| 2004 | 4070 | 25 | 0.61 | 302 | 7.42 |
| 2005 | 3915 | 24 | 0.61 | 291 | 7.43 |
| 2006 | 3848 | 15 | 0.39 | 279 | 7.25 |
| 2007 | 3767 | 19 | 0.50 | 352 | 9.34 |
| 2008 | 3676 | 59 | 1.61 | 285 | 7.75 |
| 2009 | 3586 | 67 | 1.87 | 244 | 6.80 |
| 2010 | 3396 | 25 | 0.74 | 242 | 7.13 |
| 2011 | 3224 | 21 | 0.65 | 226 | 7.01 |

The number of active firms is computed by averaging over the number of active firms across all months of the year.
*Source:* Duan, Sun, and Wang (2012, Table 1).

respectively with spot (instantaneous) exit intensity at time $t$ for the period $[t, t+\tau]$ being

$$\mu_{it}(\tau) = \frac{1}{\tau}[\mu_{it}(\tau-1)(\tau-1) + \psi_{it}(\tau-1)]. \tag{12.1.37}$$

Default is only one of the possibilities for a firm to exit; $\phi_{it}(\tau)$ must be no greater than $\psi_{it}(\tau)$. Suppose $\psi_{it}(\tau)$ and $\phi_{it}(\tau)$ depend on a set of macroeconomic factors and firm-specific attribute, $\mathbf{x}_{it}$, a convenient specification to ensure $\phi_{it}(\tau) \leq \psi_{it}(\tau)$ is to let

$$\phi_{it}(\tau) = \exp(\mathbf{x}_{it}'\boldsymbol{\gamma}(\tau)), \tag{12.1.38}$$

and

$$\psi_{it}(\tau) = \phi_{it}(\tau) + \exp(\mathbf{x}_{it}'\boldsymbol{\beta}(\tau)). \tag{12.1.39}$$

Duan, Sun, and Wang (2012) use the monthly data ($\Delta t = \frac{1}{12}$) of 12,268 publicly traded firms for the period 1991 to 2011 to predict the multiperiod ahead default probabilities for the horizon $\tau$ from 0 to 35 months. Table 12.1 provides

the summaries of the number of active companies, defaults/bankruptcies, and other exits each year. The overall default rate ranges between 0.37 percent and 3.26 percent of the firms in each sample year. Other forms of exit are significantly higher, ranging from 4.91 percent to 13.61 percent. The macro and firm-specific attributes for $\phi_{it}(\tau)$ and $\psi_{it}(\tau)$ include (trailing) one-year S&P 500 return (SP500); three-month US Treasury bill rate; the firm's distance to default (DTD), which is a volatility-adjusted leverage measure based on Merton (1974) (for details, see Duan and Wang 2012); ratio of cash and short-term investments to the total assets (CASH/TA); ratio of net income to the total assets (NI/TA); logarithm of the ratio of a firm's market equity value to the average market equity value of the S&P 500 firms (SIZE); market-to-book asset ratio (M/B); one-year idiosyncratic volatility, calculated by regressing individual monthly stock return on the value-weighted the Center for Research in Security Prices (CRSP) monthly return over the preceding 12 months (Sigma). Both level and trend measures for DTD, CASH/TA, NI/TA and SIZE are employed in the empirical analysis. To take account the impact of the massive US governmental interventions during the 2008–09 financial crisis, Duan et al. (2012) also include a common bail out term, $\lambda(\tau)\exp\{-\delta(\tau)(t-t_B)\}\cdot 1[(t-t_B)>0]$ for $\tau = 0, 1, \ldots, 11$, to the forward default intensity function where $t_B$ denotes the end of August 2008 and 1(A) is the indicator function that equals 1 if event A occurs and 0 otherwise. Specifically

$$\begin{aligned}\phi_{it}(\tau) = \exp\{&\lambda(\tau)\exp\left[-\delta(\tau)(t-t_B)\right]1((t-t_B)>0)\\ &+\mathbf{x}_{it}'\boldsymbol{\gamma}(\tau)\},\end{aligned} \tag{12.1.40}$$

for $\tau = 0, 1, \ldots, 11$.

Assuming the firms are cross-sectional independent conditional on $\mathbf{x}_{it}$, and ignoring the time dependence, Duan et al. (2012) obtain the estimated $\phi_{it}(\tau)$ and $\psi_{it}(\tau)$ by maximizing the pseudo-likelihood function,

$$\prod_{i=1}^{N}\prod_{t=0}^{T-1}\mathcal{L}_{it}(\tau). \tag{12.1.41}$$

Let $t_{oi}$, $\tau_{oi}$, and $\tau_{ci}$ denote the first month that appeared in the sample, default time, and combined exit time for firm $i$, respectively, $\mathcal{L}_{it}(\tau)$ is defined as,

$$\begin{aligned}\mathcal{L}_{it}(\tau) = &1\{t_{oi}\le t, \tau_{ci} > t+\tau\}P_t(\tau_{ci}>t+\tau)\\ &+1\{t_{oi}<t, \tau_{ci}=\tau_{oi}\le t+\tau\}P_t(\tau_{ci};\tau_{ci}=\tau_{oi}\le t+\tau)\\ &+1\{t_{oi}<t, \tau_{ci}\ne\tau_{oi}, \tau_{ci}\le t+\tau\}P_t(\tau_{ci};\tau_{ci}\ne\tau_{oi}\\ &\text{and } \tau_{ci}\le t+\tau)+1\{t_{oi}>t\}+1\{\tau_{ci}\le t\},\end{aligned} \tag{12.1.42}$$

$$P_t(\tau_{ci}>t+\tau) = \exp\left[-\sum_{s=0}^{\tau-1}\psi_{it}(s)\Delta t\right], \tag{12.1.43}$$

$$
\begin{aligned}
&P_t(\tau_{ci} \mid \tau_{ci} = \tau_{oi} \le t + \tau) \\
&= \begin{cases} 1 - \exp[-\phi_{it}(0)\Delta t], \text{ if } \tau_{oi} = t + 1, \\ \exp\left[-\sum_{s=0}^{\tau_{ci}-t-2} \psi_{it}(s)\Delta t\right] \end{cases} \\
&\quad \cdot [1 - \exp[-\phi_{it}(\tau_{c_i} - t - 1)\Delta t]\}, \text{ if } t + 1 < \tau_{ci} \le t + \tau,
\end{aligned} \tag{12.1.44}
$$

$$
\begin{aligned}
&P_t(\tau_{ci} \mid \tau_{ci} \ne \tau_{oi} \le t + \tau) \\
&= \begin{cases} \exp[-\phi_{it}(0)\Delta t] - \exp[-\psi_{it}(0)\Delta t], \text{ if } \tau_{oi} = t + 1, \\ \exp[-\sum_{s=0}^{\tau_{ci}-t-2} \psi_{it}(s)\Delta t] \end{cases} \\
&\quad \cdot \{\exp[-\phi_{it}(\tau_{ci} - t - 1)\Delta t] - \exp[-\psi_{it}(\tau_{ci} - t - 1)]\}, \\
&\quad \text{if } t + 1 < \tau_{ci} \le t + \tau,
\end{aligned} \tag{12.1.45}
$$

with $\Delta t = \frac{1}{12}$, and $\psi_{it}(s)$ and $\phi_{it}(s)$ take the form of (12.1.38) and (12.1.39), respectively. The first term on the right-hand side of $\mathcal{L}_{it}(\tau)$ is the probability of surviving both forms of exit. The second term is the probability that the firm defaults at a particular time point. The third term is the probability that the firm exits due to other reasons at a particular time point. If the firm does not appear in the sample in month $t$, it is set equal to 1, which is transformed to 0 in the log-pseudo-likelihood function. The forward intensity approach allows an investigator to predict the forward exiting time of interest $\tau$, $\phi_{it}(\tau)$, and $\psi_{it}(\tau)$ as functions of conditional variables available at time $t$ without the need to predict future conditional variables.

Figure 12.1 plots each of the estimated $\boldsymbol{\gamma}(\tau)$ and $\boldsymbol{\beta}(\tau)$ and its 90% confidence interval with $\tau$ ranging from 0 month to 35 months. They show that some firm-specific attributes influence the forward intensity both in terms of level and trend. Figure 12.2 plot the estimated term structure of predicted default probabilities of Lehman Brothers, Merrill Lynch, Bank of America, and the averages of the US financial sector at several time points prior to Lehman Brothers bankruptcy filing on September 15, 2008. The term structures for Lehman Brothers in June 2008, three months before its bankruptcy filing, show that the company's short-term credit risk reached its historical high. The peak of the forward default probability is one month. The one-year cumulative default probability increased sharply to 8.5 percent, which is about 35 times of the value three years earlier. This case study appears to suggest that the forward intensity model is quite informative for short prediction horizons.

## 12.2 COUNT DATA MODEL

The count data model is the dual of the duration model. The duration model considers the probability that an event stays for certain time period before another event occurs. The count data models consider the probability that a

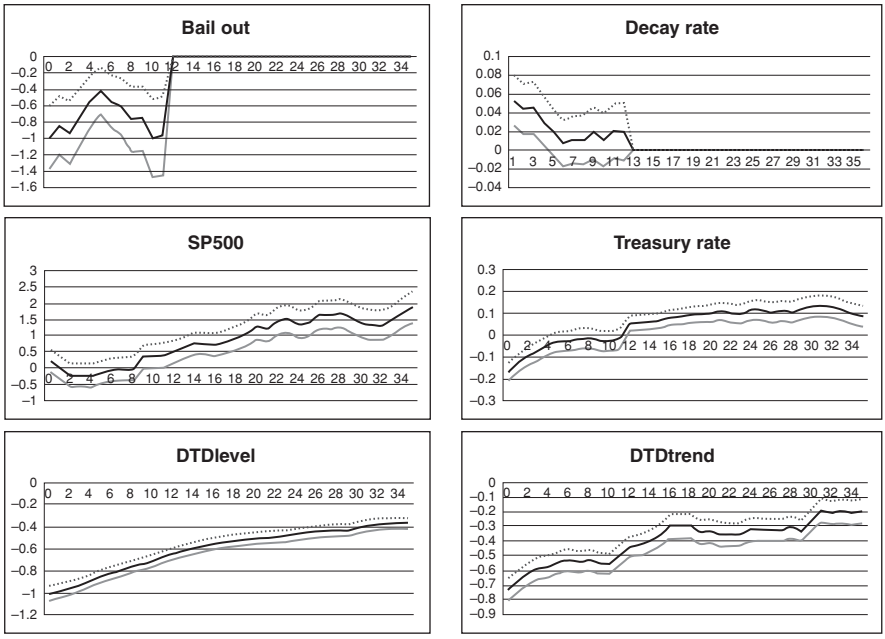


Figure 12.1. Parameter estimates for the forward default intensity function. The solid line is for the parameter estimates and the dotted lines depict the 90% confidence interval. *Source:* Duan, Sun, and Wang (2012, Fig. 1).

certain number of an event would occur during a fixed period of time. Under the assumption that the instant arrival rate is $\mu_{it}$. The probability of the nonnegative integer count number $y_{it}$ in a unit interval is given by a Poisson process.

$$P(y_{it}) = \frac{e^{-\mu_{it}}(\mu_{it})^{y_{it}}}{y_{it}!}, \quad y_{it} = 0, 1, 2, \ldots \tag{12.2.1}$$

To see this, suppose $y_{it} = 2$. Let $t + s_1$ and $t + s_1 + s_2$ be the time that the first and the second occurrence of the event of interest. Then $0 \leq s_1 < 1, 0 < s_2 < 1 - s_1$, and the probability that $y_{it} = 2$ is equal to the probability that one event occurs at $t + s_1$, another at $t + s_1 + s_2$ (or $s_2$ between 0 and $1 - s_1$), and no event occurs between $t + s_1 + s_2$ and $t + 1$),

$$\begin{aligned} P(y_{it} = 2) &= \int_0^1 \mu_{it} \exp(-\mu_{it} s_1) \left\{ \int_0^{1-s_1} \mu_{it} \exp[-\mu_{it} s_2] \right. \\ &\quad \left. \cdot \exp[-\mu_{it}(1 - s_1 - s_2)] ds_2 \right\} ds_1 \\ &= \frac{(\mu_{it})^2 e^{-\mu_{it}}}{2}. \end{aligned} \tag{12.2.2}$$

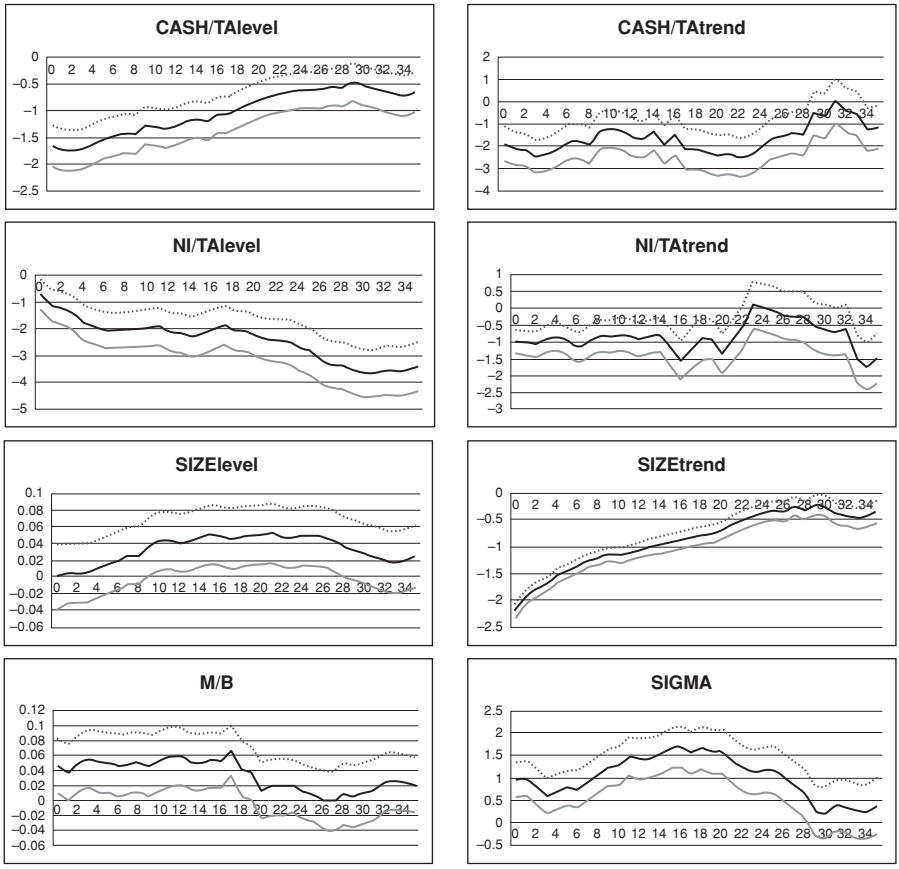


Figure 12.1 (*continued*).

Similarly, one can show that

$$P(y_{it} = r) = \frac{(\mu_{it})^r \exp(-\mu_{it})}{r!}. \tag{12.2.3}$$

The Poisson model implies $y_{it}$ is independent over time,

$$\text{Prob}(y_{it} = r \mid y_{i,t-s} = \ell) = P(y_{it} = r), \tag{12.2.4}$$

$$E(y_{it}) = \mu_{it}, \tag{12.2.5}$$

and

$$\text{Var}(y_{it}) = \mu_{it}. \tag{12.2.6}$$

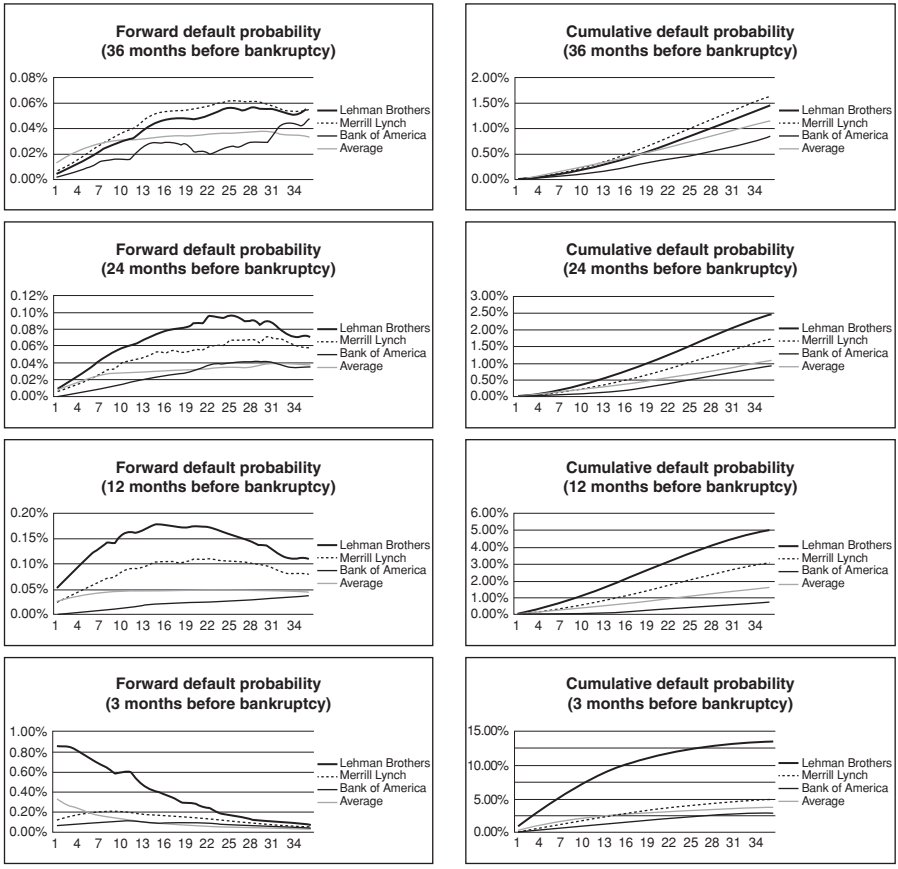


Figure 12.2. Lehman Brothers' term structure of forward and cumulative default probabilities. This figure shows the estimated term structure of forward default probabilities and that of cumulative default probabilities for Lehman Brothers, Merrill Lynch, Bank of America as well as the average values of the financial sector at 36 months, 24 months, 12 months, and 3 months before Lehman Brothers' bankruptcy filing date (September 15, 2008). *Source:* Duan, Sun, and Wang (2012, Fig. 4).

Therefore, under the assumption that $y_{it}$ is independently distributed across $i$, the log-likelihood function is given by

$$\log L = \sum_{i=1}^{N} \sum_{t=1}^{T} [y_{it} \log(\mu_{it}) - \mu_{it} - \log(y_{it}!)] \tag{12.2.7}$$

The intensity $\mu_{it}$ is often assumed a function of $K$ strictly exogenous variables, $\mathbf{x}_{it}$, and individual-specific effects, $\alpha_i$. Because $\mu_{it}$ has to be nonnegative, two popular specifications to ensure nonnegative $\mu_{it}$ without the need to impose

restrictions on the parameters are to let

$$\mu_{it} = \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i), \tag{12.2.8}$$

or to let

$$\mu_{it} = \alpha_i \exp(\mathbf{x}_{it}'\boldsymbol{\beta}), \quad \alpha_i > 0 \tag{12.2.9}$$

and $E(\alpha_i) = 1$.

When $\alpha_i$ is treated random and independent of $\mathbf{x}_{it}$ with known density function $g(\alpha)$, the marginal distribution of $(y_{i1}, \ldots, y_{iT})$ takes the form

$$f(y_{i1}, \ldots, y_{iT}) = \int \prod_{t=1}^{T} \left[ \frac{(\mu_{it})^{y_{it}} \exp(-\mu_{it})}{y_{it}!} \right] g(\alpha) d\alpha. \tag{12.2.10}$$

The MLE of $\boldsymbol{\beta}$ is consistent and asymptotically normally distributed either $N$ or $T$ or both tend to infinity. However, the computation can be tedious because the need to take multiple integration. For instance, suppose $g(\alpha)$ has gamma density $g(\alpha) = \alpha^{\nu-1} \exp(-\alpha)/\Gamma(\nu)$ with $E(\alpha) = 1$ and variance $\nu(\nu > 0)$. When $T = 1$ (cross-sectional data),

$$f(y_i) = \frac{[\exp(\mathbf{x}_i'\boldsymbol{\beta})]^{y_i} \Gamma(y_i + \nu)}{y_i! \Gamma(\nu)} \left( \frac{1}{\exp(\mathbf{x}_i'\boldsymbol{\beta}) + \nu} \right)^{y_i + \nu}, \quad i = 1, \ldots, N, \tag{12.2.11}$$

has a *negative binomial* distribution. But if $T > 1$, (12.2.10) no longer has the closed form. One computationally simpler method to obtain consistent estimator of $\boldsymbol{\beta}$ is to ignore the serial dependence of $y_{it}$ because of the presence of $\alpha_i$ by considering the marginal (or unconditional) distribution of $y_{it}$. For instance, if $\mu_{it}$ takes the form of (12.2.9) and $\alpha$ is gamma distributed, then the unconditional distribution of $y_{it}$ takes the form of (12.2.11). Maximizing the pseudo-joint likelihood function $\prod_{i=1}^{N} \prod_{t=1}^{T} f(y_{it})$ yields consistent estimator of $\boldsymbol{\beta}$ either $N$ or $T$ or both tend to infinity. The pseudo-MLE can also be used as initial values of the iterative schemes to obtain the MLE.

Conditional on $\alpha_i$, the log-likelihood function remains of the simple form (12.2.7). When $\alpha_i$ is treated as fixed and $\mu_{it}$ takes either the form (12.2.8) or (12.2.9), the maximum-likelihood estimator of $\boldsymbol{\beta}$ and $\eta_i$, where $\eta_i = \exp(\alpha_i)$ or $\eta_i = \alpha_i$ if $\mu_{it}$ takes the form (12.2.8) or (12.2.9), respectively, are obtained by simultaneously solving the first-order conditions

$$\frac{\partial \log L}{\partial \eta_i} = \sum_{t=1}^{T} \left[ \frac{y_{it}}{\eta_i} - \exp(\mathbf{x}_{it}'\boldsymbol{\beta}) \right] = 0, \quad i = 1, \ldots, N, \tag{12.2.12}$$

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \sum_{t=1}^{T} [y_{it} - \eta_i \exp(\mathbf{x}_{it}'\boldsymbol{\beta})] \mathbf{x}_{it} = \mathbf{0}. \tag{12.2.13}$$

Solving (12.2.12) yields the MLE of $\eta_i$ conditional on $\boldsymbol{\beta}$ as

$$\hat{\eta}_i = \frac{\bar{y}_i}{\bar{\mu}_i}, \quad i = 1, \ldots, N. \tag{12.2.14}$$

where $\bar{y}_i = T^{-1}\sum_{t=1}^{T} y_{it}$, $\bar{\mu}_i = T^{-1}\sum_{t=1}^{T} \exp(\mathbf{x}_{it}'\boldsymbol{\beta})$. Substituting $\hat{\eta}_i$ for $\eta_i$ in (12.2.13), the MLE of $\boldsymbol{\beta}$ is the solution to

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\left[y_{it} - \frac{\bar{y}_i}{\hat{\bar{\mu}}_i}\exp(\mathbf{x}_{it}'\hat{\boldsymbol{\beta}})\right]\mathbf{x}_{it} = \mathbf{0}, \tag{12.2.15}$$

where $\hat{\bar{\mu}}_i = \frac{1}{T}\sum_{t=1}^{T} \exp(\mathbf{x}_{it}'\hat{\boldsymbol{\beta}})$. When $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, (12.2.15) is equivalent to

$$\operatorname*{plim}_{N\longrightarrow\infty}\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\left[u_{it} - \frac{\bar{u}_i}{\bar{\mu}_i}\exp(\mathbf{x}_{it}'\boldsymbol{\beta})\right]\mathbf{x}_{it} = \mathbf{0}, \tag{12.2.16}$$

where

$$\begin{aligned} u_{it} &= y_{it} - E(y_{it} \mid \mathbf{x}_{it}, \alpha_i), \\ &= y_{it} - \eta_i \exp(\mathbf{x}_{it}'\boldsymbol{\beta}), \end{aligned} \tag{12.2.17}$$

and $\bar{u}_i = \frac{1}{T}\sum_{t=1}^{T} u_{it}$.

The strict exogeneity of $\mathbf{x}_{it}$ implies that

$$E(y_{it} \mid \mathbf{x}_{it}, \alpha_i) = E(y_{it} \mid \mathbf{x}_i, \alpha_i), \tag{12.2.18}$$

where $\mathbf{x}_i = (\mathbf{x}_{i1}', \ldots, \mathbf{x}_{iT}')'$. Therefore, $E(u_{it} \mid \mathbf{x}_i) = 0$, and hence (12.2.17) follows. However, the MLE of $\alpha_i$ (or $\eta_i$) is consistent only when $T \longrightarrow \infty$.

The sufficient statistic for $\eta_i$ is $\sum_{t=1}^{T} y_{it}$. Conditional on $\sum_{t=1}^{T} y_{it}$, the Poisson conditional log-likelihood function is given by (Hausman, Hall, and Griliches 1984)

$$\log L^* = \sum_{i=1}^{N}\sum_{t=1}^{T}\Gamma(y_{it}+1) - \sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}\log\left\{\sum_{s=1}^{T}\exp[-(\mathbf{x}_{it} - \mathbf{x}_{is})'\boldsymbol{\beta}]\right\}, \tag{12.2.19}$$

where $\Gamma(\cdot)$ is the gamma function. Equation (12.2.19) no longer involves the incidental parameters $\alpha_i$. Maximizing (12.2.19) yields consistent and asymptotic normally distributed estimator under the usual regularity conditions. As a matter of fact, $\frac{\partial \log L^*}{\partial \boldsymbol{\beta}}$ is identical to (12.2.16) (for details, see Windmeijer (2008)).

The limitations of Poisson models are the mean-variance equality restriction ((12.2.5) and (12.2.6)) and conditional on $\alpha_i$, $y_{it}$ independent of $y_{i,t-1}$. These features often contradict to the observed phenomena that the (conditional) variance usually exceeds the (conditional) mean and $y_{it}$ are not independent of $y_{i,t-1}$. The introduction of individual-specific effects, $\alpha_i$, partially get around the overdispersion problem. For instance, under the assumption that $\mu_{it}$ takes the form of (12.2.9) and $\alpha_i$ follows a gamma distribution, (12.2.11) leads to

$$E(y \mid \mathbf{x}) = \exp(\mathbf{x}'\boldsymbol{\beta}) \tag{12.2.20}$$

and

$$\operatorname{Var}(y \mid \mathbf{x}) = \exp(\mathbf{x}'\boldsymbol{\beta})[1 + \nu\exp(\mathbf{x}'\boldsymbol{\beta})] > E(y \mid \mathbf{x}). \tag{12.2.21}$$

One way to explicitly take account of serial dependence of $y_{it}$ is to include lagged $y_{i,t-1}$ into the specification of the mean arrival function $\mu_{it}$. However, inclusion of the lagged dependent variable in an exponential mean function may lead to rapidly exploding series. Crépon and Duguet (1997) suggest specifying

$$\mu_{it} = h(y_{i,t-1}) \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i). \tag{12.2.22}$$

Possible choice for $h(\cdot)$ could be

$$h(y_{i,t-1}) = \exp(\gamma(1 - d_{i,t-1})), \tag{12.2.23}$$

or

$$h(y_{i,t-1}) = \exp(\gamma_1 \ell n(y_{i,t-1} + cd_{i,t-1}) + \gamma_2 d_{i,t-1}), \tag{12.2.24}$$

where $c$ is a pre-specified constant, $d_{it} = 1$ if $y_{it} = 0$ and 0 otherwise. In this case, $\ell n y_{i,t-1}$ is included as a regressor for positive $y_{i,t-1}$, and 0 values of $y_{i,t-1}$ have a separate effect on current values of $y_{it}$. Alternatively, Blundell, Griffith, and Windmeijer (2002) propose a linear feedback model of the form

$$\mu_{it} = \gamma y_{i,t-1} + \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i). \tag{12.2.25}$$

Unfortunately, neither specification leads to easy to device MLE (because of the complications in formulating the distribution for the initial values) or moment conditions (because of the nonlinear nature of the moment functions).

Another often observed phenomena in data is that there is a much larger probability mass at the 0 (count) value than predicted by the Poisson model. One way to deal with this "excess zeros" in the data is to assume a two-part model or zero-inflated model in which the 0's and the positives come from two different data-generating process (e.g., Gurmu and Trivedi 1996; Harris and Zhao 2007), in which the probability of $y = 0$ or not is given by a binary process, say $F_1^*(0)$ and $(1 - F_1^*(0))$, and the probability $y$ takes the count values of 0,1,2,... from the count probability $F_2(y = r)$. Then the two part model assumes

$$\text{Prob}(y = 0) = F_1^*(0) \tag{12.2.26}$$

$$\text{Prob}(y = r) = [1 - F_1^*(0)]F_2(y = r), \quad r \geq 1. \tag{12.2.27}$$

The zero-inflated model assumes the zero-inflated model has probability

$$P(y = 0) = F_1^*(0) + [1 - F_1^*(0)]F_2(y = 0) \tag{12.2.28}$$

and

$$P(y = r) = [1 - F_1^*(0)]F_2^*(y = r) \text{ if } r \geq 1. \tag{12.2.29}$$

For additional discussions on modifying Poisson model to take account endogeneity, etc., see Trivedi and Munkin (2011) and Windmeijer (2008).

## 12.3 PANEL QUANTILE REGRESSION

The $\tau$th quantile of a random variable $y$, $y_\tau$, for $0 < \tau < 1$ is defined as

$$\text{Prob}(y \leq y_\tau) = \int_{-\infty}^{y_\tau} f(y)dy = F(y_\tau) = \tau, \tag{12.3.1}$$

where $f(y)$ denotes the probability density function of $y$. The sample location quantiles estimator for the $\tau$th sample quantile, $0 < \tau < 1$, for $N$ random sample $y_i$ is the solution to the minimization problem

$$\underset{c}{\text{Min}} \left\{ \sum_{i \in \psi_c} \tau \mid y_i - c \mid + \sum_{i \in \bar{\psi}_c} (1-\tau) \mid y_i - c \mid \right\} \tag{12.3.2}$$

where $\psi_c = \{i \mid y_i \geq c\}$ and $\bar{\psi}_c = \{i \mid y_i < c\}$.

As $N \to \infty$, eq. (12.3.2) divided by $N$ converges to

$$\begin{aligned} S(c) = (1-\tau) &\int_{-\infty}^{c} \mid y - c \mid f(y)\, dy \\ &+ (\tau) \int_{c}^{\infty} \mid y - c \mid f(y)\, dy. \end{aligned} \tag{12.3.3}$$

Suppose $0 < c < y_\tau$. For $y < c$, $\mid y - c \mid = \mid y - y_\tau \mid - \mid y_\tau - c \mid$. For $c < y < y_\tau$, $\mid y - c \mid = \mid y_\tau - c \mid - \mid y - y_\tau \mid$. For $y > y_\tau$, $\mid y - c \mid = \mid y - y_\tau \mid + \mid y_\tau - c \mid$. Equation (12.3.3) can be written as

$$\begin{aligned} S(c) &= (1-\tau) \int_{-\infty}^{c} \mid y - c \mid f(y)\, dy \\ &\quad + \tau \int_{c}^{y_\tau} \mid y - c \mid f(y)\, dy \\ &\quad + \tau \int_{y_\tau}^{\infty} \mid y - c \mid f(y)\, dy \\ &= S(y_\tau) + \mid y_\tau - c \mid (\tau - F(c)) - \int_{c}^{y_\tau} \mid y - y_\tau \mid f(y)dy \\ &\geq S(y_\tau), \end{aligned} \tag{12.3.4}$$

where $S(y_\tau) = (1-\tau) \int_{-\infty}^{y_\tau} \mid y - y_\tau \mid f(y)dy + \tau \int_{y_\tau}^{\infty} \mid y - y_\tau \mid f(y)dy$. Similarly, one can show that for other values of $c$ where $c \neq y_\tau$, $S(c) \geq S(y_\tau)$. Therefore, as $N \to \infty$, the solution (12.3.2) yields a consistent estimator of $y_\tau$.

Koenker and Bassett (1978) generalize the ordinary notion of sample quantiles based on an ordering sample observations to the regression framework

$$\min_{\mathbf{b}} \left\{ \sum_{i \in \psi_b} \tau \mid y_i - \mathbf{x}_i' \mathbf{b} \mid + \sum_{i \in \bar{\psi}_b} (1-\tau) \mid y_i - \mathbf{x}_i' \mathbf{b} \mid \right\}, \tag{12.3.5}$$

where $\psi_b = \{i \mid y_i \geq \mathbf{x}_i'\mathbf{b}\}$ and $\bar{\psi}_b = \{i \mid y_i < \mathbf{x}_i'\mathbf{b}\}$. When $\tau = \frac{1}{2}$, the quantile estimator (12.3.5) is the least absolute deviation estimator. Minimizing (12.3.5) can also be written in the form

$$\underset{\mathbf{b}}{\text{Min}} \sum_{i=1}^{N} \rho_\tau(y_i - \mathbf{x}_i\mathbf{b}), \tag{12.3.6}$$

where $\rho_\tau(u) := [\tau - 1(u \leq 0)]u$, and $1(A) = 1$ if $A$ occurs and 0 otherwise. Equation (12.3.6) is equivalent to the linear programming form,

$$\text{Min } [\tau \mathbf{e}'\mathbf{u}^+ + (1-\tau)\mathbf{e}'\mathbf{u}^-] \tag{12.3.7}$$

subject to

$$\mathbf{y} = X\mathbf{b} + \mathbf{u}^+ - \mathbf{u}^-, \tag{12.3.8}$$

$$(\mathbf{u}^+, \mathbf{u}^-) \in R_+^{2N}, \tag{12.3.9}$$

where $\mathbf{e}$ is an $N \times 1$ vector of $(1, \ldots, 1)$, $R_+^{2N}$ denotes the positive quadrant of the $2N$ dimensional real space such that if $u_i^+ > 0, u_i^- = 0$ and if $u_i^- > 0, u_i^+ = 0$. Sparse linear algebra and interior point methods for solving large linear programs are essential computational tools.

The quantile estimator for the panel data model,

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + u_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{matrix} \tag{12.3.10}$$

is the solution of

$$\underset{\mathbf{b}(\tau), \alpha_i(\tau)}{\text{Min}} \sum_{i=1}^{N} \sum_{t=1}^{T} \rho_\tau(y_{it} - \mathbf{x}_{it}'\mathbf{b}(\tau) - \alpha_i(\tau)), \tag{12.3.11}$$

where

$$Q_\tau(y_{it} \mid \mathbf{x}_{it}, \alpha_i) = \mathbf{x}_{it}'\mathbf{b}(\tau) + \alpha_i(\tau) \tag{12.3.12}$$

is the $\tau$th conditional quantile.

The main idea of regression quantile is to break up the common assumption that $u_{it}$ are independently, identically distributed. The conditional quantile (12.3.12) provides information on how $\mathbf{x}$ influence the location, scale and shape of the conditional distribution of the response. For instance,

$$u_{it} = (1 + \mathbf{x}_{it}'\boldsymbol{\gamma})\epsilon_{it}, \tag{12.3.13}$$

where $\mathbf{x}_{it}'\boldsymbol{\gamma} > 0$ and $\boldsymbol{\epsilon}_{it}$ has distribution function $F_\epsilon(\cdot)$. Then

$$\begin{aligned} Q_\tau(y_{it} \mid \mathbf{x}_{it}, \alpha_i) &= \mathbf{x}_{it}'\left(\boldsymbol{\beta} + \boldsymbol{\gamma} F_\epsilon^{-1}(\tau)\right) + \left(\alpha_i + F_\epsilon^{-1}(\tau)\right) \\ &= \mathbf{x}_{it}'\boldsymbol{\beta}(\tau) + \alpha_i(\tau). \end{aligned} \tag{12.3.14}$$

In other words, (12.3.14) is just a straightline describing the $\tau$th quantile of $y_{it}$ given $\mathbf{x}_{it}$. One should not confuse (12.3.14) with the traditional meaning of $E(y_{it} \mid \mathbf{x}_{it}, \alpha_i)$.

Kato, Galvao, and Montes-Rojas (2012) show that the quantile estimator of $(\mathbf{b}(\tau), \alpha_i(\tau))$ of (12.3.11) is consistent and asymptotically normally distributed provided $\frac{N^2(\log N)^3}{T} \longrightarrow 0$ as $N \longrightarrow \infty$. The requirement that the time-dimension of a panel, $T$, to grow much faster than the cross-sectional dimension, $N$, as $N$ increases is because directly estimating the individual-specific effects significantly increases the variability of the estimates of $\mathbf{b}(\tau)$. Standard linear transformation procedures such as first-differencing or mean differencing are not applicable in quantile regression. Koenker (2004) noted that shrinking the individual-specific effects toward a common mean can reduce the variability due to directly estimating the large number of individual-specific effects. He suggested a penalized version of (12.3.11),

$$\underset{\mathbf{b}(\tau),\alpha_i(\tau)}{\text{Min}} \sum_{i=1}^{N}\sum_{t=1}^{T} \rho_\tau(y_{it} - \mathbf{x}_{it}'\mathbf{b}(\tau) - \alpha_i(\tau)) + d\sum_{i=1}^{N} \mid \alpha_i(\tau) \mid \qquad (12.3.15)$$

The penalty $d \mid \sum_{i=1}^{N} \alpha_i(\tau) \mid$ serves to shrink the individual effects estimates toward zero. When $d \longrightarrow 0$, (12.3.15) yields the quantile fixed effects estimator (12.3.11). When $d \longrightarrow \infty, \hat{\alpha}_i(\tau) \longrightarrow 0$ for all $i = 1, \ldots, N$. Minimizing (12.3.15) leads to improved performance for the estimates of the slope parameter $\boldsymbol{\beta}(\tau)$.

One trouble with (12.3.11) or (12.3.15) is that the individual-specific effects could change because the realized value of $y_{it}$ at different time periods could fall into different quantiles. One way to get around this problem is to view the individual-specific effect summarizing the impact of some time-invariant latent variables while the error, $u_{it}$, bounces the responses $y_{it}$ around from quantile to quantile. In other words, we condition not only on the observed covariates, $\mathbf{x}_{it}$, but also on the individual fixed effects, $\alpha_i$, and replace the objective function (12.3.15) by pooling the estimates of individual quantile through

$$\text{Min} \sum_{j=1}^{J}\sum_{i=1}^{N}\sum_{t=1}^{T} \omega_j \rho_{\tau_j}(y_{it} - \mathbf{x}_{it}'\mathbf{b}(\tau) - \alpha_i) + d\sum_{i=1}^{N} \mid \alpha_i \mid, \quad (12.3.16)$$

where $\omega_j$ is a relative weight given to the $\tau_j$-th quantile. Monte Carlo studies show that shrinking the unconstrained individual-specific effects toward a common value helps to achieve improved performance for the estimates of the individual-specific effects and $\mathbf{b}(\tau_j)$.

Although introducing the penalty factor $d\sum_{i=1}^{N} \mid \alpha_i(\tau) \mid$ achieves the improved performance of panel quantile estimates, deciding $d$ is a challenging question. Lamarche (2010) shows that when the individual-specific effects $\alpha_i$ are independent of $\mathbf{x}_{it}$ the penalized quantile estimator is asymptotically unbiased and normally distributed if the individual-specific effects, $\alpha_i$, are drawn from a class of zero-median distribution functions. The regularization parameter, $d$, can thus be selected accordingly to minimize the estimated asymptotic variance.

## 12.4 SIMULATION METHODS

Panel data contains two dimensions – a cross-sectional dimension and a time dimension. Models using panel data also often contain unobserved heterogeneity factors. To transform a latent variable model involving missing data, random coefficients, heterogeneity, etc., into an observable model often requires the integration of latent variables over multiple dimensions (e.g., Hsiao 1989, 1991a,b, 1992a). The resulting panel data model estimators can be quite difficult to compute. Simulation methods have been suggested to get around the complex computational issues involving multiple integrations (e.g., Geweke 1991; Gourieroux and Monfort 1996; Hajivassiliou 1990; Hsiao and Wang 2000; Keane 1994; McFadden 1989; Pakes and Pollard 1989; and Richard and Zhang 2007).

The basic idea of simulation approach is to rely on the law of large numbers to obtain the approximation of the integrals through taking the averages of random drawings from a known probability distribution function. For instance, consider the problem of computing the conditional density function of $\mathbf{y}_i$ given $\mathbf{x}_i$, $f(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta})$ or some conditional moments $\mathbf{m}(\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\theta})$ say $E(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta})$ or $E(\mathbf{y}_i \mathbf{y}_i' \mid \mathbf{x}_i; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the vector of parameters characterizing these functions. In many cases, it is difficult to compute these functions because they do not have closed forms. However, if the conditional density or moments conditional on $\mathbf{x}$ and another vector $\boldsymbol{\eta}$, $f^*(\mathbf{y}_i \mid \mathbf{x}_i, \boldsymbol{\eta}; \boldsymbol{\theta})$ or $\mathbf{m}^*(\mathbf{y}, \mathbf{x} \mid \boldsymbol{\eta}; \boldsymbol{\theta})$, have closed forms and the probability distribution of $\boldsymbol{\eta}$, $P(\boldsymbol{\eta})$, is known, then from

$$f(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta}) = \int f^*(\mathbf{y}_i \mid \mathbf{x}_i, \boldsymbol{\eta}; \boldsymbol{\theta}) dP(\boldsymbol{\eta}), \tag{12.4.1}$$

and

$$\mathbf{m}(\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\theta}) = \int \mathbf{m}^*(\mathbf{y}_i, \mathbf{x}_i \mid \boldsymbol{\eta}; \boldsymbol{\theta}) dP(\boldsymbol{\eta}), \tag{12.4.2}$$

we may approximate (12.4.1) and (12.4.2) by

$$\tilde{f}_H(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta}) = \frac{1}{H} \sum_{h=1}^{H} f^*(\mathbf{y}_i \mid \mathbf{x}_i, \boldsymbol{\eta}_{ih}; \boldsymbol{\theta}), \tag{12.4.3}$$

and

$$\tilde{\mathbf{m}}_H(\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\theta}) = \frac{1}{H} \sum_{h=1}^{H} \mathbf{m}^*(\mathbf{y}_i, \mathbf{x}_i \mid \boldsymbol{\eta}_{ih}; \boldsymbol{\theta}), \tag{12.4.4}$$

where $(\boldsymbol{\eta}_{i1}, \ldots, \boldsymbol{\eta}_{iH})$ are $H$ random draws from $P(\boldsymbol{\eta})$.

For example, consider the random effects panel Probit and Tobit models defined by the latent response function

$$y_{it}^* = \boldsymbol{\beta}' \mathbf{x}_{it} + \alpha_i + u_{it}, \tag{12.4.5}$$

where $\alpha_i$ and $u_{it}$ are assumed to be independently normally distributed with mean 0 and variance $\sigma_\alpha^2$ and 1, respectively, and are mutually independent. The

Probit model assumes that the observed $y_{it}$ takes the form

$$y_{it} = \begin{cases} 1 & \text{if } y_{it}^* > 0, \\ 0 & \text{if } y_{it}^* \le 0. \end{cases} \tag{12.4.6}$$

The Tobit model assumes that

$$y_{it} = \begin{cases} y_{it}^* & \text{if } y_{it}^* > 0, \\ 0 & \text{if } y_{it}^* \le 0. \end{cases} \tag{12.4.7}$$

We note that the density function of $\alpha_i$ and $u_{it}$ can be expressed as transformations of some standard distributions, here, standard normal, so that the density function of $\mathbf{y}_i' = (y_{i1}, \ldots, y_{iT})$ becomes an integral of a conditional function over the range of these standard distributions A:

$$f(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta}) = \int_A f^*(\mathbf{y}_i \mid \mathbf{x}_i, \eta; \boldsymbol{\theta})\, dP(\eta) \tag{12.4.8}$$

with $p(\eta) \sim N(0, 1)$. For instance, in the case of Probit model,

$$f^*(\mathbf{y}_i \mid \mathbf{x}_i, \eta; \boldsymbol{\theta}) = \sum_{t=1}^{T} \Phi(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma_\alpha \eta_i)^{y_{it}} [1 - \Phi(\mathbf{x}_{it}'\boldsymbol{\beta} + \sigma_\alpha \eta_i)]^{1-y_{it}}, \tag{12.4.9}$$

and in the case of Tobit model,

$$\begin{aligned} f^*(\mathbf{y}_i \mid \mathbf{x}_i, \eta; \boldsymbol{\theta}) = & \prod_{t \epsilon \Psi_1} \phi(y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta} - \sigma_\alpha \eta_i) \\ & \cdot \prod_{t \epsilon \Psi_0} \Phi(-\mathbf{x}_{it}'\boldsymbol{\beta} - \sigma_\alpha \eta_i), \end{aligned} \tag{12.4.10}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal density and integrated normal, respectively, and $\Psi_1 = \{t \mid y_{it} > 0\}$ and $\Psi_0 = \{t \mid y_{it} = 0\}$. Since conditional on $\mathbf{x}_{it}$ and each of the $H$ random draws of $\eta$ from a standard normal distribution, $\eta_{ih}, h = 1, \ldots, H$, the conditional density function (12.4.9) on (12.4.10) is well defined in terms of $\boldsymbol{\beta}, \sigma_\alpha^2$, the approximation of $f(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\beta}, \sigma_\alpha^2)$ can be obtained by taking their averages as in (12.4.3).

Random draws of $\eta_h$ from $P(\eta)$ can be obtained through the *inversion* technique from a sequence of independent uniform [0, 1] pseudo-random draws

$$\eta_h = P^{-1}(\epsilon_h),$$

where $P^{-1}(\cdot)$ denote the inverse of $P$. For instance, if $\epsilon$ is normally distributed with mean $\mu$ and variance $\sigma_\epsilon^2$, then $\eta_h = \Phi^{-1}(\frac{\epsilon_h - \mu}{\sigma_\epsilon})$. If $\eta$ is a Weibull random variable with parameters $a$ and $b$, $P(\eta_h) = 1 - \exp(-b\eta_h^a)$, then $\eta_h = \left[-\frac{1}{b} ln \epsilon_h\right]^{\frac{1}{a}}$.

The generation of a multivariate $\boldsymbol{\eta}_h$ can be obtained through recursive factorization of its density into lower dimensional density (e.g., Liesenfeld and Richard 2008). The basic idea of factorization of a $k$-dimensional

$\boldsymbol{\eta}_h = (\eta_{1h}, \ldots, \eta_{kh})$ is to write

$$P(\boldsymbol{\eta}_h) = P(\eta_{kh} \mid \boldsymbol{\eta}^*_{k-1,h}) P(\eta_{k-1,h} \mid \boldsymbol{\eta}^*_{k-2,h}) \ldots P(\eta_{2h} \mid \eta_{1h}) P(\eta_{1h}), \tag{12.4.11}$$

where $\eta^*_{j,h} = (\eta_{1h}, \ldots, \eta_{jh})$. For example, random draws from a multivariate normal density are typically obtained based on Cholesdy decomposition of its covariance matrix $\sum = \bigwedge \bigwedge'$, $\boldsymbol{\eta}_h = \bigwedge \boldsymbol{\eta}^*_h$, where $\bigwedge$ is a lower triangular matrix and $\boldsymbol{\eta}^*$ is standard multivariate normal with identity covariance matrix.

A particularly useful technique for evaluating high-dimensional integrals is known as *Importance Sampling*. The idea of importance sampling is to replace $P(\boldsymbol{\eta}_i)$ by an alternative simulator with density $\mu(\cdot)$. Substituting $\mu(\cdot)$ into (12.4.11)

$$f(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta}) = \int f^*(\mathbf{y}_i \mid \mathbf{x}_i, \boldsymbol{\eta}; \boldsymbol{\theta}) \omega(\boldsymbol{\eta}_i) \mu(\boldsymbol{\eta}_i)\, d\boldsymbol{\eta}_i, \tag{12.4.12}$$

where $dP(\boldsymbol{\eta}_i) = p(\boldsymbol{\eta}_i)\, d\boldsymbol{\eta}_i$,

$$\omega(\boldsymbol{\eta}_i) = \frac{p(\boldsymbol{\eta}_i)}{\mu(\boldsymbol{\eta}_i)}. \tag{12.4.13}$$

Then the corresponding Monte Carlo simulator of (12.4.3), known as the "importance sampling" estimator, is given by

$$\tilde{f}_H(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta}) = \frac{1}{H} \sum_{h=1}^{H} \omega(\boldsymbol{\eta}^*_{ih}) \mu(\boldsymbol{\eta}^*_{ih}) f^*(\mathbf{y}_i \mid \mathbf{x}_i, \boldsymbol{\eta}^*_{ih}; \boldsymbol{\theta}), \tag{12.4.14}$$

where $\boldsymbol{\eta}^*_{ih}$ are random draws from $\mu(\boldsymbol{\eta}^*_i)$.

If $u_{it}$ in the above example follows a first-order autoregressive process

$$u_{it} = \rho u_{i,t-1} + \epsilon_{it}, \quad \mid \rho \mid < 1, \tag{12.4.15}$$

then we can rewrite (12.4.5) as

$$y^*_{it} = \boldsymbol{\beta}' \mathbf{x}_{it} + \sigma_\alpha \eta_i + \sum_{\tau=1}^{t} a_{t\tau} \eta^*_{i\tau}, \tag{12.4.16}$$

where $\eta^*_{i\tau}, \tau = 1, \ldots, T$ are random draws from independent $N(0, 1)$, and $a_{t\tau}$ are the entries of the lower triangular matrix $\Lambda$. It turns out that here $a_{t\tau} = (1 - \rho^2)^{-\frac{1}{2}} \rho^{t-\tau}$ if $t \geq \tau$ and $a_{t\tau} = 0$ if $t < \tau$.

Using the approach described above, we can obtain an unbiased, differentiable and positive simulator of $f(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta})$, $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma_\alpha, \rho)'$, in the Probit case by considering the following drawings:

$\eta_{ih}$ is drawn from N(0,1).
$\eta^*_{i1h}$ is drawn from N(0,1) restricted to

$[-(\boldsymbol{\beta}'\mathbf{x}_{i1} + \sigma_\alpha \eta_{ih})/a_{11}, \infty]$ if $y_{i1} = 1$ or $[-\infty, -(\boldsymbol{\beta}'\mathbf{x}_{i1} + \sigma_\alpha \eta_{ih})/a_{11}]$ if $y_{i1} = 0$, $\eta^*_{i2h}$ is drawn from N(0,1) restricted to

$$[-(\boldsymbol{\beta}'\mathbf{x}_{i2} + \sigma_\alpha \eta_{ih} + a_{21}\eta^*_{i1h})/a_{22}, \infty] \text{ if } y_{i2} = 1,$$

and

$$[-\infty, -(\boldsymbol{\beta}'\mathbf{x}_{i2} + \sigma_\alpha \eta_{ih} + a_{21}\eta^*_{i1h})/a_{22}] \text{ if } y_{i2} = 0,$$

and so on. The simulator of $f(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta})$ is

$$\tilde{f}_H(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta}) = \frac{1}{H}\sum_{h=1}^{H}\prod_{t=1}^{T}\Phi\left[(-1)^{1-y_{it}}\left(\boldsymbol{\beta}'\mathbf{x}_{it} + \sigma_\alpha\eta_{ih} + \sum_{\tau=1}^{t-1} a_{t\tau}\eta^*_{i\tau h}\right)/a_{tt}\right], \tag{12.4.17}$$

where for $t = 1$, the sum over $\tau$ disappears.

In the Tobit case, the same kind of method can be used. The only difference is that the simulator of $f(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta})$ becomes

$$\tilde{f}_H(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta}) = \frac{1}{H}\sum_{i=1}^{H}\left[\prod_{t\in\Psi_1}\frac{1}{a_{tt}}\phi\left(\left[y_{it} - \left(\boldsymbol{\beta}'\mathbf{x}_{it} + \sigma_\alpha\eta_{ih} + \sum_{\tau=1}^{t-1} a_{t\tau}\eta^*_{i\tau h}\right)\right]/a_{tt}\right)\right]$$
$$\cdot \prod_{t\in\Psi_0}\Phi\left[-\left(\boldsymbol{\beta}'\mathbf{x}_{it} + \sigma_\alpha\eta_{ih} + \sum_{\tau=1}^{t-1} a_{t\tau}\eta^*_{i\tau h}\right)/a_{tt}\right]. \tag{12.4.18}$$

The simulated maximum likelihood estimator (SMLE) is obtained from maximizing the simulated log-likelihood function. The simulated method of moments estimator (SMM) is obtained from the simulated moments. The simulated least squares estimator (SLS) is obtained if we let $\mathbf{m}(\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\theta}) = E(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta})$ and minimize $\sum_{i=1}^{N}[\mathbf{y}_i - E(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta})]^2$.

Although we need $H \to \infty$ to obtain consistent simulator of $f(\mathbf{y}_i \mid \mathbf{x}_i; \boldsymbol{\theta})$ and $\mathbf{m}(\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\theta})$, it is shown by McFadden (1989) that when finite $H$ vectors $(\boldsymbol{\eta}_{i1}, \ldots, \boldsymbol{\eta}_{iH})$ are drawn by simple random sampling and independently for different $i$ from the marginal density $P(\boldsymbol{\eta})$, the simulation errors are independent across observations; hence the variance introduced by simulation will be controlled by the law of large numbers operating across observations, making it unnecessary to consistently estimate each theoretical $\mathbf{m}(\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\theta})$ for the consistency of SMM, $\hat{\boldsymbol{\theta}}_{\text{SGMM}}$, as $N \to \infty$.

The asymptotic covariance matrix of $\sqrt{N}(\hat{\boldsymbol{\theta}}_{\text{SMM}} - \boldsymbol{\theta})$ obtained from minimizing $[\hat{\mathbf{m}}(\boldsymbol{\theta}) - \mathbf{m}(\boldsymbol{\theta})]'A[\hat{\mathbf{m}}(\boldsymbol{\theta}) - \mathbf{m}(\boldsymbol{\theta})]$ where $A$ is a positive definite matrix such as moments of the form (4.3.38) can be approximated by

$$(R'AR)^{-1}R'AG_{NH}AR(R'AR)^{-1}, \tag{12.4.19}$$

where

$$
\begin{aligned}
R &= \frac{1}{N}\sum_{i=1}^{N} W_i' \frac{\partial \tilde{\mathbf{m}}_H(\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}, \\
G_{NH} &= \frac{1}{N}\sum_{i=1}^{N} W_i \left(\Omega + \frac{1}{H}\Delta_H\right) W_i', \qquad (12.4.20) \\
\Omega &= \text{Cov}\,(\mathbf{m}_i(\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\theta})) \\
\Delta_H &= \text{Cov}\,[\tilde{\mathbf{m}}_H(\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\theta}) - \mathbf{m}(y_i, \mathbf{x}_i; \boldsymbol{\theta})],
\end{aligned}
$$

and $W_i$ is given by (4.3.41). When $A = [\text{plim Cov}\,(\hat{\mathbf{m}}_i(y_i, \mathbf{x}_i; \boldsymbol{\theta})]^{-1}$, the SMM is the simulated generalized method of moments estimator (SGMM). It is clear that as $H \to \infty$, the SGMM has the same asymptotic efficiency as the GMM. However, even with finite $H$, the relative efficiency of SGMM is quite high. For instance, for the simple frequency simulator, $\Delta_H = \Omega$, one draw per observation gives fifty percent of the asymptotic efficiency of the corresponding GMM estimator, and nine draws per observation gives 90 percent relative efficiency.

The consistency of SMLE or SLS needs consistently estimated conditional density or moments. With a finite $H$, the approximation error of the conditional density or moments is of order $H^{-1}$. This will lead to the asymptotic bias of $O(1/H)$ (e.g., Gourieroux and Monfort 1996; Hsiao, Wang, and Wang 1997). Nevertheless, with a finite H it is still possible to propose SLS estimator that is consistent and asymptotically normally distributed as $N \to \infty$ by noting that for the sequence of 2H random draws $(\boldsymbol{\eta}_{i1}, \ldots, \boldsymbol{\eta}_{iH}, \boldsymbol{\eta}_{i,H+1}, \ldots, \boldsymbol{\eta}_{i,2H})$ for each $i$,

$$
\begin{aligned}
E\left[\frac{1}{H}\sum_{h=1}^{H} \mathbf{m}^*(\mathbf{y}_i, \mathbf{x}_i \mid \boldsymbol{\eta}_{ih}; \boldsymbol{\theta})\right] &= E\left[\frac{1}{H}\sum_{h=1}^{H} \mathbf{m}^*(\mathbf{y}_i, \mathbf{x}_i \mid \boldsymbol{\eta}_{i,H+h}; \boldsymbol{\theta})\right] \\
&= \mathbf{m}(\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\theta}), \qquad (12.4.21)
\end{aligned}
$$

and

$$
\begin{aligned}
&E\left[\mathbf{y}_i - \frac{1}{H}\sum_{h=1}^{H} \mathbf{m}^*(\mathbf{y}_i, \mathbf{x}_i \mid \boldsymbol{\eta}_{ih}; \boldsymbol{\theta})\right]' \left[\mathbf{y}_i - \frac{1}{H}\sum_{h=1}^{H} \mathbf{m}^*(\mathbf{y}_i, \mathbf{x}_i \mid \boldsymbol{\eta}_{i,H+h}; \boldsymbol{\theta})\right] \qquad (12.4.22) \\
&\quad = E\left[\mathbf{y}_i - \mathbf{m}(\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\theta})\right]' \left[\mathbf{y}_i - \mathbf{m}(\mathbf{y}_i, \mathbf{x}_i; \boldsymbol{\theta})\right],
\end{aligned}
$$

because of the independence between $(\boldsymbol{\eta}_{i1}, \ldots, \boldsymbol{\eta}_{iH})$ and $(\boldsymbol{\eta}_{i,H+1}, \ldots, \boldsymbol{\eta}_{i,2H})$. Then the SLS estimator that minimizes

$$
\sum_{i=1}^{N}\left[\mathbf{y}_i - \frac{1}{H}\sum_{h=1}^{H} \mathbf{m}^*(\mathbf{y}_i, \mathbf{x}_i \mid \boldsymbol{\eta}_{ih}; \boldsymbol{\theta})\right]' \left[\mathbf{y}_i - \frac{1}{H}\sum_{h=1}^{H} \mathbf{m}^*(y_i, \mathbf{x}_i \mid \boldsymbol{\eta}_{i,H+h}; \boldsymbol{\theta})\right] \qquad (12.4.23)
$$

is consistent as $N \to \infty$ even $H$ is fixed (e.g., Gourieroux and Monfort 1996; Hsiao and Wang 2000).

## 12.5 DATA WITH MULTILEVEL STRUCTURES

We have illustrated panel data methodology by assuming the presence of individual and/or time effects only. However, panel data need not be restricted to two dimensions. We can have a more complicated "clustering" or "hierarchical" structure. For example, Antweiler (2001), Baltagi, Song, and Jung (2001), and Davis (2002), following the methodology developed by Wansbeek (1982) Wansbeek and Kapteyn (1978), consider the multiway error components model of the form

$$y_{ij\ell t} = \mathbf{x}'_{ij\ell t}\boldsymbol{\beta} + v_{ij\ell t}, \tag{12.5.1}$$

for $i = 1, \ldots, N$, $j = 1, \ldots, M_i$, $\ell = 1, \ldots, L_{ij}$, and $t = 1, \ldots, T_{ij\ell}$. For example, the dependent variable $y_{ij\ell t}$ could denote the air pollution measured at station $\ell$ in city $j$ of country $i$ in time period $t$. This means that there are $N$ countries, and each country $i$ has $M_i$ cities in which $L_{ij}$ observation stations are located. At each station, air pollution is observed for $T_{ij\ell}$ periods. The $\mathbf{x}_{ij\ell t}$ denotes a vector of $K$ explanatory variables, and the disturbance is assumed to have a multiway error components structure,

$$v_{ij\ell t} = \alpha_i + \lambda_{ij} + \nu_{ij\ell} + \epsilon_{ij\ell t} \tag{12.5.2}$$

where $\alpha_i$, $\lambda_{ij}$, $\nu_{ij\ell}$ and $\epsilon_{ij\ell t}$ are assumed to be independently, identically distributed and are mutually independent with mean 0 and variances $\sigma_\alpha^2$, $\sigma_\lambda^2$, $\sigma_\nu^2$, and $\sigma_\epsilon^2$, respectively.

In the case that the data are balanced, the variance–covariance matrix of $\mathbf{v}$, has the form

$$\Omega = \sigma_\alpha^2(I_N \otimes J_{MLT}) + \sigma_\lambda^2(I_{NM} \otimes J_{LT}) + \sigma_\nu^2(I_{NML} \otimes J_T) + \sigma_\epsilon^2 I_{LMNT}, \tag{12.5.3}$$

where $J_s$ be a square matrix of dimension $s$ with all elements equal to 1. Rewrite (12.5.3) in the form representing the spectral decomposition $\Omega$ (e.g., as in Appendix 3B), we have

$$\begin{aligned}\Omega &= MLT\sigma_\alpha^2(I_N \otimes P_{MLT}) + LT\sigma_\lambda^2(I_{NM} \otimes P_{LT}) \\ &\quad + T\sigma_\nu^2(I_{NML} \otimes P_T) + \sigma_\epsilon^2 I_{LMNT} \\ &= \sigma_\epsilon^2(I_{NML} \otimes Q_T) + \sigma_1^2(I_{NM} \otimes Q_L \otimes P_T) \\ &\quad + \sigma_2^2(I_N \otimes Q_M \otimes P_{LT}) + \sigma_3^2(I_N \otimes P_{MLT})\end{aligned} \tag{12.5.4}$$

where $P_s \equiv \frac{1}{s}J_s$, $Q_s = I_s - P_s$, and

$$\sigma_1^2 = T\sigma_\nu^2 + \sigma_\epsilon^2, \tag{12.5.5}$$

$$\sigma_2^2 = LT\sigma_\lambda^2 + T\sigma_\nu^2 + \sigma_\epsilon^2, \tag{12.5.6}$$

$$\sigma_3^2 = MLT\sigma_\alpha^2 + LT\sigma_\lambda^2 + T\sigma_\nu^2 + \sigma_\epsilon^2, \tag{12.5.7}$$

$\sigma_\epsilon^2$ are the characteristic roots of $\Omega$. As each of the terms of (12.5.4) is orthogonal to each other and sum to $I_{NMLT}$, it follows that

$$\begin{aligned}\Omega^{-1/2} = {} & \sigma_\epsilon^{-1}(I_{NML} \otimes Q_T) + \sigma_1^{-1}(I_{NM} \otimes Q_L \otimes P_T) \\ & + \sigma_2^{-1}(I_N \otimes Q_M \otimes P_{LT}) + \sigma_3^{-1}(I_N \otimes P_{MLT}) \end{aligned} \tag{12.5.8}$$

Expanding all the $Q$ matrices as the difference of $I$ and $P$, multiplying both sides of the equation by $\sigma_\epsilon$, and collecting terms yield

$$\begin{aligned}\sigma_\epsilon \Omega^{-1/2} = {} & I_{NMLT} - \left(1 - \frac{\sigma_\epsilon}{\sigma_1}\right)(I_{\mathrm{NML}} \otimes P_T) \\ & - \left(\frac{\sigma_\epsilon}{\sigma_1} - \frac{\sigma_\epsilon}{\sigma_2}\right)(I_{NM} \otimes P_{\mathrm{LT}}) \\ & - \left(\frac{\sigma_\epsilon}{\sigma_2} - \frac{\sigma_\epsilon}{\sigma_3}\right)(I_N \otimes P_{\mathrm{MLT}})\,. \end{aligned} \tag{12.5.9}$$

The generalized least-squares estimator (GLS) of (12.5.1) is equivalent to the least squares estimator of

$$y^*_{ij\ell t} = y_{ij\ell t} - \left(1 - \frac{\sigma_\epsilon}{\sigma_1}\right)\bar{y}_{ij\ell\cdot} - \left(\frac{\sigma_\epsilon}{\sigma_1} - \frac{\sigma_\epsilon}{\sigma_2}\right)\bar{y}_{ij\cdot\cdot} - \left(\frac{\sigma_\epsilon}{\sigma_2} - \frac{\sigma_\epsilon}{\sigma_3}\right)\bar{y}_{i\cdots}, \tag{12.5.10}$$

on

$$\mathbf{x}^*_{ij\ell t} = x_{ij\ell t} - \left(1 - \frac{\sigma_\epsilon}{\sigma_1}\right)\bar{\mathbf{x}}_{ij\ell\cdot} - \left(\frac{\sigma_\epsilon}{\sigma_1} - \frac{\sigma_\epsilon}{\sigma_2}\right)\bar{\mathbf{x}}_{ij\cdot\cdot} - \left(\frac{\sigma_\epsilon}{\sigma_2} - \frac{\sigma_\epsilon}{\sigma_3}\right)\bar{\mathbf{x}}_{i\cdots}, \tag{12.5.11}$$

where $\bar{y}_{ij\ell\cdot}(\bar{\mathbf{x}}_{ij\ell\cdot})$, $\bar{y}_{ij\cdot\cdot}(\bar{\mathbf{x}}_{ij\cdot\cdot})$ and $\bar{y}_{i\cdots}(\bar{\mathbf{x}}_{i\cdots})$ indicate group averages. The application of feasible GLS can be carried out by replacing the variances in (12.5.10) and (12.5.11) by their estimates obtained from the three groupwise between estimates and the within estimate of the innermost group.

The pattern exhibited in (12.5.10) and (12.5.11) is suggestive of solutions for higher order hierarchy with a balanced structure. If the hierarchical structure is unbalanced, Kronecker product operation can no longer be applied. It introduces quite a bit of notational inconvenience into the algebra (e.g., Baltagi (1995, Chapter 9) and Wansbeek and Kapteyn (1978)). Neither can the GLS estimator be molded into a simple transformation for least-squares estimator. However, an unbalanced panel is made up of $N$ top level groups, each containing $M_i$ second-level groups, the second-level groups containing the innermost $L_{ij}$ subgroups, which in turn containing $T_{ij\ell}$ observations, the number of observations in the higher-level groups are thus $T_{ij} = \sum_{\ell=1}^{L_{ij}} T_{ij\ell}$ and $T_i = \sum_{j=1}^{M_i} T_{ij}$, and the total number of observations is $H = \sum_{i=1}^{N} T_i$. The number of top-level groups is $N$, the number of second level groups is $F = \sum_{i=1}^{N} M_i$, and the bottom-level groups is $G = \sum_{i=1}^{N}\sum_{j=1}^{M_i} L_{ij}$. We can redefine $J$ matrices to

be block diagonal of size $H \times H$, corresponding in structure to the groups or subgroups they represent. They can be constructed explicitly by using "group membership" matrices consisting of 1's and 0's that uniquely assign each of the $H$ observations to one of the $G$ (or $F$ or $N$) groups. Antweiler (2001) has derived the maximum-likelihood estimator for the panels with unbalanced hierarchy.

When data contains a multilevel hierarchical structure, the application of a simple error component estimation, although inefficient, remains consistent under the assumption that the error component is independent of the regressors. However, the estimated standard errors of the slope coefficients are usually biased downward.

## 12.6 ERRORS OF MEASUREMENT

Thus far we have assumed that variables are observed without errors. Economic quantities, however, are frequently measured with errors, particularly if longitudinal information is collected through one-time retrospective surveys, which are notoriously susceptible to recall errors. If variables are indeed subject to measurement errors, exploiting panel data to control for the effects of unobserved individual characteristics using standard differenced estimators (deviations from means, etc.) may result in even more biased estimates than simple least-squares estimators using cross-sectional data alone.

Consider, for example, the following single-equation model (Solon 1985):

$$y_{it} = \alpha_i^* + \beta x_{it} + u_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{matrix} \tag{12.6.1}$$

where $u_{it}$ is independently identically distributed, with mean 0 and variance $\sigma_u^2$, and $\text{Cov}(x_{it}, u_{is}) = \text{Cov}(\alpha_i^*, u_{it}) = 0$ for any $t$ and $s$, but $\text{Cov}(x_{it}, \alpha_i^*) \neq 0$. Suppose further that we observe not $x_{it}$ itself, but rather the error-ridden measure

$$x_{it}^* = x_{it} + \tau_{it}, \tag{12.6.2}$$

where $\text{Cov}(x_{is}, \tau_{it}) = \text{Cov}(\alpha_i^*, \tau_{it}) = \text{Cov}(u_{it}, \tau_{is}) = 0$, and $\text{Var}(\tau_{it}) = \sigma_\tau^2$, $\text{Cov}(\tau_{it}, \tau_{i,t-1}) = \gamma_\tau \sigma_\tau^2$.

If we estimate (12.6.1) by ordinary least-squares (OLS) with cross-sectional data for period $t$, the estimator converges to (as $N \to \infty$)

$$\operatorname*{plim}_{N\to\infty} \hat{\beta}_{LS} = \beta + \frac{\text{Cov}(x_{it}, \alpha_i^*)}{\sigma_x^2 + \sigma_\tau^2} - \frac{\beta \sigma_\tau^2}{\sigma_x^2 + \sigma_\tau^2}, \tag{12.6.3}$$

where $\sigma_x^2 = \text{Var}(x_{it})$. The inconsistency of the least-squares estimator involves two terms, the first due to the failure to control for the individual effects $\alpha_i^*$ and the second due to measurement error.

If we have panel data, say $T = 2$, we can alternatively first difference the data to eliminate the individual effects, $\alpha_i^*$,

$$y_{it} - y_{i,t-1} = \beta(x_{it}^* - x_{i,t-1}^*) + [(u_{it} - \beta\tau_{it}) - (u_{i,t-1} - \beta\tau_{i,t-1})], \tag{12.6.4}$$

and then apply least squares. The probability limit of the differenced estimator as $N \to \infty$ becomes

$$\begin{aligned} \operatorname*{plim}_{N\to\infty} \hat{\beta}_d &= \beta\left[1 - \frac{2(1-\gamma_\tau)\sigma_\tau^2}{\operatorname{Var}(x_{it}^* - x_{i,t-1}^*)}\right] \\ &= \beta - \frac{\beta\sigma_\tau^2}{[(1-\gamma_x)/(1-\gamma_\tau)]\sigma_x^2 + \sigma_\tau^2}, \end{aligned} \tag{12.6.5}$$

where $\gamma_x$ is the first-order serial-correlation coefficient of $x_{it}$. The estimator $\hat{\beta}_d$ eliminates the first source of inconsistency, but may aggravate the second. If $\gamma_x > \gamma_\tau$, the inconsistency due to measurement error is larger for $\hat{\beta}_d$ than for $\hat{\beta}_{LS}$. This occurs because if the serial correlation of the measurement error is less than that of the true $x$ (as seems often likely to be the case), first differencing increases the noise-to-signal ratio for the measured explanatory variable.

The standard treatment for the errors-in-variables models requires extraneous information in the form of either additional data (replication and/or instrumental variables) or additional assumptions to identify the parameters of interest (e.g., Aigner et al. (1984)). The repeated measurement property of panel data allows a researcher to use different transformations of the data to induce different and deducible changes in the biases in the estimated parameters that can then be used to identify the importance of measurement errors and recover the "true" parameters (Ashenfelter, Deaton, and Solon (1984); Griliches and Hausman (1986). For instance, if the measurement error, $\tau_{it}$, is independently identically distributed across $i$ and $t$ and $x$ is serially correlated, then in the foregoing example we can use $x_{i,t-2}^*$ or $(x_{i,t-2}^* - x_{i,t-3}^*)$ as instruments for $(x_{it}^* - x_{i,t-1}^*)$ as long as $T > 3$. Thus, even though $T$ may be finite, the resulting IV estimator is consistent when $N$ tends to infinity.

Alternatively, we can obtain consistent estimates through a comparison of magnitudes of the bias arrived at by subjecting a model to different transformations (Griliches and Hausman 1986). For instance, if we use a covariance transformation to eliminate the contributions of unobserved individual components, we have

$$(y_{it} - \bar{y}_i) = \beta(x_{it}^* - \bar{x}_i^*) + [(u_{it} - \bar{u}_i) - \beta(\tau_{it} - \bar{\tau}_i)], \tag{12.6.6}$$

where $\bar{y}_i, \bar{x}_i^*, \bar{u}_i$, and $\bar{\tau}_i$ are individual time means of respective variables. Under the assumption that the measurement errors are independently identically distributed, the LS regression of (12.6.6) converges to

$$\operatorname*{plim}_{N\to\infty} \beta_w = \beta\left[1 - \frac{T-1}{T}\frac{\sigma_\tau^2}{\operatorname{Var}(x_{it}^* - \bar{x}_i^*)}\right]. \tag{12.6.7}$$

Then consistent estimators of $\beta$ and $\sigma_\tau^2$ can be solved from (12.6.5) and (12.6.7),

$$\hat{\beta} = \left[ \frac{2\hat{\beta}_w}{\text{Var}\,(x_{it}^* - x_{i,t-1}^*)} - \frac{(T-1)\hat{\beta}_d}{T\ \text{Var}(x_{it}^* - \bar{x}_i^*)} \right] \cdot \left[ \frac{2}{\text{Var}(x_{it}^* - x_{i,t-1}^*)} - \frac{T-1}{T\ \text{Var}(x_{it}^* - \bar{x}_i^*)} \right]^{-1}, \tag{12.6.8}$$

$$\hat{\sigma}_\tau^2 = \frac{\hat{\beta} - \hat{\beta}_d}{\hat{\beta}} \cdot \frac{\text{Var}(x_{it}^* - x_{i,t-1}^*)}{2}. \tag{12.6.9}$$

In general, if the measurement errors are known to possess certain structures, consistent estimators may be available from a method of moments and/or from an IV approach by utilizing the panel structure of the data. Moreover, the first difference and the within estimators are not the only ones that will give us an implicit estimate of the bias. In fact, there are $T/2$ such independent estimates. For a six-period cross section with $\tau_{it}$ independently identically distributed, we can compute estimates of $\beta$ and $\sigma_\tau^2$ from $y_6 - y_1$, $y_5 - y_2$, and $y_4 - y_3$ using the relationships

$$\begin{aligned} \operatorname*{plim}_{N\to\infty} \hat{\beta}_{61} &= \beta - 2\beta\sigma_\tau^2/\text{Var}(x_{i6}^* - x_{i1}^*), \\ \operatorname*{plim}_{N\to\infty} \hat{\beta}_{52} &= \beta - 2\beta\sigma_\tau^2/\text{Var}(x_{i5}^* - x_{i2}^*), \\ \operatorname*{plim}_{N\to\infty} \hat{\beta}_{43} &= \beta - 2\beta\sigma_\tau^2/\text{Var}(x_{i4}^* - x_{i3}^*). \end{aligned} \tag{12.6.10}$$

Thus, there are alternative consistent estimators. This fact can be exploited to test the assumption with regard to measurement errors,which provide the rationale for the validity of the instruments, by comparing whether or not the alternative estimates of $\beta$ are mutually coherent (e.g., Griliches and Hausman 1986). The moment conditions (12.6.5), (12.6.7), and (12.6.10) can also be combined together to obtain efficient estimates of $\beta$ and $\sigma_\tau^2$ by the use of Chamberlain $\pi$ method (Chapter 3, Section 3.8) or generalized method of moments estimator.

For instance, transforming $\mathbf{y}$ and $\mathbf{x}$ by the transformation matrix $P_s$ such that $P_s\mathbf{e}_T = \mathbf{0}$ eliminates the individual effects from the model (12.6.1). Regressing the transformed $\mathbf{y}$ on transformed $\mathbf{x}$ yields estimator that is a function of $\beta, \sigma_x^2, \sigma_\tau$ and the serial correlations of $x$ and $\tau$. Wansbeek and Koning (1989) have provided a general formula for the estimates based on various transformation of the data by letting

$$Y^* = \mathbf{e}_{NT}\mu + X^*\boldsymbol{\beta} + \mathbf{v}^* \tag{12.6.11}$$

where $Y^* = (\mathbf{y}_1^{*'}, \ldots, \mathbf{y}_T^{*'})'$, $\mathbf{y}_t^* = (y_{1t}, \ldots, y_{Nt})'$, $X^* = (\mathbf{x}_1^*, \ldots, \mathbf{x}_T^*)'$, $\mathbf{x}_t^* = (\mathbf{x}_{1t}', \ldots, \mathbf{x}_{Nt}')$, $\mathbf{v}^* = (\mathbf{v}_1^{*'}, \ldots, \mathbf{v}_T^{*'})'$, and $\mathbf{v}_t^* = (v_{1t}. \ldots, v_{NT})'$. Then

$$\begin{aligned}\hat{\mathbf{b}}_s &= [X^{*'}(Q_s \otimes I_N)X^*]^{-1}[X^{*'}(Q_s \otimes I_N)Y^*] \\ &= \boldsymbol{\beta} + [X^{*'}(Q_s \otimes I_N)X^*]^{-1}[X^{*'}(Q_s \otimes I_N)(\mathbf{u}^* - \boldsymbol{\tau}^*\boldsymbol{\beta})],\end{aligned} \tag{12.6.12}$$

where $Q_s = P_s'P_s$, $\mathbf{u}^* = (\mathbf{u}_1^{*'}, \ldots, \mathbf{u}_T^{*'})'$, $\mathbf{u}_t^* = (u_{1t}, \ldots, u_{Nt})'$, $\boldsymbol{\tau}^* = (\boldsymbol{\tau}_1^*, \ldots, \boldsymbol{\tau}_T^{*'})'$, and $\boldsymbol{\tau}_t^* = (\boldsymbol{\tau}_{1t}, \ldots, \boldsymbol{\tau}_{Nt})$. In the case of $K = 1$ and measurement errors are serially uncorrelated, Wansbeek and Koning (1989) show that the $m$ different transformed estimators $\mathbf{b} = (b_1, \ldots, b_m)'$

$$\sqrt{N}(\mathbf{b} - \beta(\mathbf{e}_m - \sigma_\tau^2\boldsymbol{\phi})] \sim N(\mathbf{0}, V), \tag{12.6.13}$$

where $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)'$, $\phi_s = (tr\, Q_s / tr\, Q_s \Sigma_{x^*})$,

$$\Sigma_{x^*} = \text{Cov}\,(\mathbf{x}_i^*),\ x_i^* = (\mathbf{x}_{i1}^*. \ldots, \mathbf{x}_{iT}^*)',$$

$$V = F'\{\sigma_u^2\Sigma_{x^*} \otimes I_T + \beta^2\sigma_\tau^2(\Sigma_{x^*} + \sigma_\tau^2 I_T) \otimes I_T\}F,$$

and $F$ is the $T^2 \times m$ matrix with the $s$th column $\mathbf{f}_s = \text{vec}\,(Q_s)/(tr\, Q_s \Sigma_{x^*})$, where vec (A) denotes the operation of transforming an $m \times n$ matrix A into the $mn \times 1$ vector by stacking the columns of A one underneath the other (Magnus and Neudecker (1999, p. 30). Then one can obtain an efficient estimator by minimizing

$$[\mathbf{b} - \beta(\mathbf{e}_m - \sigma_\tau^2\boldsymbol{\phi})]'V^{-1}[\mathbf{b} - \beta(\mathbf{e}_m - \sigma_\tau^2\boldsymbol{\phi})], \tag{12.6.14}$$

with respect to $\beta$ and $\sigma_\tau^2$, which yields

$$\hat{\beta} = \left\{\frac{\boldsymbol{\phi}'V^{-1}\mathbf{b}}{\boldsymbol{\phi}'V^{-1}\boldsymbol{\phi}} - \frac{\mathbf{e}_m'V^{-1}\mathbf{b}}{\mathbf{e}_m'V^{-1}\boldsymbol{\phi}}\right\} \Big/ \left\{\frac{\boldsymbol{\phi}'V^{-1}\mathbf{e}}{\boldsymbol{\phi}'V^{-1}\boldsymbol{\phi}} - \frac{\mathbf{e}_m'V^{-1}\mathbf{e}_m}{\mathbf{e}_m'V^{-1}\boldsymbol{\phi}}\right\} \tag{12.6.15}$$

and

$$\sigma_\tau^2 = \left\{\frac{\boldsymbol{\phi}'V^{-1}\mathbf{e}_m}{\boldsymbol{\phi}'V^{-1}\mathbf{b}} - \frac{\mathbf{e}_m'V^{-1}\mathbf{e}_m}{\mathbf{e}_m'V^{-1}\mathbf{b}}\right\} \Big/ \left\{\frac{\boldsymbol{\phi}'V^{-1}\boldsymbol{\phi}}{\boldsymbol{\phi}'V^{-1}\mathbf{b}} - \frac{\mathbf{e}_m'V^{-1}\boldsymbol{\phi}}{\mathbf{e}_m'V^{-1}\mathbf{b}}\right\}. \tag{12.6.16}$$

Extensions of this simple model to the serially correlated measurement errors are given by Biørn (2000), Hsiao and Taylor (1991). Wansbeek and Kapteyn (1978) consider simple estimators for dynamic panel data models with measurement errors. In the case of only one regressor for a linear panel data model, Wansbeek (2001) has provided a neat framework to derive the moment conditions under a variety of measurement errors assumption by stacking the matrix of covariances between the vector of dependent variables over time and the regressors, then projecting out nuisance parameters. To illustrate the basic idea, consider a linear model,

$$y_{it} = \alpha_i^* + \beta x_{it} + \boldsymbol{\gamma}'\mathbf{w}_{it} + \mathbf{u}_{it} \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{matrix} \tag{12.6.17}$$

where $x_{it}$ is not observed. Instead one observes $x_{it}^*$, which is related to $x_{it}$ by (12.6.2). Suppose that the $T \times 1$ measurement error vector $\boldsymbol{\tau}_i = (\tau_{i1}, \ldots, \tau_{iT})'$ is i.i.d. with mean zero and covariance matrix $\Omega = E(\boldsymbol{\tau}_i \boldsymbol{\tau}_i')$.

Suppose $\Omega$ has a structure of the form

$$\text{vec } \Omega = R_0 \boldsymbol{\lambda}, \tag{12.6.18}$$

where vec denotes the operation that stacks the rows of a matrix one after another in a column vector form, $R$ is a matrix of order $T^2 \times m$ with known elements, and $\boldsymbol{\lambda}$ is an $m \times 1$ vector of unknown constants. Using the covariance transformation matrix $Q = I_T - \frac{1}{T}\mathbf{e}_T \mathbf{e}_T'$ to eliminate the individual effects, $\alpha_i^*$, yields

$$Q\mathbf{y}_i = Q\mathbf{x}_i + QW_i\boldsymbol{\gamma} + Q\mathbf{u}_i, \tag{12.6.19}$$

$$Q\mathbf{x}_i^* = Q\mathbf{x}_i + Q\boldsymbol{\tau}_i, \tag{12.6.20}$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{iT})'$, $W_i = (\mathbf{w}_{it}')$. Let

$$R \equiv (I_T \otimes Q)R_0. \tag{12.6.21}$$

From (12.6.2), we have

$$\begin{aligned} E(\boldsymbol{\tau}_i \otimes Q\boldsymbol{\tau}_i) &= (I_T \otimes Q)E(\boldsymbol{\tau}_i \otimes \boldsymbol{\tau}_i) \\ &= (I_T \otimes Q)R_0\boldsymbol{\lambda} \\ &= R\boldsymbol{\lambda}. \end{aligned} \tag{12.6.22}$$

It follows that

$$\begin{aligned} E(\mathbf{x}_i^* \otimes Q\mathbf{x}_i) &= E(\mathbf{x}_i^* \otimes Q\mathbf{x}_i^*) - E[(\mathbf{x}_i + \boldsymbol{\tau}_i) \otimes Q\boldsymbol{\tau}_i] \\ &= E(\mathbf{x}_i^* \otimes Q\mathbf{x}_i^*) - R\boldsymbol{\lambda}. \end{aligned} \tag{12.6.23}$$

Therefore

$$E(\mathbf{x}_i^* \otimes Q\mathbf{y}_i) = E(\mathbf{x}_i^* \otimes Q\mathbf{x}_i^*)\beta + E(\mathbf{x}_i^* \otimes QW_i)\boldsymbol{\gamma} - \mathbf{R}\boldsymbol{\lambda}\boldsymbol{\beta}. \tag{12.6.24}$$

Equation (12.6.24) contains the nuisance parameter $\boldsymbol{\lambda}$. To eliminate $\boldsymbol{\lambda}$ from (12.6.24), multiplying $M_R = I_{T^2} - R(R'R)^{-1}R'$ to both sides of (12.6.24), we have the orthogonality conditions:

$$M_R E\{\mathbf{x}_i^* \otimes Q(\mathbf{y}_i - \mathbf{x}_i^*\beta - W_i\gamma)\} = \mathbf{0} \tag{12.6.25}$$

Combining (12.6.25) with the moment conditions $E(W_i'Q\mathbf{u}_i) = \mathbf{0}$, we have the moment conditions for the measurement error model (12.6.17)

$$E[M(\mathbf{d}_i - C_i\boldsymbol{\theta})] = \mathbf{0}, \tag{12.6.26}$$

where

$$M = \begin{bmatrix} M_R & \mathbf{0} \\ \mathbf{0} & I_K \end{bmatrix}, \mathbf{d}_i = \begin{bmatrix} \mathbf{x}_i^* \otimes I_T \\ W_i' \end{bmatrix} Q\mathbf{y}_i,$$

$$C_i = \begin{bmatrix} \mathbf{x}_i^* \otimes I_T \\ W_i' \end{bmatrix} Q[\mathbf{x}_i^*, W_i], \quad \boldsymbol{\theta}' = (\beta, \boldsymbol{\gamma}').$$

A GMM estimator is obtained by minimizing

$$\frac{1}{N}\left[\sum_{i=1}^{N} M(\mathbf{d}_i - C_i\boldsymbol{\theta})\right]' A_N \left[\sum_{i=1}^{N} M(\mathbf{d}_i - C_i\boldsymbol{\theta})\right]. \tag{12.6.27}$$

An optimal GMM estimator is to let

$$A_N^{-1} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{d}_i - C_i\hat{\boldsymbol{\theta}})(\mathbf{d}_i - \mathbf{C}_i\hat{\boldsymbol{\theta}})', \tag{12.6.28}$$

where $\hat{\boldsymbol{\theta}}$ is some consistent estimator of $\boldsymbol{\theta}$ such as

$$\hat{\boldsymbol{\theta}} = \left[\left(\sum_{i=1}^{N} C_i'\right) M \left(\sum_{i=1}^{N} C_i\right)\right]^{-1} \left[\left(\sum_{i=1}^{N} C_i\right)' M \left(\sum_{i=1}^{N} \mathbf{d}_i\right)\right]. \tag{12.6.29}$$

In the case when $\tau_{it}$ is i.i.d. across $i$ and over $t$, $\Omega$ is diagonal with equal diagonal element. Then $m = 1$ and $R_0 = \text{vec } I_T$, $R = (I_T \otimes Q) \text{ vec } I_T = \text{vec}\, Q$, $R'R = tr\, Q = T - 1$, and $M_R = I_{T^2} - \frac{1}{T-1}(\text{vec } Q)(\text{vec } Q)'$. When $\Omega$ is diagonal with distinct diagonal elements, $m = T$ and $R_0 = \mathbf{i}_t\mathbf{i}_t' \otimes \mathbf{i}_t$, where $\mathbf{i}_t$ is the $t$th unit vector of order $T$. When $\tau_{it}$ is a first-order moving average process and $T = 4$,

$$\Omega = \begin{bmatrix} a & c & 0 & 0 \\ c & b & c & 0 \\ 0 & c & b & c \\ 0 & 0 & c & a \end{bmatrix},$$

then

$$R_0 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \end{bmatrix},$$

and $\boldsymbol{\lambda} = (a, b, c)'$.

In general, the identifiability of the slope parameters $\boldsymbol{\beta}$ for a linear regression model depends on whether the moment equations involving observables in levels and differences for different order of lags are sufficient to obtain a unique solution for $\boldsymbol{\beta}$ given the assumption about the autocorrelation patterns of measurement errors. For additional references, see Biørn (2000), Biørn and Klette (1998), Biørn and Krishnakumar (2008), Wansbeek (2001), and Wansbeek and Meijer (2000, Chapter 6, Section 6.6).

The measurement errors for nonlinear models are much more difficult to handle (e.g., Hsiao 1992c). For binary choice models with measurement errors, see Kao and Schnell (1987a,b) and Hsiao (1991b).

## 12.7 NONPARAMETRIC PANEL DATA MODELS

Our discussion of panel data models have been confined to parametrically or semiparametrically specified models. The generalization to panel nonparametric models can be very complicated. However, in a static framework, the generalization to the nonparametric setup is fairly straightforward although the computation can be tedious. To see this, let

$$y_{it} = m(\mathbf{x}_{it}) + \nu_{it}, \quad \begin{matrix} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{matrix} \tag{12.7.1}$$

$$\nu_{it} = \alpha_i + u_{it}, \tag{12.7.2}$$

where $\mathbf{x}_{it}$ denotes the $K \times 1$ strictly exogenous variables with respect to $u_{it}$, $E(u_{it} \mid \mathbf{x}_{is}) = 0$ for all $t$ and $s$.

If $\alpha_i$ is treated as random and uncorrelated with $\mathbf{x}_{it}$, then $m(\mathbf{x}_{it})$ can be estimated by kernel method or the local linear least-squares method ($\min \sum_{i=1}^{N} \sum_{t=1}^{I} K(\frac{\mathbf{x}_{it}-\mathbf{x}^*}{\sigma_N})(y_{it} - m(\mathbf{x}_{it}))^2$) in which

$$m(\mathbf{x}_{it}) = m(\mathbf{x}^*) + (\mathbf{x}_{it} - \mathbf{x}^*)'\boldsymbol{\beta}(\mathbf{x}^*) \tag{12.7.3}$$

for $\mathbf{x}_{it}$ close to $\mathbf{x}^*$, where the "closeness" is defined in terms of some kernel function, $\sigma_N^{-K} K\left(\frac{\mathbf{x}_{it}-\mathbf{x}^*}{\sigma_N}\right)$, with $K(\mathbf{v}) \geq 0$, $K(\mathbf{v}) \longrightarrow 0$ as $\mathbf{v} \longrightarrow \pm\infty$ and $\sigma_N$ is a bandwidth parameter. Substituting (12.7.3) into (12.7.1), one can estimate $m(\mathbf{x}^*)$ and $\boldsymbol{\beta}(\mathbf{x}^*)$ by the least squares method (Li and Racine (2007, ch. 2)). However, the least squares method ignores the error components structure of $\nu_{it}$. Martins-Filho and Yao (2009), Su, Ullah, and Wang (2010), etc. have considered more efficient two-step estimators.

When $\alpha_i$ is treated as fixed constant, Kernel approach is not a convenient method to estimate $m(\mathbf{x}_{it})$ because linear difference of $y_{it}$ has to be used to eliminate $\alpha_i$. (e.g., Li and Stengos 1996). A convenient approach is to put $m(\mathbf{x}_{it})$ into the following general index format,

$$m(\mathbf{x}_{it}) = \nu_0(\mathbf{x}_{it}, \boldsymbol{\theta}^0) + \sum_{j=1}^{m} h_{j0}(\nu_j(\mathbf{x}_{it}, \boldsymbol{\theta}^0)), \tag{12.7.4}$$

where $\nu_j(\mathbf{x}_{it}, \boldsymbol{\theta}^0)$ for $j = 0, 1, \ldots, m$ are known functions of $\mathbf{x}_{it}$ and $h_{j0}(\cdot)$ for $j = 1, \ldots, m$ are unknown functions.

However, to uniquely identify the parameters of interest of the index model (12.7.4) one needs to impose the following normalization conditions:

(1) $h_{j0}(0) = 0$ for $j = 1, \ldots, m$.
(2) The scaling restriction, say $\boldsymbol{\theta}^{0'}\boldsymbol{\theta}^0 = 1$ or the first element of $\boldsymbol{\theta}^0$ be normalized to 1 if it is known different from 0.
(3) The exclusion restriction when $\nu_j(\mathbf{x}, \boldsymbol{\theta})$ and $\nu_s(\mathbf{x}, \boldsymbol{\theta})$ are homogeneous of degree 1 in the regressors for some $s \neq j$.

When (1) does not hold, it is not possible to distinguish $(h_{j0}(\cdot)), \alpha_i)$ from $h_{j0}(\cdot) - \mu, \alpha_i + \mu)$ for any constant $\mu$ and for any $j$ in (1). When (2) does not hold, it is not possible to distinguish $(\boldsymbol{\theta}^0, h_0(\cdot))$ from $(c\boldsymbol{\theta}^0, \tilde{h}_0(\cdot) = h_0(\cdot/c))$ for any nonzero constant $c$. When (3) does not hold, say $(h_{10}(\cdot), h_{20}(\cdot)$ containing a common element $\mathbf{x}_{3it}$, then $(h_{10}, h_{20})$ is not distinguishable from $(h_{10} + g(\mathbf{x}_{3it}), h_{20} - g(\mathbf{x}_{3it}))$ for any function $g$ (For further details, see Ai and Li (2008)).

A finite sample approximations for $h_j(\cdot)$ is to use series approximations

$$h_{j0}(\cdot) \simeq \mathbf{p}_j(\cdot)'\boldsymbol{\pi}_j \tag{12.7.5}$$

The simplest series base function is to use the power series. However, power series can be sensitive to outliers. Ai and Li (2008) suggest using the piecewise local polynomial spline as a base function in nonparametric series estimation. An $t$th order univariate $B$-spline base function is given by (see Chui (1992, Chapter 4).

$$B_r(x \mid t_0, \ldots, t_r) = \frac{1}{(r-1)!} \sum_{j=1}^{r} (-1)^j \binom{r}{j} [\max(0, x - t_j)]^{r-1}, \tag{12.7.6}$$

where $t_0, t_1, \ldots, t_r$ are the evenly spaced design knots on the support of $X$. When $r = 2$, (12.7.6) gives a piecewise linear spline, and when $r = 4$, it gives piecewise cubic splines (i.e., the third-order polynomials). Substituting the parametric specification (12.7.5) in lieu of $h_{j0}(\cdot)$ into (12.7.1), we obtain the parametric analog of (12.7.4). Then, just like in the parametric case, one can remove $\alpha_i$ by taking the deviation of $y_{it}$ from the $i$th individual time series mean $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$. Therefore one can obtain consistent estimators of $\boldsymbol{\theta}^0$ and $\boldsymbol{\pi}_j$, $j = 1, \ldots, m$ by minimizing

$$\begin{aligned} \sum_{i=1}^{N}\sum_{t=1}^{T}\{(y_{it} - \nu_0(\mathbf{x}_{it}, \boldsymbol{\theta}) - \sum_{j=1}^{m} \mathbf{p}_j(\nu_j(\mathbf{x}_{it}, \boldsymbol{\theta}))'\boldsymbol{\pi}_j) \\ - \frac{1}{T}\sum_{t=1}^{T}(y_{is} - \nu_0(\mathbf{x}_{is}, \boldsymbol{\theta}) - \sum_{j=1}^{m} \mathbf{p}_j(\nu_j(\mathbf{x}_{is}, \boldsymbol{\theta})'\boldsymbol{\pi}_j))\}^2 \end{aligned} \tag{12.7.7}$$

Shen (1997), Newey (1997), and Chen and Shen (1998) show that both $\hat{\boldsymbol{\theta}}$ and $\hat{h}_j(\cdot)$, $j = 1, \ldots, m$ are consistent and asymptotically normally distributed if $k_j \longrightarrow \infty$ while $\frac{k_j}{N} \longrightarrow 0$ (at certain rate) where $k_j$ denotes the dimension of $\boldsymbol{\pi}_j$.

The series approach can also be extended to the sample selection model (or partial linear model) discussed in Chapter 8,

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + m(\mathbf{z}_{it}) + \alpha_i + u_{it}, \tag{12.7.8}$$

where $y_{it}$ is observed if the dummy variable $d_{it} = 1$. The sample selection effect $m(\mathbf{z}_{it})$ given $d_{it} = 1$ can be approximated

$$m(\mathbf{z}_{it}) \sim \sum_{j=1}^{m} h_j(\mathbf{z}_{jit}), \tag{12.7.9}$$

where $h_j(\cdot)$ are unknown function. For identification purpose, $h_j(\cdot)$ is commonly assumed to satisfy the local restriction $h_j(0) = 0$ for all $j$ and the exclusive restriction that $\mathbf{z}_{1it}, \ldots, \mathbf{z}_{mit}$ are mutually exclusive. Then each $h_j(\cdot)$ can be approximated by the linear sieve $\mathbf{p}_j^{k_j}(\cdot)'\boldsymbol{\pi}_j$, where $\mathbf{p}_j^{k_j}(\cdot)$ is a vector of approximating functions satisfying $\mathbf{p}_j^{k_j}(0) = \mathbf{0}$. The unknown parameters $\boldsymbol{\beta}$ and the coefficients $\boldsymbol{\pi} = (\boldsymbol{\pi}_1', \ldots, \boldsymbol{\pi}_m')'$ can be estimated by the generalized least-squares estimator if $\alpha_i$ are treated as random and uncorrelated with $(\mathbf{x}_{it}, \mathbf{z}_{it})$, or by minimizing

$$\sum_{i=1}^{N} \sum_{s<t} \left[ (y_{it} - y_{is}) - (\mathbf{x}_{it} - \mathbf{x}_{is})'\boldsymbol{\beta} - \sum_{j=1}^{m} (\mathbf{p}_j^{k_j}(\mathbf{z}_{jit}) - \mathbf{p}_j^{k_j}(\mathbf{z}_{is}))'\boldsymbol{\pi}_j \right]^2 . \tag{12.7.10}$$

Ai and Li (2005) show that the resulting estimator is consistent and derive its asymptotic distribution.[1]

The nonparametric estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\pi}_j$ can be used to test the parametric specification of the model following the idea of Hong and White (1995). However, the strict exogeneity assumption of $\mathbf{x}_{it}$ excludes the inclusion of lagged dependent variables. Neither is this approach of replacing unknown $h_j(\cdot)$ by series expansion easily generalizable to censored or nonlinear panel data models [for further discussion, see Ai and Li (2008), and Su and Ullah (2011)].

[1] Ai and Li (2005) show that the nonlinear least-squares estimator of $\boldsymbol{\beta}$ is asymptotically normally distributed, but not $\boldsymbol{\pi}_j$.

CHAPTER 13

# A Summary View

## 13.1 BENEFITS OF PANEL DATA

As discussed in Chapter 1, panel data provides major benefits for econometric estimation in at least six areas: (1) increasing degrees of freedom and reducing problems of data multicollinarity, (2) constructing more realistic behavioral models and discriminating between competing economic hypotheses, (3) eliminating or reducing estimation bias, (4) obtaining more precise estimates of micro relations and generating more accurate micro predictions, (5) providing information on appropriate level of aggregation, and (6) simplifying cross sections or time series data inferential procedures. In this section we provide a summary view on how different methods discussed in this monograph can be used to achieve these benefits.

### 13.1.1 Increasing Degrees of Freedom and Lessening the Problem of Multicollinearity

In empirical studies investigators often encounter problems of shortage of degrees of freedom and multicollinearity. That is, the information provided by the sample is not rich enough to meet the requirement of the specified model. To narrow this gap, investigators either often have to impose ad hoc prior restrictions (e.g., Hsiao, Mountain, and Ho-Illman 1995) or to augment sample information. Panel data have many more degrees of freedom than cross-sectional or time series data. Moreover, panel data containing information on both interindividual differences across cross-sectional units and intraindividual dynamics over time can substantially increase the sample information. Pooling procedures to obtain more accurate estimation of common parameters for linear static and dynamic models are discussed in Chapters 3, 4, 9, 11 (Section 11.4), and 12 (Section 12.3); static and dynamic system of equations are in Chapters 5 and 10; nonlinear models are in Chapters 7, 8, and 12 (Sections 12.1 and 12.2). Pooling for heterogeneous individuals is discussed in Chapter 6.

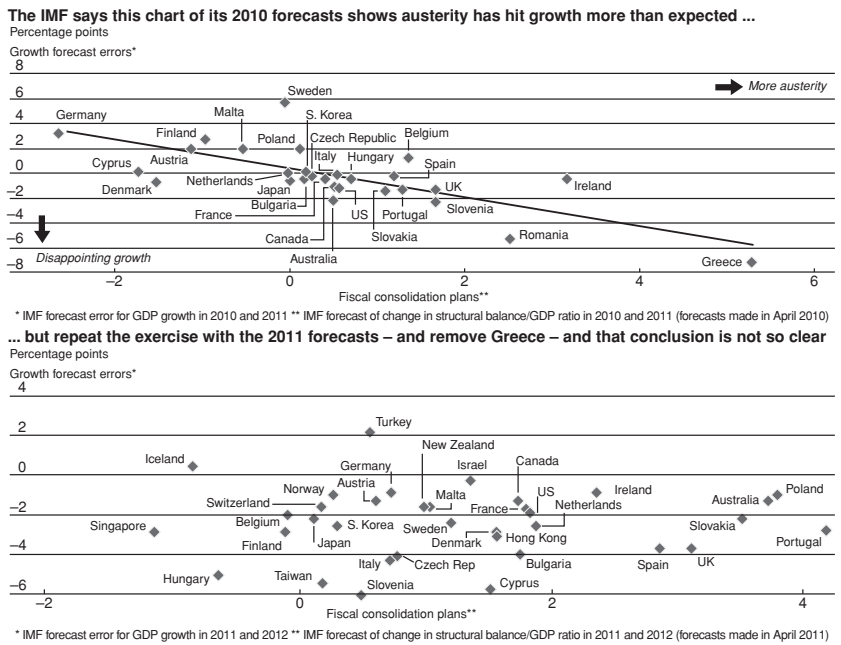


Figure 13.1. Robustness of IMF data scrutinized. * *Source:* The top figure is from IMF World Economic Outlook, Oct. 2012. The bottom figure is from Financial Times, Oct 13/14, 2012.

### 13.1.2 Identification and Discrimination between Competing Hypotheses

Aggregate time series data are not particularly useful for discriminating between hypotheses that depend on individual attributes. A single individual time series data set is not possible to provide information on the effects of different socio-demographic factors. Cross-sectional data, though containing information on microeconomic and demographic variables, cannot take account the (unobserved) heterogeneity across individuals. A fundamental assumption inherent in studies using cross-sectional data is that $E(y_i \mid \mathbf{x} = \mathbf{x}^*) = E(y_j \mid \mathbf{x} = \mathbf{x}^*)$. This homogeneity (conditional on $\mathbf{x}$) assumption can lead to grossly misleading or sensitive inference on the impact of $\mathbf{x}$ on $y$. For instance, with many economics in fiscal consolidation model since financial crises broke out in 2008, a debate has been raging about the size of fiscal multipliers. The smaller the multipliers, the less costly the fiscal consolidation. Under rational expectations and if the correct forecast model has been used, there should be no relation between the forecast error for real GDP growth and planned fiscal consolidation (measured as changes in the structural fiscal balance (due to tax rises and spending cuts) as a percentage of potential GDP). The top figure of Figure 13.1 shows the International Monetary Fund (IMF) estimate of the effect of austerity plans

on the 2010–11 forecast error of the real GDP growth rate (World Economic Outlook IMF, October 9, 2012). The figures shows a large, negative relation. The baseline estimate suggests that a planned consolidation of 1 percent of GDP is associated with a growth forecast error of about 1 percentage point. Based on this analysis, IMF concludes that the assumed multipliers of about 0.5 underlying the forecast models have been too low by about 1. The short-term fiscal multipliers should be in the range of 0.9 to 1.7. The bottom figure shows the Financial Times (October 13/14, 2012) estimates after removing Greece and Germany. The relationship between deficit reduction efforts and forecast error of the growth rate is simply not there. One reason that these results are so sensitive to the includion and exclusion of certain countries is because the cross-sectional analysis cannot take account the effects of (unobserved) country-specific factors.

In economics, as in other branches of the social and behavioral sciences, often there are competing theories. Examples of these include the effect of collective bargaining on wages, the appropriate short-term policy to alleviate unemployment (Chapters 1 and 7), the effects of schooling on earnings (Chapter 5), and the question of causal ordering such as "Does delinquency lead to low self-esteem or does low self-esteem lead to delinquency?" (e.g., Jang and Thornberry 1998). Economists on opposite sides of these issues generally have very different views on the operation of the economy and the influence of institutions on economic performance. Some economists believe unions indeed raise wages or that advertising truly generates greater sales. Adherents of the opposite view tend to regard the effects more as epiphenomena than as substantive forces and believe that observed differences are due mainly to sorting of workers or firms by characteristics (e.g., Allison 2000).

Proper recognition of the sources of variation can provide very useful information for discriminating individual behavior from average behavior or for identifying an otherwise unidentified model. For instance, in the foregoing collective bargaining example, even if information on worker quality is not available, if a worker's ability stays constant or changes only slowly, the within correlation between the union-status dummy and the worker-quality variable is likely to be negligible. Thus, the impact of worker quality can be controlled through the use of within estimates (Chapter 3). The resulting coefficient for the union-status dummy then will provide a measure of the effect of unionism. In the income schooling model, the availability of family groupings can provide an additional set of cross-sibling covariances via a set of common omitted variables. These additional restrictions can be combined with the conventional slope restrictions to identify what would otherwise be unidentified structure parameters (Chapter 5, Section 5.4).

Panel data providing sequential observations for a number of individuals allow an investigator to distinguish interindividual differences from intraindividual differences and construct a proper causal structure (Chapters 5, 9 [Sections 9.3, 9.4], and 10). Furthermore, addition of the cross-sectional dimension to the time series dimension provides a distinct possibility to identify the pattern of serial correlations in the residuals or to identify the lag adjustment

patterns when the conditioning variables are changed without having to resort to imposing prior parametric restrictions (Chapters 3 [Section 3.8] and 11 [Section 11.4]) or to identify a model subject to measurement errors (Chapter 12, Section 12.6).

### 13.1.3 Reducing Estimation Bias

A fundamental statistical problem facing every econometrician is the *specification problem*. By that we mean the selection of variables to be included in a behavioral relationship as well as the manner in which these variables are related to the variables that affect the outcome but appear in the equation only through the error term. Empirical findings are often criticized on the grounds that the researcher has not explicitly recognized the effects of omitted variables that are correlated with the included explanatory variables (in the union example, the omitted variable, worker quality, can be correlated with the included variable, union status). If the effects of the omitted variables are correlated with the included explanatory variables, and if these correlations are not explicitly allowed for, the resulting regression estimates could be seriously biased (Chapter 3, Sections 3.4 and 3.5). To minimize the bias, it is helpful to distinguish four types of correlations between the included variables and the error term. The first type is due to the correlation between the included exogenous variables and those variables that should be included in the equation but are not, either because of a specification error or because of unavailability of data (Chapters 3, 7, and 8). The second type is due to the dynamic structure of the model and the persistence of the shocks that give rise to the correlation between lagged dependent variables and the error term (Chapters 4 and 10). The third type is due to the simultaneity of the model, which gives rise to the correlation between the jointly dependent variables and the error terms (Chapters 5 and 10 [Section 10.4]). The fourth type is due to measurement errors in the explanatory variables (Chapter 12, Section 12.6). Knowing the different sources of correlations provides important information for devising consistent estimators. It also helps one avoid the possibility of eliminating one source of bias while aggravating another (e.g., Chapter 5, Section 5.1).

Panel data can help identify these four sources of correlations. For instance, if the effects of these omitted variables stay constant for a given individual through time or are the same for all individuals in a given time period and the model is linear (e.g., Chapters 3 and 4), the omitted-variable bias can be eliminated by one of the following three methods when panel data are available: (1) differencing the sample observations to eliminate the individual-specific and/or time-specific effects, (2) using dummy variables to capture the effects of individual invariant and/or time-invariant variables; and (3) postulating a conditional distribution of unobserved effects given observed exogenous variables, then integrating out the unobserved effects to make inferences based on the marginal distribution of observables. The first two approaches are commonly referred as the *fixed-effects inference* and the third approach is referred as the *random-effects inference*.

Panel data can also help to identify the correlations between the regressors and errors that are due to simultaneity or to the correlations between the unobserved individual- or time-specific effects and the regressors. The standard approach to eliminate simultaneity bias is to use instrumental variables to purge the correlations between the joint dependent variables and the error of the equation. However, if there exist correlations between the regressors and the unobserved individual- or time-specific effects, what are generally considered as valid instruments may not be valid any more (Chapter 5, Sections 5.3 and 5.4.)

Measurement errors in the explanatory variables create correlations between the regressors and the errors of the equation. If variables are subject to measurement errors, the common practice of differencing out individual effects eliminates one source of bias but may aggravate the bias due to measurement errors. However, different transformation of the data can induce different and deducible changes in the estimated regression parameters, which can be used to determine the importance of measurement errors and obtain consistent estimators of parameters of interest (Chapter 12, Section 12.6).

### 13.1.4 Generating More Accurate Predictions for Individual Outcomes

If individual behaviors are similar conditional on certain variables, panel data provide the possibility of learning an individual's behavior by observing the behavior of others. Thus, it is possible to obtain a more accurate description of an individual's behavior by supplementing observations of the individual in question with data on other individuals (e.g., Chapter 6).

### 13.1.5 Providing Information on Appropriate Level of Aggregation

A model is a simplification of reality, not a slavish reproduction of all real-world data. The real-world detail is reduced through aggregation of "homogeneous" units or through the "representative agent" assumption. However, if micro units are heterogeneous, not only can the time series properties of aggregate data be very different from those of disaggregate data (e.g., Granger 1980; Lewbel 1992, 1994; Stoker 1993), policy evaluation based on aggregate data can be grossly misleading. Furthermore, the prediction of aggregate outcomes using aggregate data can be less accurate than the prediction based on micro-equations (e.g., Chapter 6, Section 6.8.2 or Hsiao, Shen, and Fujiki 2005). Panel data containing time series observations for a number of individuals is ideal for investigating the "homogeneity" versus "heterogeneity" issue. Moreover, when "homogeneity" in panel is rejected, the variable coefficient models discussed in Chapter 6 provides a feasible alternative to make inferences about the population while taking account of the heterogeneity among micro units.

### 13.1.6 Simplifying Computation and Statistical Inference

Panel data involve at least two dimensions, a cross-sectional dimension and a time series dimension. Under normal circumstances the computation of panel data estimator or inference would be more complicated than cross-sectional or time series data. However, on many occasions, the availability of panel data actually simplifies computation and inference. For instance, in the analysis of time series properties of a variable, first one will need large number of time series observations to properly distinguish stationary time series from nonstationary time series. Second, when time series data are not stationary, the large sample approximation of the distributions of the least-squares or maximum likelihood estimators are no longer normally distributed (e.g., Anderson 1959; Dickey and Fuller 1979, 1981; Phillips and Durlauf 1986). But if panel data are available, one can invoke the central limit theorem across cross-sectional units to show that the limiting distributions of many estimators remain asymptotically normally distributed. Moreover, even only a small number of time series observations are available, an investigator making use of information on cross-sectional dimension may be able to distinguish unit roots or cointegration processes from stationary process (Chapter 10).

Another example is in the evaluation of the impact of social program. When only cross-sectional data are available, the control of the impact of *selection on observables or unobservables* could be complicated (Chapter 9, Section 9.6.2). However, if panel data are available and if individual units are cross-sectionally dependent, then one can use cross-sectional units information to construct the counterfactuals for the evaluation of the impact of social program without the need to worry about the issues of selection on observables or unobservables which may considerably simplify the analysis (Chapter 9, Sections 9.6.3 and 9.6.4).

## 13.2 CHALLENGES FOR PANEL DATA ANALYSIS

Although panel data offer many advantages over a cross-sectional or time series data set, there are many interesting and unresolved issues remain such as (1) how best to model unobserved heterogeneity across individuals and/or over time; (2) controlling the impact of unobserved heterogeneity to obtain valid inference for nonlinear models; (3) modeling cross-sectional dependence; (4) multidimensional asymptotics; and (5) sample attrition, etc.

### 13.2.1 Modeling Unobserved Heterogeneity

As discussed in the introduction (Chapter 1, Section 1.3), panel data focus on individual outcomes over time. Factors affecting individual outcomes could be numerous. One of the most challenging issues in panel data modeling is how to model the unobserved heterogeneity across individuals and over time that are not captured by the conditional variables $\mathbf{x}$. This monograph essentially follows

the approach of letting part of the parameters characterizing the conditional distribution of $y_{it}$ given $\mathbf{x}_{it}$ to vary across $i$ and over $t$, $f(y_{it} \mid \mathbf{x}_{it}, \boldsymbol{\beta}, \boldsymbol{\gamma}_{it})$, where $\boldsymbol{\beta}$ is assumed identical over $i$ and $t$ and $\boldsymbol{\gamma}_{it}$ vary across $i$ and over $t$. To control the impact of $\boldsymbol{\gamma}_{it}$ on the inference of $\boldsymbol{\beta}$, $\boldsymbol{\gamma}_{it}$ is further decomposed into components that are individual-specific, $\boldsymbol{\alpha}_i$, and component that are time-specific, $\boldsymbol{\lambda}_t$. Is this the best way to model unobserved heterogeneity? When time series dimension or cross-sectional dimension becomes large, is the assumption of time-invariance in $\boldsymbol{\alpha}_i$ or individual-invariance of $\boldsymbol{\lambda}_t$ still reasonable? Further, the function of a variable could have very different meanings at different time. For instance, Friedman (1969) found that there was a stable relation between the $M2$ and nominal GDP in the 1960s. However, with technological development there are many financial instruments today that can also perform the function of currency and demand deposits. Do today's M2 still have the same economic implication as M2 in the 1960s or should these close substitutes be also included in the analysis of money?

There is also an issue of whether to treat the unobserved heterogeneity as fixed and different (fixed effects) or as random draws from a common (conditional) distribution (random effects). In general, if the unobserved heterogeneity can be viewed as random draws from a common population, then it is more appropriate to postulate a random-effects model. If the unobserved heterogeneity is correlated with explanatory variables or comes from heterogeneous population, then it is more appropriate to postulate a fixed-effects model unless the interaction between observables and unobservables are known to investigators. The fixed-effects formulation makes inference conditional on the specific effects; hence it has the advantage of not requiring one to postulate the distribution of the effects. However, there is also a loss of efficiency in conditional inference because of the loss of degrees of freedom in estimating the specific effects. It may even introduce incidental parameters problem if the dimension of the effects increase at the same rate as sample size (Chapters 4, 7, and 8). The advantages of a random effects specification is that the probability function of the effects in general depends only on a finite number of parameters and there is no incidental-parameter problem, and efficient inference is possible. The disadvantage is that it requires explicit knowledge about the way in which observables and unobservables interact. In general, the advantages of fixed effects formulation are the disadvantages of random effects formulation and the disadvantages of the fixed effects formulation are the advantages of the random effects formulations (e.g., see the discussions in Chapters 3 and 4). Unfortunately, without explicit knowledge about the way in which observables and unobservables interact it is hard to decide which approach to adopt.

### 13.2.2 Controlling the Impact of Unobserved Heterogeneity in Nonlinear Models

There is a very fundamental difference between the linear and nonlinear models. If a model is linear, one can condition the effects on the observables and apply a minimum distance type estimator. If the model is nonlinear, the assumptions

for the conditional distribution of the effects need to be very specific. However, the effects are unobservable. It is hard to specify the conditional distribution of the effects without explicit assumptions about how the observables and unobservables interact. Moreover, the derivation of random-effect estimator often would involve multidimensional integration which can be very complicated even with today's computing capacity.

If the effects are treated as fixed, and if the number of unknown specific effects increases at the same rate as the sample size, attempts to estimate the specific effects creates the incidental-parameter problem. For general nonlinear models, there does not exist a generally applicable framework to implement the Neyman–Scott (1948) principle of separating the estimation of the common coefficients from the estimation of the specific effects. To devise consistent estimators of the structural parameters, one has to exploit the specific structure of a nonlinear model. The three most commonly used approaches are: (1) the conditional approach that conditions on the minimum sufficient statistics of the effects, (2) the semiparametric approach that exploits the latent linear structure of a model (Chapters 7 and 8), and (3) reparameterization of the model so that the information matrix of the reparameterized individual effects are uncorrelated with the reparameterized structural parameters (Lancaster 2001). The first two approaches apply classical sampling inference to a model that no longer involve incidental parameters. The transformation of a model containing incidental parameters to a model without incidental parameters is obtained through exploiting the specific structure of the original model. The third approach is from a Bayesian perspective. It can be shown that when the information matrix of the structural (or common) parameters are orthogonal to the incidental (or individual-specific) parameters, taking a uniform prior for the incidental parameter reduces the bias (Arellano and Bonhomme 2009). However, for most nonlinear models there does not appear that simple transformations to achieve information orthogonality exist. Whether any of these approaches will yield consistent estimators has to be considered case by case. Moreover, even in the case that consistent estimators exist, the conditions imposed on the data are so restrictive that hardly any data set can meet them (e.g., Chapter 7, Section 7.5).[1]

### 13.2.3 Modeling Cross-Sectional Dependence

If panel data are not conditional independent across cross-sectional units, ignoring cross-sectional dependence can lead to misleading inference. Contrary to time series observations there is no natural ordering of cross-sectional units. Chapter 9 surveyed some of the popular approaches that have been tried econometrically. Each approach has its merits and also limitations. In particular, in

[1] For instance, in the dynamic logit model considered in Chapter 7, Section 7.5, the conditions for the existence of consistent estimator requires at least (1) four times series observations for each individual; (2) indivdiuals switch position during the two intermediate periods; and (3) the value of the exogenous variable has to be equal in period 3 and period 4.

the case when $N$ is large and $T$ is small or the model is nonlinear, methods to take account cross-sectional dependence remain to be developed.

### 13.2.4 Multidimensional Asymptotics

This monograph focuses on panels that contain a cross-sectional dimension ($N$) and a time-series dimension ($T$). The majority of the discussions are on the case that there are a few observations in one dimension (usually the time dimension) and a great many observations in another dimension (usually the cross-sectional dimension), but there are panels where $N$ and $T$ are of similar magnitude. It is important to understand the properties of inferential procedures when a panel with only one dimension observations that are large or a panel that both or multidimensional observations are large, say both $N$ and $T \to \infty$, and the relative speed of their increase. On the basis of this information, one can then determine which parameters can, and which parameters cannot, be consistently estimated from a given panel or where the asymptotic bias comes from. For instance, in a linear dynamic model with the error composed of the sum of two components, one being individually time-invariant and the other being independently distributed, then the individual time-invariant effects can be eliminated by differencing successive observations of an individual. We can then use lagged dependent variable (of sufficiently high order) as instruments for the transformed model to circumvent the issues of the serial dependence of the residual (Chapter 4, Sections 4.3 and 4.5). When $T$ is fixed and $N$ is large, the resulting estimator is consistent and asymptotically normally distributed. However, when $T$ increases with $N$ and $\frac{T}{N} \to c \neq 0$ as $N \to \infty$, although the resulting estimator is consistent, there is an asymptotic bias term when the estimator is multiplied by the scale factor, $\sqrt{NT}$, that needs to be corrected to obtain asymptotic valid inference (e.g., Chapters 4 and 10, Appendix 4B or Alvarez and Arellano 2003; Phillips and Moon 1999).

Computing speed and storage capability have enabled researchers to collect, store and analyze data sets of very high dimensions. Multidimensional panel will become more available. Classical asymptotic theorems under the assumption that the dimension of data is fixed (e.g., Anderson (1985)) appear to be inadequate to analyze issues arising from finite sample of very high dimensional data (e.g., Bai and Silverstein 2004). For example, Bai and Saranadasa (1996) proved that when testing the difference of means of two high-dimensional populations, Dempsters (1959) nonexact test is more powerful than Hotellings (1931) $T^2$-test even though the latter is well defined. Many interesting and important issues remain to be worked out. Statistic theorems providing insight to finite sample issues for high dimensional data analysis can be very useful to economists and/or social scientists (e.g., Bai and Silverstein 2006).

### 13.2.5 Sample Attrition

Panel data follows a number of individuals over time. As Table 1.1 shows, as time goes on, a number of individuals drop out. If sample attrition is random,

it does not pose serious issues on panel data model, as one can simply focus on the remaining samples that have complete history. If a test (Chapter 11, Section 11.1) indicates that sample attrition is behaviorably related, ignoring the attrition issues could result in misleading inference. Baltagi and Song (2006), and Hirano et al. (2001) show the potential of using refreshment samples to distinguish between various forms of attrition. However, to properly take account of sample attrition, one will have to have explicit knowledge of why individuals drop out. Moreover, as Hausman and Wise (1977) (see Chapter 8, Section 8.2) or Ridder (1990) illustrates, computationally it could be a formidable task to take into account the sample attrition issue.

## 13.3 A CONCLUDING REMARK

This monograph hopes to provide an overview of the many statistical tools developed to analyze panel data and demonstrate the many advantages panel data may possess. In choosing the proper method to exploit the richness and unique property of panel data, it is helpful to keep several factors in mind. First, what advantages do panel data offer us in adapting economic theory for empirical investigation over data sets consisting of a single cross section or time series? Second, what are the limitations of panel data and the econometric methods that have been proposed for analyzing such data? Third, the usefulness of panel data in providing particular answers to certain issues depends critically on the compatibility between the assumptions underlying the statistical inference procedures and the data-generating process. Fourth, when using panel data, how can we increase the efficiency of parameter estimates? "Analyzing economic data requires skills of synthesis, interpretation and empirical imagination. Command of statistical methods is only a part, and sometimes a very small part, of what is required to do a first-class empirical research" (Heckman 2001). Panel data are no panacea. Nevertheless, if "panel data are only a little window that opens upon a great world, they are nevertheless the best window in econometrics" (Mairesse 2007).

# References


Abramowitz, M., and J. Stegun. (1965). *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*. New York: Dover.

Abrevaya, J. (1999). "Leapfrog Estimation of a Fixed-Effects Model with Unknown Transformation of the Dependent Variable." *Journal of Econometrics*, 93, 203–228.

_____. (2000). "Rank Estimation of a Generalized Fixed-Effects Regression Model." *Journal of Econometrics*, 95, 1–24.

Ahn, H., and J.L. Powell. (1993). "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism." *Journal of Econometrics*, 58, 3–30.

Ahn, S.C., and H.R. Moon. (2001). "On Large-$N$ and Large-$T$ Properties of Panel Data Estimators and the Hausman Test." Mimeo, Arizon State University.

Ahn, S.C., and P. Schmidt. (1995). "Efficient Estimation of Models for Dynamic Panel Data." *Journal of Econometrics*, 68, 5–27.

Ahn, S.C., Y.H. Lee and P. Schmidt. (2001). "GMM Estimation of Linear Panel Data Models with Time-Varying Individual Effects." *Journal of Econometrics*, 101, 219–55.

_____. (2013). "Panel Data Models with Multiple Time-Varying Individual Effects." *Journal of Econometrics*, 174, 1–14.

Ai, C., and Q. Li. (2005). "Estimation of Partly Specified Panel Data Tobit Models." Mimeo, Texas A&M University.

Ai, C., and Q. Li. (2008). "Semi-Parametric and Non-Parametric Models in Panel Data Models," in *The Econometrics of Panel Data*, 3rd ed., edited by L. Mátyás and P. Sevestre. Berlin: Springer-Verlag, pp. 451–78.

Aigner, D.J., and P. Balestra. (1988). "Optimal Experimental Design for Error Components Models." *Econometrica*, 56, 955–72.

Aigner, D.J., C. Hsiao, A. Kapteyn, and T. Wansbeek. (1984). "Latent Variable Models in Econometrics," in *Handbook of Econometrics*, Vol. II, edited by Z. Griliches and M. Intriligator, pp. 1322–93. Amsterdam: North-Holland.

Akaike, H. (1973). "Information Theory and an Extension of the Maximum Likelihood Principle," in *Proceedings of the 2nd International Symposium Information Theory*, edited by B.N. Petrov and F. Csaki, pp. 267–281. Budapest: Akademiai Kiado.

Akashi, K., and N. Kunitomo. (2011). "The Limited Information Maximum Likelihood Approach to Dynamic Structural Equation Models." Mimeo, University of Tokyo.

_____. (2012). "Some Properties of the LIML Estimator in a Dynamic Panel Structural Equation." *Journal of Econometrics*, 166, 167–83.

Allison, P. (2000). "Inferring Causal Order from Panel Data." Paper presented at the 9th International Conference on Panel Data, Geneva, Switzerland.

Alvarez, J., and M. Arellano. (2003). "The Time Series and Cross-Section Asymptotics of Dynamic Panel Data Estimators." *Econometrica*, 71, 1121–59.

Amemiya, T. (1971). "The Estimation of the Variance in a Variance-Component Model." *International Economic Review*, 12, 1–13.

_____. (1974). "Bivariate Probit Analysis: Minimum Chi-Square Methods." *Journal of the American Statistical Association*, 69, 940–44.

_____. (1976). "The Maximum Likelihood, the Minimum Chi-Square and the Nonlinear Weighted Least Squares Estimator in the General Qualitative Response Model." *Journal of the American Statistical Association*, 71, 347–51.

_____. (1978a). "The Estimation of a Simultaneous Equation Generalized Probit Model." *Econometrica*, 46, 1193–205.

_____. (1978b). "A Note on a Random Coefficients Model." *International Economic Review*, 19, 793–96.

_____. (1980a). "Selection of Regressors." *International Economic Review*, 21, 331–54.

_____. (1980b). "The $n^{-2}$-Order Mean Squared Errors of the Maximum Likelihood and the Minimum Logit Chi-Square Estimator." *Annals of Statistics*, 8, 488–505.

_____. (1981). "Qualitative Response Models: A Survey." *Journal of Economic Literature*, 19, 1483–536.

_____. (1983). "Nonlinear Regression Models," in *Handbook of Econometrics*, Vol. I, edited by Z. Griliches and M. Intriligator, pp. 333–89. Amsterdam: North-Holland.

_____. (1984). "Tobit Models: A Survey." *Journal of Econometrics*, 24, 3–62.

_____. (1985). *Advanced Theory of Econometrics*. Cambridge, MA: Harvard University Press.

Amemiya, T., and W.A. Fuller. (1967). "A Comparative Study of Alternative Estimators in a Distributed-Lag Model." *Econometrica*, 35, 509–29.

Amemiya, T., and T.E. MaCurdy. (1986). "Instrumental Variable Estimation of An Error Components Model." *Econometrica*, 54, 869–80.

Andersen, E.B. (1970). "Asymptotic Properties of Conditional Maximum Likelihood Estimators." *Journal of the Royal Statistical Society B*, 32, 283–301.

_____. (1973). *Conditional Inference and Models for Measuring*. København: Mentalhygiejnish Farlag.

Anderson, T.W. (1959). "On Asymptotic Distributions of Estimates of Parameters of Stochastic Differences Equations." *Annals of Mathematical Statistics*, 30, 676–87.

_____. (1969). "Statistical Inference for Covariance Matrices with Linear Structure," in *Multivariate Analysis*, Vol. 2, edited by P. R. Krishnaiah, pp. 55–66. New York: Academic Press.

_____. (1970). "Estimation of Covariance Matrices Which Are Linear Combinations or Whose Inverses Are Linear Combinations of Given Matrices," in *Essays in Probability and Statistics*, edited by R. C. Bose, pp. 1–24. Chapel Hill: University of North Carolina Press.

_____. (1971). *The Statistical Analysis of Time Series*. New York: John Wiley & Sons.

_____. (1978). "Repeated Measurements on Autoregressive Processes." *Journal of the American Statistical Association*, 73, 371–78.

———. (1985). *An Introduction to Multivariate Analysis*, 2nd ed. New York: John Wiley & Sons.

Anderson, T.W., and C. Hsiao. (1981). "Estimation of Dynamic Models with Error Components," *Journal of the American Statistical Association*, 76, 598–606.

———. (1982). "Formulation and Estimation of Dynamic Models Using Panel Data." *Journal of Econometrics*, 18, 47–82.

Angrist, J.D., and J. Hahn. (1999). "When to Control for Covariates? Panel-Asymptotic Results for Estimates of Treatment Effects." NBER Technical Working Paper 241.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic.

Anselin, L., and D.A. Griffith. (1988). "Do Spatial Effects Really Matter in Regression Analysis?" *Papers of the Regional Science Association*, 65, 11–34.

Anselin, L., J. Le Gallo, and H. Jayet. (2008). "Spatial Panel Econometrics," in *The Econometrics of Panel Data*, 3rd ed., edited by L. Mátyás and P. Sevestre, pp. 625–60. Berlin: Springer-Verlag.

Antweiler, W. (2001). "Nested Random Effects Estimation in Unbalanced Panel Data." *Journal of Econometrics*, 101, 295–313.

Arellano, M. (2003). *Panel Data Econometrics*. Oxford: Oxford University Press.

Arellano, M., and S. Bond. (1991). "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *Review of Economic Studies*, 58, 277–97.

Arellano, M., and S. Bonhomme. (2009). "Robust Priors in Nonlinear Panel Data Models." *Econometrica*, 77, 489–536.

Arellano, M., and O. Bover. (1995). "Another Look at the Instrumental Variable Estimation of Error-Components Models." *Journal of Econometrics*, 68, 29–51.

Arellano, M., and R. Carrasco. (2003). "Binary Choice Panel Data Models with Predetermined Variables." *Journal of Econometrics*, 357–81.

Arellano, M., and B. Honoré. (2001). "Panel Models: Some Recent Development," in *Handbook of Econometrics*, Vol. 5, edited by J. Heckman and E. Leamer, pp. 3229–296. Amsterdam: North-Holland.

Arellano, M., O. Bover, and J. Labeaga. (1999). "Autoregressive Models with Sample Selectivity for Panel Data," in *Analysis of Panels and Limited Dependent Variable Models*, edited by C. Hsiao, K. Lahiri, L.F. Lee, and M.H. Pesaran, pp. 23–48. Cambridge: Cambridge University Press.

Ashenfelter, O. (1978). "Estimating the Effect of Training Programs on Earnings." *Review of Economics and Statistics*, 60, 47–57.

Ashenfelter, O., and G. Solon. (1982). "Longitudinal Labor Market Data-Sources, Uses and Limitations," in *What's Happening to American Labor Force and Productivity Measurements?* pp. 109–26. Proceedings of a June 17, 1982, conference sponsored by the National Council on Employment Policy, W. E. Upjohn Institute for Employment Research.

Ashenfelter, O., A. Deaton, and G. Solon. (1984). "Does It Make Sense to Collect Panel Data in Developing Countries?" Mimeo, World Bank.

Avery, R.B. (1977). "Error Components and Seemingly Unrelated Regressions." *Econometrica*, 45, 199–209.

Bai, J. (2003). "Inferential Theory for Factor Models of Large Dimensions." *Econometrica*, 71, 135–71.

———. (2009). "Panel Data Models with Interactive Fixed Effects." *Econometrica*, 77, 1229–79.

Bai, J., and J.L. Carrion-i-Silvestre. (2009). "Structural Changes, Common Stochastic Trends, and Unit Roots in Panel Data." *Review of Economic Studies*, 76, 471–501.

Bai, J., and S. Ng. (2002). "Determining the Number of Factors in Approximate Factor Models." *Econometrica*, 70, 191–221.

_____. (2004). "A Panic on Unit Root Tests and Cointegration." *Econometrica*, 72, 1127–77.

_____. (2010). "Panel Unit Root Tests with Cross-Section Dependence: A Further Investigation." *Econometric Theory*, 26, 1088–114.

Bai, Z.D., and H. Saranadasa. (1996), "Effect of High Dimension: by an Example of a Two Sample Problem." *Statistical Sinica*, 6, 311–29.

Bai, Z.D., and J.W. Silverstein. (2004), "CLT of Linear Spectral Statistics of Large-Dimensional Sample Covariance Matrices." *Annals of Probability*, 32(1A), 553–605.

_____. (2006). *Spectral Analysis of Large Dimensional Random Matrices*. Beijing: Science Press.

Baillie, R.T., and B.H. Baltagi. (1999). "Prediction from the Regression Model with One-Way Error Components," in *Analysis of Panels and Limited Dependent Variable Models*, edited by C. Hsiao, K. Lahiri, L.F. Lee, and M.H. Pesaran, pp. 255–67. Cambridge: Cambridge University Press.

Balestra, P., and M. Nerlove. (1966). "Pooling Cross-Section and Time Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas." *Econometrica*, 34, 585–612.

Baltagi, B.H. (1980). "On Seemingly Unrelated Regressions with Error Components." *Econometrica*, 48, 1547–51.

_____. (1981a). "Simultaneous Equations with Error Components." *Journal of Econometrics*, 17, 189–200.

_____. (1981b). "Pooling: An Experimental Study of Alternative Testing and Estimation Procedures in a Two-Way Error Components Model." Mimeo, University of Houston.

_____. (1995). "Econometric Analysis of Panel Data." New York: John Wiley & Sons.

Baltagi, B.H., and J.M. Griffin. (1983). "Gasoline Demand in the OECD: An Application of Pooling and Testing Procedures." *European Economic Review*, 22, 117–37.

Baltagi, B.H., and C. Kao. (2000). "Nonstationary Panels, Cointegration in Panels and Dynamic Panels, A Survey in Nonstationary Panels, Panel Cointegration and Dynamic Panels," *Advances in Econometrics*, Vol. 15, edited by B. Baltagi, pp. 7–52. Amsterdam: JAI Press.

Baltagi, B.H., and Q. Li. (1991). "A Transformation that will Circumvent the Problem of Autocorrelation in an Error Component Model." *Journal of Econometrics*, 48, 385–93.

_____. (1992). "A Monotonic Property for Iterative GLS in the Two-Way Random Effects Model." *Journal of Econometrics*, 53, 45–51.

Baltagi, B. and S.H. Song. (2006). "Unbalanced Panel Data: A Survey." *Statistical Papers*, 47, 493–523.

Baltagi, B.H., S. Song, and B. Jung. (2001). "The Unbalanced Nested Error Component Regression Model." *Journal of Econometrics*, 101, 357–81.

Baltagi, B.H., S. Song, B. Jung, and W. Koh. (2007). "Testing for Serial Correlation, Spatial Autocorrelation and Random Effects Using Panel Data." *Journal of Econometrics*, 140, 5–51.

Banerjee, A. (1999). "Panel Data Unit Roots and Cointegration: An Overview." *Oxford Bulletin of Economics and Statistics*, 61, 607–29.

Banerjee, A., M. Marcellino, and C. Osbat. (2005). "Testing for PPP: Should We Use Panel Methods?" *Empirical Economics*, 30, 77–91.

Barro, R., and X. Sala-i-Martin. (1995). *Economic Growth*. New York: McGraw-Hill.

Barth, J., A. Kraft, and J. Kraft. (1979). "A Temporal Cross-Section Approach to the Price Equation." *Journal of Econometrics*, 11, 335–51.

Bartolucci, F., and V. Nigro. (2010). "A Dynamic Model for Binary Panel Data with Unobserved Heterogeneity Admitting a Root-N Consistent Conditional Estimator." *Econometrica*, 78, 719–33.

_____. (2012). "Pseudo Conditional maximum Likelihood Estimation of the Dynamic Logit Model for Binary Panel Data." *Journal of Econometrics*, 170, 102–16.

Bassett, G. Jr., and R. Koenker. (1978), "Asymptotic Theory of Least Absolute Error Regression." *Journal of the American Statistical Association*, 73, 618–22.

Bates, G., and J. Neyman. (1951). "Contributions to the Theory of Accident Proneness. II: True of False Contagion." *University of California Publications in Statistics*, pp. 215–53.

Becketti, S., W. Gould, L. Lillard, and F. Welch. (1988). "The Panel Study of Income Dynamics After Fourteen Years: An Evaluation." *Journal of Labor Economics*, 6, 472–92.

Beckwith, N. (1972). "Multivariate Analysis of Sales Response of Competing Brands to Advertising." *Journal of Marketing Research*, 9, 168–76.

Ben-Porath, Y. (1973). "Labor Force Participation Rates and the Supply of Labor." *Journal of Political Economy*, 81, 697–704.

Berkson, J. (1944). "Application of the Logistic Function to Bio-Assay." *Journal of the American Statistical Association*, 39, 357–65.

_____. (1955). "Maximum Likelihood and Minimum $\chi^2$ Estimates of the Logistic Function." *Journal of the American Statistical Association*, 50, 130–62.

_____. (1957). "Tables for Use in Estimating the Normal Distribution Function by Normit Analysis." *Biometrika*, 44, 411–35.

_____. (1980). "Minimum Chi-Square, Not Maximum Likelihood!" *Annals of Statistics*, 8, 457–87.

Bernard, A., and C. Jones. (1996). "Productivity Across Industries and Countries: Time Series Theory and Evidence." *Review of Economics and Statistics*, 78, 135–46.

Bester, C.A., and C.B. Hansen. (2012). "Grouped Effects Estimators in Fixed Effects Models." *Journal of Econometrics*, **forthcoming**.

Bhargava, A., and J. D. Sargan. (1983). "Estimating Dynamic Random Effects Models from Panel Data Covering Short Time Periods." *Econometrica*, 51, 1635–59.

Binder, M., C. Hsiao, and M.H. Pesaran. (2005). "Estimation and Inference in Short Panel Vector Autoregression with Unit Roots and Cointegration." *Econometric Theory*, 21, 795–837.

Biørn, E. (1981). "Estimating Economic Relations from Incomplete Cross-Section/Time Series Data." *Journal of Econometrics*, 16, 221–36.

_____. (1992). "Econometrics of Panel Data with Measurement Errors," in *Econometrics of Panel Data: Theory and Applications*, edited by L. Mátyás and P. Sevestre, pp. 152–95. Dordrecht: Kluwer Academic.

_____. (2000). "Panel Data with Measurement Errors, Instrumental Variables and GMM Estimators Combining Levels and Differences." *Econometric Reviews*, 19, 391–424.

Biørn, E., and T.J. Klette. (1998). "Panel Data with Errors-in-Variables: Essential and Redundant Orthogonality Conditions in GMM Estimator." *Econometrics Letters*, 59, 275–82.

Biørn, E., and J. Krishnakumar. (2008). "Measurement Errors and Simultaneity," in *The Econometrics of Panel Data*, 3rd eds., edited by L. Mátyás and P. Sevestre, pp. 323–68. Berlin: Springer-Verlag.

Bishop, Y.M., S.E. Fienberg, and P.W. Holland. (1975). *Discrete Multivariate Analysis*: *Theory and Practice*. Cambridge, MA: MIT Press.

Blanchard, P. (1996). "Software Review," in *The Econometrics of Panel Data*, 2nd ed. edited by L. Matyas and P. Sevestre, pp. 879–913. Dordrecht: Kluwer Academic.

Blundell, R., and S. Bond. (1998). "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models." *Journal of Econometrics*, 87, 115–43.

Blundell, R., and R.J. Smith. (1991). "Conditions Initiales et Estimation Efficace dans les Modéles Dynamiques sur Donné es de Panel." *Annals d'Economies et de Statistique*, 20–21, 109–24.

Blundell, R., M. Browning, and C. Meghir. (1994). "Consumer Demand and the Life Cycle Allocation of Household Expenditure." *Review of Economic Studies*, 61, 57–80.

Blundell, R., R. Griffith, and F. Windmeijer. (2002). "Individual Effects and Dynamics in Count Data Models." *Journal of Econometrics*, 102, 113–31.

Bond, S., and C. Meghir. (1994). "Dynamic Investment Models and the Firm's Financial Policy." *Review of Economic Studies*, 61, 197–222.

Borus, M.E. (1981). "An Inventory of Longitudinal Data Sets of Interest to Economists." Mimeo, Ohio State University.

Box, G.E.P., and G.M. Jenkins. (1970). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

Box, G.E.P., and G.C. Tiao. (1968). "Bayesian Estimation of Means for the Random Effects Model." *Journal of the American Statistical Association*, 63, 174–81.

———. (1973). *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley.

Brainard, W.C., and J. Tobin. (1968). "Pitfalls in Financial Model Building." *American Economic Review*, 58, 99–122.

Breitung, J., and S. Das. (2008). "Panel Unit Roots Under Cross Sectional Dependence." *Statistica Nerrlandica*, 59, 414–33.

Breitung, J., and M.H. Pesaran. (2008). "Unit Roots and Cointegration in Panels," in *The Econometrics of Panel Data*, edited by L. Matyas and P. Sevestre, pp. 279–322. Berlin: Springer-Verlag.

Bresson, G., and C. Hsiao. (2011). "A Functional Connectivity Approach for Modelling Cross-Sectional Dependence with an Application to the Estimation of Hedonic Housing Prices in Paris." *Advances in Statistical Analysis*, 95, 501–9.

Bresson, G., C. Hsiao, and A. Pirotte. (2011). "Assessing the Contribution of R&D to Total Productivity – A Bayesian Approach to Account for Heterogeneity and Heteroscedasticity." *Advances in Statistical Analysis*, 95, 435–52.

Breusch, T.S. (1987). "Maximum Likelihood Estimation of Random Effects Models." *Journal of Econometrics*, 36, 383–89.

Breusch, T.S., and A. R. Pagan. (1979). "A Simple Test for Heteroscedasticity and Random Coefficient Variation." *Econometrica*, 47, 1287–94.

———. (1980). "The Lagrange Multiplier Test and His Application to Model Specification in Econometrics." *Review of Economic Studies*, 47, 239–54.

Breusch, T.S., G.E. Mizon, and P. Schmidt. (1989). "Efficient Estimation Using Panel Data." *Econometrica*, 51, 695–700.

Burridge, P. (1980). "On the Cliff-Ord Test for Spatial Autocorrelation." *Journal of Royal Statistical Society* B, 42, 107–8.

Butler, J.S., and R. Moffitt. (1982). "A Computationally Efficient Quadrature Procedure for the One Factor Multinominal Probit Model." *Econometrica*, 50, 761–64.

Cameron, A.C., and P.K. Trevedi. (1998). *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.

Canova, F. (1999). "Testing for Convergence Clubs in Income Per Capita: A Predictive Density Approach." Mimeo, Universitat Pompeu Fabra.

Card, D. (1996). "The Effect of Unions on the Structure of Wages: A Longitudinal Analysis." *Econometrica*, 64, 957–79.

Carro, J.M. (2007). "Estimating Dynamic Panel Data Discrete Choice Models with Fixed Effects." *Journal of Econometrics*, 140, 503–28.

Case, A.C. (1991). "Spatial Patterns in Household Demand." *Econometrica*, 59, 953–65.

Chamberlain, G. (1976). "Identification in Variance Components Models." Discussion Paper No. 486, Harvard Institute of Economic Research.

_____. (1977a). "Education, Income, and Ability Revisited," in *Latent Variables in Socio-Economic Models*, edited by D. J. Aigner and A. S. Goldberger, pp. 143–61. Amsterdam: North-Holland.

_____. (1977b). "An Instrumental Variable Interpretation of Identification in Variance-Components and MIMIC Models," in *Kinometrics: Determinants of Social-Economic Success within and between Families*, edited by P. Taubman, pp. 235–54. Amsterdam: North-Holland.

_____. (1978a). "Omitted Variable Bias in Panel Data: Estimating the Returns to Schooling." *Annales de l'INSEE*, 30–31, 49–82.

_____. (1978b). "On the Use of Panel Data." Paper presented at the Social Science Research Council conference on life-cycle aspects of employment and the labor market, Mt. Kisco, NY.

_____. (1980). "Analysis of Covariance with Qualitative Data." *Review of Economic Studies*, 47, 225–38.

_____. (1982). "Multivariate Regression Models for Panel Data." *Journal of Econometrics*, 18, 5–46.

_____. (1984). "Panel Data," in *Handbook of Econometric*, Vol. II, edited by Z. Griliches and M. Intriligator, pp. 1247–318. Amsterdam: North-Holland.

_____. (1992). "Efficiency Bounds for Semiparametric Regression." *Econometrica*, 60, 567–96.

_____. (1993). "Feedback in Panel Data Models." Mimeo, Department of Economics, Harvard University.

_____. (2010). "Binary Response Models for Panel Data: Identification and Information." *Econometrica*, 78, 159–68.

Chamberlain, G., and Z. Griliches. (1975). "Unobservables with a Variance-Components Structure: Ability, Schooling and the Economic Success of Brothers." *International Economic Review*, 16, 422–50.

Chang, Y. (2002). "Nonlinear IV Unit Root Tests in Panels with Cross-Sectional Dependency." *Journal of Econometrics*, 110, 261–92.

Charlier, E., B. Melenberg, and A. van Soest. (2000). "Estimation of a Censored Regression Panel Data Model Using Conditional Moment Restrictions Efficiently." *Journal of Econometrics*, 95, 25–56.

———. (2001). "An Analysis of Housing Expenditure Using Semiparametric Models and Panel Data." *Journal of Econometrics*, 101, 71–108.

Chen, S. (1999). "Distribution-Free Estimation of the Random Coefficient Dummy Endogenous Variable Model." *Journal of Econometrics*, 91, 171–99.

———. (2000), "Efficient Estimation of Binary Choice Models Under Symmetry." *Journal of Econometrics*, 96, 183–99.

Chen, X., and X. Shen. (1998). "Sieve Extremum Estimates for Weakly Dependent Data." *Econometrica*, 66, 289–314.

Chesher, A.D. (1983). "The Information Matrix Test: Simplified Calculation via a Score Test Interpretation." *Economics Letters*, 13, 45–48.

———. (1984). "Testing for Neglected Heterogeneity." *Econometrica*, 52, 865–72.

Chesher, A.D., and T. Lancaster. (1983). "The Estimation of Models of Labor Market Behavior." *Review of Economic Studies*, 50, 609–24.

Chetty, V.K. (1968). "Pooling of Time Series and Cross-Section Data." *Econometrica*, 36, 279–90.

Chiang, C.L. (1956). "On Regular Best Asymptotically Normal Estimates." *Annals of Mathematical Statistics*, 27, 336–51.

Chintagunta, P., E. Kyriazidou, and J. Perktold. (2001). "Panel Data Analysis of Household Brand Choices." *Journal of Econometrics*, 101, 111–53.

Choi, I. (2001). "Unit Root Tests for Panel Data." *Journal of International Money and Finance*, 20, 249–72.

———. (2002a). "Combination Unit Root Tests for Cross-Sectionally Correlated Panels," in *Econometric Theory and Practice: Frontiers of Analysis and Applied Research, Essays in Honor of P.C.B. Phillips*. Cambridge: Cambridge University Press.

———. (2002b). "Instrumental Variable Estimation of a Nearly Nonstationary, Heterogeneous Error Components Model." *Journal of Econometrics*, 109, 1–32.

———. (2006). "Nonstationary Panels," in *Palgrave Handbooks of Econometrics*, Vol. I, edited by K. Patterson and T.C. Mills, pp. 511–39. New York: Palgrave Macmillan.

Choi, I., and T.K. Chue. (2007). "Subsampling Hypothesis Tests for Nonstationary Panels with Applications to Exchange Rates and Stock Prices." *Journal of Applied Econometrics*, 22, 223–64.

Chow, G.C. (1983). *Econometrics*. New York: McGraw-Hill.

Chui, C.K. (1992). *An Introduction to Wavelets*. San Diego: Academic Press.

Coakley, J., A. Fuertes, and R. Smith. (2006). "Unobserved Heterogeneity in Panel Time Series Models." *Computational Statistics and Data Analysis*, 50, 2361–80.

Coleman, J.S. (1964). *Models of Change and Response Uncertainty*. Englewood Cliffs, NJ: Prentice-Hall.

Collado, M.D. (1997). "Estimating Dynamic Models from Time Series of Independent Cross-Sections." *Journal of Econometrics*, 82, 37–62.

Conley, T.G. (1999). "GMM Estimation with Cross-sectional Dependence." *Journal of Econometrics*, 92, 1–45.

Cooley, T.F., and E.C. Prescott. (1976). "Estimation in the Presence of Stochastic Parameter Variation." *Econometrica*, 44, 167–84.

Cornwell, C., and P. Schmidt. (1984). "Panel Data with Cross-Sectional Variation in Slopes as Well as Intercepts." Mimeo, Michigan State University.

Cosslett, S.R. (1981). "Maximum Likelihood Estimator for Choice-Based Samples." *Econometrica*, 49, 1289–316.

Cox, D.R. (1957). "Note on Grouping." *Journal of the American Statistical Association*, 52, 543–47.

_____. (1962). "Further Results on Tests of Separate Families of Hypotheses." *Journal of the Royal Statistical Society B*, 24, 406–24.

_____. (1970). *Analysis of Binary Data*. London: Methuen.

_____. (1972). "The Analysis of Multivariate Binary Data." *Journal of the Royal Statistical Society C*: *Applied Statistics*, 21, 113–20.

_____. (1975). "Partial Likelihood." *Biometrika*, 62, 269–76.

Crépon, B., and E. Duguet. (1997). "Estimating the Innovation from Patent Numbers: GMM on Count Panel Data." *Journal of Applied Econometrics*, 12, 243–63.

Crépon, B., and J. Mairesse. (1996). "The Chamberlain Approach," in *The Econometrics of Panel Data: A Handbook of the Theory with Applications*, edited by L. Matyas and P. Sevestre, pp. 323–91. Dordrecht: Kluwer Academic.

Cripps, T., and R. Tarling. (1974). "An Analysis of the Duration of Male Unemployment in Great Britain, 1932–1973." *Economic Journal*, 84, 289–316.

Damrongplasit, K., and C. Hsiao. (2009), "Decriminalization Policy and Marijuana Smoking Prevalence: A Look at the Literature." *Singapore Economic Review*, 59, 621–44.

Damrongplasit, K., C. Hsiao, and X. Zhao. (2010). "Decriminalization and Marijuana Smoking Prevalence: Evidence from Australia." *Journal of Business and Economic Statistics*, 28, 344–56.

Davis, P. (2002). "Estimating Multi-Way Error Components Models with Unbalanced Data Structures." *Journal of Econometrics*, 106, 67–95.

Deaton, A. (1985). "Panel Data from Time Series of Cross-Sections." *Journal of Econometrics*, 30, 109–26.

de Finetti, B. (1964). "Foresight: Its Logical Laws, Its Subjective Sources," in *Studies in Subjective Probability*, edited by H.E. Kyburg, Jr. and H.E. Smokler, pp. 93–158. New York: John Wiley & Sons.

Dehejia, R.H., and S. Wahba. (1999). "Propensity Score-Matching Methods for Nonexperimental Causal Studies." *The Review of Economics and Statistics*, 84, 151–61.

Dhrymes, P. (1971). *Distributed Lags: Problems of Estimation and Formulation*. San Francisco: Holden-Day.

Dickey, D.A., and W.A. Fuller. (1979). "Distribution of the Estimators for Autoregressive Time Series with a Unit Root." *Journal of the American Statistical Association*, 74, 427–31.

_____. (1981). "Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root." *Econometrica*, 49, 1057–72.

Dielman, T., T. Nantell, and R. Wright. (1980). "Price Effects of Stock Repurchasing: A Random Coefficient Regression Approach." *Journal of Financial and Quantitative Analysis*, 15, 175–89.

Donald, S., and W. Newey. (2001). "Choosing the Number of Instruments." *Econometrica*, 69, 1161–91.

Driscoll, J., and A. Kraay. (1998). "Consistent Covariance Matrix Estimation with Spatically Dependent Panel Data." *Review of Economics and Statistics*, 80, 549–560.

Duan, J.C., and T. Wang. (2012). "Measuring Distance-to-Default for Financial and Non-Financial Firms." *Global Credit Review*, 2, 95–108.

Duan, J.C., J. Sun, and T. Wang. (2012). "Multiperiod Corporate Default Prediction – A Forward Intensity Approach." *Journal of Econometrics*, 170, 191–209.

Dufour, J.M., and C. Hsiao. (2008). "Identification." In the *New Palgrave*, 2nd ed., edited by L. Blume and S. Durlauf. London and New York: MacMillan.

Duncan, G.M. (1980). "Formulation and Statistical Analysis of the Mixed Continuous/Discrete Dependent Variable Model in Classical Production Theory." *Econometrica*, 48, 839–52.

Durbin, J. (1953). "A Note on Regression When There Is Extraneous Information About One of the Coefficients." *Journal of the American Statistical Association*, 48, 799–808.

_____. (1960). "Estimation of Parameters in Time-Series Regression Models." *Journal of the Royal Statistical Society B*, 22, 139–53.

Durlauf, S.N. (2001). "Manifesto for a Growth Econometrics." *Journal of Econometrics*, 100, 65–69.

Durlauf, S.N., and P. Johnson. (1995). "Multiple Regimes and Cross-Country Growth Behavior." *Journal of Applied Econometrics*, 10, 365–84.

Durlauf, S., and D. Quah. (1999). "The New Empirics of Economic Growth," in *Handbook of Macroeconomics*, edited by J. Taylor and M. Woodford, pp. 235–308, Amsterdam: North-Holland.

Eicker, F. (1963). "Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regression." *Annals of Mathematical Statistics*, 34, 447–56.

Engle, R.F., and C.W.J. Granger. (1987). "Cointegration and Error Correction: Representation, Estimation and Testing." *Econometrica*, 55, 251–76.

Eurostat. (1996). *European Community Household Panel (ECHP)*. Office for Official Publications of the European Communities.

Fan, J., and I. Gijbels. (1992). "Variable Bandwidth and Local Linear Regression Smoothers." *Annals of Statistics*, 20, 1669–2195.

Fan, K.T. and Y.T. Zhang. (1990). *Generalized Multivariate Analysis*. Berlin: Springer.

Fazzari, S.M., R.G. Hubbard, and B.C. Petersen. (1988). "Financing Constraints and Corporate Investment." *Brookings Papers on Economic Activity*, 1, 141–95.

Ferguson, T.S. (1958). "A Method of Generating Best Asymptotically Normal Estimates with Application to the Estimation of Bacterial Densities." *Annals of Mathematical Statistics*, 29, 1046–162.

Fisher, R.A. (1932). *Statistical Methods for Research Workers*, 4th ed. Edinburgh: Oliver and Boyd.

Flinn, C., and J. Heckman. (1982). "New Methods for Analyzing Structural Models of Labour Force Dynamics." *Journal of Econometrics*, 8, 115–68.

Florens, J.P., D. Fougére, and M. Mouchart. (1996). "Duration Models," in *The Econometrics of Panel Data*, 2nd ed., edited by L. Matyas and P. Sevestre, pp. 491–536. Dordrecht: Kluwer Academic.

_____. (2008). "Duration Models and Point Processes," in *The Econometrics of Panel Data*, 3rd ed., edited by L. Mátyás and P. Sevestre, pp. 547–602. Berlin: Springer.

Fougére, G., and T. Kamionka. (1996). "Individual Labour Market Transitions," in *The Econometrics of Panel Data: A Handbook of the Theory with Applications*, 2nd ed., pp. 771–809. Dordrecht: Kluwer Academic.

Frankel, J.A., and A.K. Rose. (1996). "A Panel Project on Purchasing Power Parity: Mean Revision Between and Within Countries." *Journal of International Economics*, 40, 209–44.

Freeman, R.B., and J.L. Medoff. (1981). "The Impact of Collective Bargaining: Illusion or Reality?" Mimeo, Harvard University.

Friedman, M. (1953). *Essays in Positive Economics*. Chicago: University of Chicago Press.

Fujiki, H., C. Hsiao, and Y. Shen. (2002). "Is There a Stable Money Demand Function Under the Low Interest Rate Policy? A Panel Data Analysis." *Monetary and Economic Studies*, Bank of Japan, 20, 1–23.

Fuller, W.A., and G.E. Battese. (1974). "Estimation of Linear Models with Cross-Error Structure." *Journal of Econometrics*, 2, 67–78.

Gelfand, A.E., and A.F.M. Smith. (1990). "Sampling-Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association*, 85, 398–409.

Geweke, J. (1991). "Efficient Simulation from the Multivariate Normal and Student-*t* Distributions Subject to Linear Constraints." *Computer Science and Statistics: Proceedings of the Twenty Third Symposium on the Interface*, 571–78.

Girma, S. (2000). "A Quasi-differencing Approach to Dynamic Modelling from a Time Series of Independent Cross-Sections." *Journal of Econometrics*, 98, 365–83.

Goldberger, A.S. (1964). *Econometric Theory*. New York: John Wiley & Sons.

———. (1972). "Maximum Likelihood Estimation of Regressions Containing Unobservable Independent Variables." *International Economic Review*, 13, 1–15.

Goodman, L.A. (1961). "Statistical Methods for the Mover-Stayer Model." *Journal of the American Statistical Association*, 56, 841–68.

Goodrich, R.L., and P.E Caines. (1979). "Linear System Identification from Nonstationary Cross-Sectional Data." *IEEE Transaction on Automatic Control*, AC-24, 403–11.

Gorseline, D.E. (1932). *The Effect of Schooling Upon Income*. Bloomington: Indiana University Press.

Gourieroux, C., and J. Jasiak. (2000). "Nonlinear Panel Data Models with Dynamic Heterogeneity," in *Panel Data Econometrics*, edited by J. Krishnakumar and E. Ronchetti, pp. 127–48. Amsterdam: North-Holland.

Gourieroux, C., and A. Monfort. (1996). *Simulation-Based Econometric Methods*, Oxford: Oxford University Press.

Granger, C.W.J. (1980). "Long Memory Relationships and the Aggregation of Dynamic Models." *Journal of Econometrics*, 14, 227–38.

Grassetti, L. (2011). "A Note on Transformed Likelihood Approach in Linear Dynamic Panel Models." *Statistical Methods and Applications*, 20, 221–40.

Graybill, F.A. (1969). *Introduction to Matrices with Applications in Statistics*. Belmont, CA: Wadsworth.

Greenaway-McGrevy, R., C. Han, and D. Sul. (2012). "Standardization and Estimation of the Number of Factors for Panel Data." *Journal of Economic Theory and Econometrics*, 79–88.

Griliches, Z. (1957). "Specification Bias in Estimates of Production Functions." *Journal of Farm Economics*, 39, 8–20.

———. (1977). "Estimating the Returns to Schooling: Some Econometric Problems." *Econometrica*, 45, 1–22.

———. (1979). "Sibling Models and Data in Economics: Beginning of a Survey." *Journal of Political Economy*, 87 (Supplement 2), S37–S64.

Griliches, Z., and J.A. Hausman. (1986). "Errors-in-Variables in Panel Data." *Journal of Econometrics*, 31, 93–118.

Griliches, Z., and Y. Grunfeld. (1960). "Is Aggregation Necessarily Bad?" *Review of Economics and Statistics*, 42, 1–13.

Griliches, Z., B. Hall, and J.A. Hausman. (1978). "Missing Data and Self-selection in Large Panels." *Annales de l'INSEE*, 30–31, 137–76.

Groen, J.J.J., and F. Kleibergen. (2003). "Likelihood-Based Cointegration Analysis in Panels of Vector Error-Correction Models." *Journal of Business and Economic Statistics*, 21, 295–318.

Gronau, R. (1976). "The Allocation of Time of Israeli Women." *Journal of Political Economy*, 84, 4, Part II.

Grunfeld, Y. (1958). "The Determinants of Corporate Investment." Unpublished Ph.D. thesis, University of Chicago.

Gupta, A.K., and T. Varga (1993). *Elliptically Contour Models in Statistics*. Dordrecht: Kluwer.

Gurmu, S., and P.K. Trivedi. (1996). "Excess Zeros in Count Models for Recreational Trips." *Journal of Business and Economic Statistics*, 14, 469–77.

Hadri, K. (2000). "Testing for Stationarity in Heterogeneous Panel Data." *Econometrics Journal*, 3, 148–61.

Hadri, K., and R. Larsson. (2005). "Testing for Stationarity in Heterogeneous Panel Data Where the Time Dimension Is Fixed." *Econometric Journal*, 8, 55–69.

Hahn, J. (1998). "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects." *Econometrica*, 66, 315–31.

_____. (1999). "How Informative is the Initial Condition in a Dynamic Panel Model with Fixed Effects?" *Journal of Econometrics*, 93, 309–26.

Hahn, J., and G. Kuersteiner. (2002). "Asymptotically Unbiased Inference for a Dynamic Panel Model with Fixed Effects When Both $n$ and $T$ are Large." *Econometrica*, 70, 1639–57.

_____. (2011). "Bias Reduction in Dynamic Nonnlinear Panel Data Models." *Econometric Theory*, 27, 1152–91.

Hahn, J., and H.R. Moon. (2006). "Reducing Bias of MLE in a Dynamic Panel Model." *Econometric Theory*, 499–512.

Hahn, J., and W. Newey. (2004). "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models." *Econometrica*, 72, 1295–1319.

Hahn, J., J. Hausman, and G. Kuersteiner. (2007). "Long Difference Instrumental Variable Estimation for dynamic Models with Fixed Effects." *Journal of Econometrics*, 140, 574–617.

Hall, B., Z. Griliches, and J.A. Hausman. (1996). "Patents and R&D: Is There a Lag?" *International Economic Review*, 27, 265–83.

Hajivassiliou, V. (1990). "Smooth Simulation Estimation of Panel Data LDV Models." Mimeo, Yale University.

Hall, P., N.I. Fisher, and B. Hoffman. (1992). "On the Nonparametric Estimation of Covariance Functions." Working paper, Australian National University.

Han, A., and J.A. Hausman. (1990). "Flexible Parametric Estimation of Duration and Computing Risk Models." *Journal of Applied Econometrics*, 5, 1–28.

Han, C., and P.C.B. Phillips. (2013). "First Difference Maximum Likelihood and Dynamic Panel Estimation." *Journal of Econometrics*, 175, 35–45.

Hansen, B. (1982). "Efficient Estimation and Testing of Cointegrating Vectors in the Presence of Deterministic Trends." *Journal of Econometrics*, 53, 87–121.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

Harris, M.N., and X. Zhao. (2007). "A Zero-Inflated Ordered Probit Model with an Application to Modeling Tobacco Consumption." *Journal of Econometrics*, 141, 1073–99.

Harris, R.D.F., and E. Tzaralis. (1999). "Inference for Unit Roots in Dynamic Panels Where the Time Dimension Is Fixed." *Journal of Econometrics*, 91, 201–26.

Hartley, H.O., and J.N.K. Rao. (1967). "Maximum Likelihood Estimation for the Mixed Analysis of Variance Model." *Biometrika*, 54, 93–108.

Harvey, A.C. (1978). "The Estimation of Time-Varying Parameters from Panel Data." *Annales de l'INSEE*, 30–31, 203–6.

Harvey, A.C., and G.D.A. Phillips. (1982). "The Estimation of Regression Models with Time-Varying Parameters," in *Games, Economic Dynamics, and Time Series Analysis*, edited by M. Deistler, E. Fürst, and G.S. Schwödiauer, pp. 306–21. Cambridge, MA: Physica-Verlag.

Hausman, J.A. (1978). "Specification Tests in Econometrics." *Econometrica*, 46, 1251–71.

Hausman, J.A., and D. McFadden. (1984). "Specification Tests for the Multinominal Logit Models." *Econometrica*, 52, 1219–40.

Hausman, J.A., B.H. Hall, and Z. Griliches. (1984). "Econometric Models for Count Data with an Application to the Patents-R and D Relationship." *Econometrica*, 52, 909–938.

Hausman, J.A., and W.E. Taylor. (1981). "Panel Data and Unobservable Individual Effects." *Econometrica*, 52, 1219–40.

Hausman, J.A., and D. Wise. (1977). "Social Experimentation, Truncated Distributions, and Efficient Estimation." *Econometrica*, 45, 919–38.

_____. (1978). "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences." *Econometrica*, 46, 403–26.

_____. (1979). "Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment," *Econometrica*, 47, 455–73.

Hayakawa, K. (2009). "On the Effect of Mean-Nonstationary Initial Conditions in Dynamic Panel Data Models." *Journal of Econometrics* 153, 133–35.

Hayashi, F. (1982). "Tobin's Marginal $q$ and Average $q$: A Neoclassical Interpretation." *Econometrica*, 50, 213–24.

Heckman, J.J. (1976a). "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models." *Annals of Economic and Social Measurement*, 5, 475–92.

_____. (1976b). "Simultaneous Equations Models with Continuous and Discrete Endogenous Variables and Structural Shifts," in *Studies in Nonlinear Estimation*, edited by S.M. Goldfeld and R.E. Quandt, pp. 235–72. Cambridge, MA: Ballinger.

_____. (1978a). "Simple Statistical Models for Discrete Panel Data Developed and Applied to Test the Hypothesis of True State Dependence Against the Hypothesis of Spurious State Dependence." *Annales e l'INSEE*, 30–31, 227–69.

_____. (1978b). "Dummy Endogenous Variables in a Simultaneous Equation System." *Econometrica*, 46, 931–59.

_____. (1979). "Sample Selection Bias as a Specification Error." *Econometrica*, 47, 153–61.

_____. (1981a). "Statistical Models for Discrete Panel Data," in *Structural Analysis of Discrete Data with Econometric Applications*, edited by C.F. Manski and D. McFadden; pp. 114–78. Cambridge, MA: MIT Press.

_____. (1981b). "The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Discrete Data Stochastic Process," in *Structural Analysis of Discrete Data with Econometric Applications*, edited by C.F. Manski and D. McFadden, pp. 179–95. Cambridge, MA: MIT Press.

_____. (1981c). "Heterogeneity and State Dependence," in *Studies in Labor Markets*, edited by S. Rosen, pp. 91–139. Chicago: University of Chicago Press.

———. (2001). "Econometrics and Empirical Economics." *Journal of Econometrics*, 100, 3–6.

———. (1997). "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations." *Journal of Human Resources*, 32, 441–462.

Heckman, J.J., and G. Borjas. (1980). "Does Unemployment Cause Future Unemployment? Definitions, Questions and Answers from a Continuous Time Model of Heterogeneity and State Dependence." *Economica*, 47, 247–83.

Heckman, J.J., and R. Robb. (1985). "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data*, edited by J. Heckman and B. Singer. New York: Cambridge University Press.

Heckman, J.J., and B. Singer. (1982). "The Identification Problem in Econometric Models for Duration Data," in *Advances in Econometrics*, edited by W. Hildenbrand, pp. 39–77. Cambridge: Cambridge University Press.

———. (1984). "Econometric Duration Analysis." *Journal of Econometrics*, 24, 63–132.

Heckman, J.J., and E.J. Vytacil. (2001). "Local Instrumental Variables," in *Nonlinear Statistical Inference*, edited by C. Hsiao, K. Morimune and J.L Powell. New York: Cambridge University Press, 1–46.

———. (2005). "Structural Equations, Treatment Effects and Economic Policy Evaluating." *Econometrica*, 669–738.

———. (2007). "Econometric Evaluation of Social Programs," in *Handbook of Econometrics*, Vol. 6B. Amsterdam: North-Holland.

Heckman, J.J., and R. Willis. (1977). "A Beta-Logistic Model for the Analysis of Sequential Labor Force Participation by Married Women." *Journal of Political Economy*, 85, 27–58.

Heckman, J.J., H. Ichimura, and P. Todd. (1998). "Matching as an Econometric Evaluations Estimator." *Review of Economic Studies*, 65, 261–94.

Heckman, J.J., D.A. Schmierer and S.S. Urzua. (2010). "Testing the Correlated Random Coefficient Model." *Journal of Econometrics*, 158, 177–203.

Henderson, C.R., Jr. (1971). "Comment on 'The Use of Error Components Models in Combining Cross-Section with Time Series Data'." *Econometrica*, 39, 397–401.

Hendricks, W., R. Koenker, and D.J. Poirier. (1979). "Residential Demand for Electricity: An Econometric Approach." *Journal of Econometrics*, 9, 33–57.

Hildreth, C., and J.P. Houck. (1968). "Some Estimators for a Linear Model with Random Coefficients." *Journal of the American Statistical Association*, 63, 584–95.

Hirano, K., G.W. Imbens and G. Ridder. (2003). "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica*, 76, 1661–89.

Hirano, K., G.W. Imbens, G. Ridder, and D.B. Rubin. (2001). "Combining Panel Data Sets with Attrition and Refreshment Samples." *Econometrica*, 69, 1645–60.

Hoch I. (1962). "Estimation of Production Function Parameters Combining Time-Series and Cross-Section Data." *Econometrica*, 30, 34–53.

Holly, A. (1982). "A Remark on Hausman's Specification Test." *Econometrica*, 50, 749–59.

Holly, A., and L. Gardiol. (2000). "A Score Test for Individual Heteroscedasticity in a One-Way Error Components Model," in *Panel Data Econometrics*, edited by J. Krishnakumkar and E. Ronchetti, pp. 199–211. Amsterdam: North-Holland.

Holtz-Eakin, D., W. Newey, and H.S. Rosen. (1988). "Estimating Vector Autoregressions with Panel Data." *Econometrica*, 56, 1371–95.

Hood, Wm. C. and T.C. Koopmans. (1953). *Studies in Econometric Method*. Cowles Foundation Monographs 13. New York: John Wiley.

Hong, Y., and C. Kao. (2004). "Wavelet-Based Testing for Serial Correlation of Unknown Form in Panel Models." *Econometrica*, 72, 1519–63.

Hong, Y., and H. White. (1995). "Consistent Specification Testing via Nonparametric Series Regression." *Econometrica*, 63, 1133–59.

Honoré, B.E. (1992). "Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects." *Econometrica*, 60, 533–67.

_____. (1993). "Orthogonality Conditions for Tobit Models with Fixed Effects and Lagged Dependent Variables." *Journal of Econometrics*, 59, 35–61.

Honoré, B.E., and E. Kyriazidou. (2000a). "Panel Data Discrete Choice Models with Lagged Dependent Variables." *Econometrica*, 68, 839–74.

_____. (2000b). "Estimation of Tobit-Type Models with individual Specific Effects." *Econometrics Review*, 19.

Honoré, B.E., and J.L. Powell. (1994). "Pairwise Difference Estimators of Censored and Truncated Regression Models." *Journal of Econometrics*, 64, 241–78.

Honoré, B.E., and E. Tamer. (2006). "Bounds on Parameters in Panel Dynamic Discrete Choice Models." *Econometrica*, 74, 611–29.

Horowitz, J.L. (1992). "A Smoothed Maximum Score Estimator for the Binary Response Model." *Econometrica*, 60, 505–31.

_____. (1996). "Semiparametric Estimation of a Regression Model with an Unknown Transformation of the Dependent Variable." *Econometrica*, 64, 103–37.

Hotelling, H. (1931). "The Generalization of Student's Ratio." *Annals of Mathematical Statistics*, 2, 360–78.

Hsiao, C. (1974a). "Statistical Inference for a Model with Both Random Cross-Sectional and Time Effects." *International Economic Review*, 15, 12–30.

_____. (1974b). "The Estimation of Labor Supply of Low Income Workers – Some Econometric Considerations." Working Paper 970–71, The Urban Institute, Washington, D.C.

_____. (1975). "Some Estimation Methods for a Random Coefficients Model." *Econometrica*, 43, 305–25.

_____. (1976). "Regression Analysis with Limited Dependent Variable." 1P-186, IBER and CRMS, University of California, Berkeley.

_____. (1979a). "Causality Tests in Econometrics." *Journal of Economic Dynamics and Control*, 1, 321–46.

_____. (1979b). "Autoregressive Modelling of Canadian Money and Income Data." *Journal of the American Statistical Association*, 74, 553–60.

_____. (1982). "Autoregressive Modelling and Causal Ordering of Economic Variables." *Journal of Economic Dynamics and Control*, 4, 243–59.

_____. (1983). "Identification," in *Handbook of Econometrics*, Vol. I, edited by Z. Griliches and M. Intriligator, pp. 223–83. Amsterdam: North-Holland.

_____. (1985a). "Benefits and Limitations of Panel Data." *Econometric Reviews*, 4, 121–74.

_____. (1985b). "Minimum Chi-Square," in the *Encyclopedia of Statistical Science*, Vol. 5, edited by S. Kotz and N. Johnson, pp. 518–22. New York: John Wiley & Sons.

_____. (1989). "Consistent Estimation or Some Nonlinear Errors-in-Variables Models." *Journal of Econometrics*, 41, 159–85.

_____. (1991a). "A Mixed Fixed and Random Coefficients Framework for Pooling Cross-Section and Time Series Data." Paper presented at the Third conference on Telecommunication Demand Analysis with Dynamic Regulation, Hilton Head, S.

Carolina, in New Development in Quantitative Economics, edited by J.W. Lee and S.Y. Zhang. Beijing: Chinese Academic of Social Science.

_____. (1991b). "Identification and Estimation of Latent Binary Choice Models Using Panel Data." *Review of Economic Studies*, 58, 717–31.

_____. (1992a). "Random Coefficients Models," in *The Econometrics of Panel Data*, ed. by L. Matyas and P. Sevestres, Kluwer, 1st ed., pp. 223–41, 2nd ed. (1996), pp. 410–28.

_____. (1992b). "Logit and Probit Models," in *The Econometrics of Panel Data*, edited by L. Matyas and P. Sevestre. Dordrecht: Kluwer Academic, 223–41.

_____. (1992c). "Nonlinear Latent Variables Models," in *Econometrics of Panel Data*, edited by L. Matyas and P. Sevestre, pp. 242–61. Dordrecht: Kluwer Academic.

_____. (1995). "Panel Analysis for Metric Data." *Handbook of Statistical Modelling in the Social and Behavioral Sciences*, edited by G. Arminger, C.C. Clogg, and M.Z. Sobel, pp. 361–400. New York: Plenum Press.

_____. (2001). "Economic Panel Data," in *International Encyclopedia of the Social and Behavioral Sciences*, edited by N.J. Snelser and P.B. Bates. Vol. 6, 4114–121. Oxford: Elsevier.

_____. (2007). "Panel Data Analysis-Advantages and Challenges." *TEST*, 16, 1–22.

_____. (2011). "Dynamic Panel Data Models," in *Handbook of Empirical Economics and Finance*, edited by D. Giles and A. Ullah, pp. 373–96. Boca Raton, FL: Taylor and Francis Group.

_____. (2012). "The Creative Tension between Statistics and Econometrics." *Singapore Economic Review*, 57, 125007-1-11.

Hsiao, C., and D.C. Mountain. (1994). "A Framework for Regional Modeling and Impact Analysis – An Analysis of the Demand for Electricity by Large Municipalities in Ontario, Canada." *Journal of Regional Science*, 34, 361–85.

Hsiao, C. and M.H. Pesaran. (2008). "Random Coefficients Models," in *The Econometrics of Panel Data*, 3rd ed., edited by L. Matayas and P. Sevestre, pp. 187–216. Berlin: Springer Science+Business Media.

Hsiao, C., and B.H. Sun. (2000). "To Pool or Not to Pool Panel Data," in *Panel Data Econometrics*:*Future Directions, Papers in Honor of Professor Pietro Balestra*, edited by J. Krishnakumar and E. Ronchetti, pp. 181–98. Amsterdam: North Holland.

Hsiao, C., and A.K. Tahmiscioglu. (1997). "A Panel Analysis of Liquidity Constraints and Firm Investment." *Journal of the American Statistical Association*, 92, 455–65.

_____. (2008). "Estimation of dynamic Panel Data Models with Both Individual and Time Specific Effects." *Journal of Statistical Planning and Inference*, 138, 2698–721.

Hsiao, C., and G. Taylor. (1991). "Some Remarks on Measurement Errors and the Identification of Panel Data Models." *Statistica Neerlandica*, 45, 187–94.

Hsiao, C., and K.Q. Wang. (2000). "Estimation of Structural Nonlinear Errors-in-Variables Models by Simulated Least Squares Method." *International Economic Review*, 41, 523–42.

Hsiao, C., and J. Zhang. (2013). "IV, GMM or MLE to Estimate Dynamic Panel Data Models When Both $N$ and $T$ are Large." Mimeo, University of Southern California.

Hsiao, C., and Q. Zhou. (2013). "Statistical Inference for Panel Dynamic Simultaneous Equations Models When Both $N$ and $T$ Are Large." Mimeo, University of Southern California.

Hsiao, C., T.W. Appelbe, and C.R. Dineen. (1993). "A General Framework for Panel Data Analysis – with an Application to Canadian Customer Dialed Long Distance Service." *Journal of Econometrics*, 59, 63–86.

Hsiao, C., D.C. Mountain, and K.F. Ho-Illman. (1995). "Bayesian Integration of End-Use Metering and Conditional Demand Analysis." *Journal of Business and Economic Statistics*, 13, 315–26.

Hsiao, C., H.S. Ching, and S. Wan. (2012). "A Panel Data Approach for Program Evaluation – Measuring the Benefits of Political and Economic Integration of Hong Kong with Mainland China." *Journal of Applied Econometrics*, 27, 705–40.

Hsiao, C., K. Morimune, and J.L. Powell. (2001). *Nonlinear Statistical Inference*. New York: Cambridge University Press.

Hsiao, C., J. Nugent, I. Perrigne, and J. Qiu. (1998). "Shares versus Residual Claimant Contracts: The Case of Chinese TVEs." *Journal of Comparative Economics*, 26, 317–37.

Hsiao, C., M.H. Pesaran, and A. Picks. (2012). "Diagnostic Tests of Cross-Section Independence for Limited Dependent Variable Panel Data Models." *Oxford Bulletin of Economics and Statistics*, 74, 253–77.

Hsiao, C., M.H. Pesaran, and A.K. Tahmiscioglu. (1999). "Bayes Estimation of Short-Run Coefficients in Dynamic Panel Data Models," in *Analysis of Panels and Limited Dependent Variables Models*, edited by C. Hsiao, L.F. Lee, K. Lahiri, and M.H. Pesaran, pp. 268–96. Cambridge: Cambridge University Press.

_____. (2002). "Maximum Likelihood Estimation of Fixed Effects Dynamic Panel Data Models Covering Short Time Periods." *Journal of Econometrics*, 109, 107–50.

Hsiao, C., Y. Shen, and H. Fujiki. (2005). "Aggregate vs. Disaggregate Data Analysis – A Paradox in the Estimation of Money Demand Function of Japan Under the Low Interest Rate Policy." *Journal of Applied Econometrics*, 20, 579–601.

Hsiao, C., L.Q. Wang, and K.Q. Wang. (1997). "Estimation of Nonlinear Errors-in-Variables Models - An Approximate Solution." *Statistical Papers*, 38, 1–28.

Hsiao, C., Q. Li, Z. Liang, and W. Xie. (2012). "Correlated Random Coefficients Models." Mimeo, University of Southern California.

Hsiao, C., K. Lahiri, L.F. Lee, and M.H. Pesaran. (1999). *Analysis of Panel Data and Limited Dependent Variable Models*. Cambridge: Cambridge University Press.

Hsiao, C., D.C. Mountain, K.Y. Tsui, and M.W. Luke Chan. (1989). "Modeling Ontario Regional Electricity System Demand Using a Mixed Fixed and Random Coefficients Approach." *Regional Science and Urban Economics*, 19, 567–87.

Hsiao, C., Y. Shen, B. Wang, and G. Weeks. (2007). "Evaluating the Impacts of Washington State Repeated Job Search Services on the Earnings of Prime-Age Female TANF Recipients." *Journal of Applied Econometrics*, 22, 453–75.

_____. (2008). "Evaluating the Effectiveness of Washington State Repeated Job Search Services on the Employment Rate of Prime-age Female Welfare Recipients." *Journal of Econometrics*, 145, 98–108.

Hu, L. (1999). "Estimating of a Censored Dynamic Panel Data Model with an Application to Earnings Dynamics." Mimeo, Department of Economics, Princeton University.

_____. (2002). "Estimation of a Censored Dynamic Panel Data Model." *Econometrica*, 70, 2499–517.

Hurvich, C.M., and C.L. Tsai. (1989). "Regression and Time Series Model Selections in Small Samples." *Biometrika*, 76, 297–307.

Hurwicz, L. (1950). "Systems with Nonadditive Disturbances," in *Statistical Inference in Dynamic Economic Models*, edited by T. C. Koopmans, pp. 330–72. New York: John Wiley & Sons.

Hyslop, D. (1999). "State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Women." *Econometrica*, 52, 363–89.

Im, K.S., J. Lee, and M. Tieslau. (2005). "Panel LM Unit Root Tests with Level Shifts." *Oxford Bulletin of Economics and Statistics*, 67, 393–419.

Im, K.S., M.H. Pesaran, and Y. Shin. (2003). "Testing for Unit Roots in Heterogeneous Panels." *Journal of Econometrics*, 115, 53–74.

Imbens, G.W., and J.D. Angrist. (1994). "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62, 467–75.

Imbens, G.W., and T. Lemieux. (2008). "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics*, 142, 615–35.

Inoue, A. (2008). "Efficient Estimation and Inference in Linear Pseudo-Panel Data Models." *Journal of Econometrics*, 148, 449–66.

International Monetary Fund. (2012). World Economic Outlook, Oct. 2012.

Intriligator, M.D., R.G. Bodkin, and C. Hsiao. (1996). *Econometric Models, Techniques, and Applications*. Upper Saddle River, NJ: Prentice-Hall.

Izan, H.Y. (1980). "To Pool or not to Pool? A Reexamination of Tobin's Food Demand Problem." *Journal of Econometrics*, 13, 391–402.

Jang, S.J., and T.P. Thornberry. (1998). "Self Esteem, Delinquent Peers, and Delinquency: A Test of Self-Enhancement Thesis." *American Sociological Review*, 63, 586–98.

Janz, N., G. Ebling, S. Gottshalk, and H. Niggemann. (2001). "The Mannheim Innovation Panels (MIP and MIP-S) of the Centre for European Economic Research (ZEW). *Schmollers Jahrbuch*, 121, 123–29.

Jeong, K.J. (1978). "Estimating and Testing a Linear Model When an Extraneous Information Exists." *International Economic Review*, 19, 541–43.

Johansen, S. (1991). "Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models." *Econometrica*, 59, 1551–80.

Johansen, S. (1995). *Likelihood Based Inference on Cointegration in the Vector Autoregressive Model*. Oxford: Oxford University Press.

Johnston, J. (1972). *Econometric Methods*, 2nd ed. New York: McGraw-Hill.

Jöreskog, K.G., and A.S. Goldberger. (1975). "Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable." *Journal of the American Statistical Association*, 70, 631–39.

Jorgenson, D.W. (1971). "Econometric Studies of Investment Behavior: A Survey." *Journal of Economic Literature*, 9, 1111–47.

Jorgenson, D.W., and T.M. Stokes. (1982). "Nonlinear Three Stage Least Squares Pooling of Time Series and Cross Section Data." Discussion Paper No. 952, Harvard Institute of Economic Research.

Judge, G., W. Griffiths, R. Hill, and T. Lee. (1980). *The Theory and Practice of Econometrics*. New York: John Wiley & Sons.

Judson, R.A., and A.L. Owen. (1999). "Estimating Dynamic Panel Data Models: A Guide for Macroeconomists." *Economic Letters*, 65, 9–15.

Juster, T. (2001). "Microdata Bases: Economics," in *International Encyclopedia of Social Sciences*, ed. by N.J. Smelser and P.B. Bates, Vol. 14, 9770–777. Amsterdam: Elsevier.

Kalman, R.E. (1960). "A New Approach to Linear Filtering and Prediction Problems." *Transactions of ASME, Series D: Journal of Basic Engineering*, 82, 35–45.

Kao, C. (1999). "Spurious Regression and Residual-Based Tests for Cointegration in Panel Data." *Journal of Econometrics*, 90, 1–44.

Kao, C., and M.H. Chiang. (2000). "On the Estimation and Inference of a Cointegrated Regression in Panel Data," in *Advances in Econometrics*, Vol. 15, edited by B. Baltagi, pp. 161–78. Amsterdam: JAI Press.

Kao, C., and J.F. Schnell. (1987a). "Errors in Variables in Panel Data with Binary Dependent Variable." *Economic Letters*, 24, 45–49.

_____. (1987b). "Errors-in-Variables in a Random Effects Probit Model for Panel Data." *Economic Letters*, 24, 339–42.

Kapetanios, G., M.H. Pesaran, and T. Yamagata. (2011). "Panels with Non-Stationary Multifactor Error Structures." *Journal of Econometrics*, 160, 326–48.

Kapoor, M., H. Kelejian, and I. Prucha. (2007). "Panel Data Models with Spatially Correlated Error Components." *Journal of Econometrics*, 140, 97–130.

Karlin, S., and H. Taylor. (1975). *A First Course in Stochastic Processes*, 2nd ed. New York: Academic Press.

Kato, K., A.F. Galvao, and G.V. Montes-Rojas. (2012). "Asymptotics for Panel Quantile Regression Models with Individual Effects." *Journal of Econometrics* 170, 76–91.

Kaufman, G.M. (1977). "Posterior Inference for Structural Parameters Using Cross Section and Time Series Data," in *Studies in Bayesian Econometrics and Statistics, in Honor of L. J. Savage*, Vol. 2, edited by S. Fienberg and A. Zellner, pp. 73–94. Amsterdam: North-Holland.

Keane, M.P. (1994). "A Computationally Practical Simulation Estimator for Panel Data." *Econometrica*, 62, 95–116.

Kelejian, H.H. (1977). "Random Parameters in Simultaneous Equation Framework: Identification and Estimation." *Econometrica*, 42, 517–27.

Kelejian, J.H., and I.R. Prucha. (2001). "On the Asymptotic Distribution of the Moran I Test Statistic with Application." *Journal of Econometrics*, 104, 219–57.

Kelejian, H.H., and S. W. Stephan. (1983). "Inference in Random Coefficient Panel Data Models: A Correction and Clarification of the Literature." *International Economic Review*, 24, 249–54.

Kiefer, J., and J. Wolfowitz. (1956). "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters." *Annals of Mathematical Statistics*, 27, 887–906.

Kiefer, N.M. (1979). "Population Heterogeneity and Inference from Panel Data on the Effects of Vocational Education." *Journal of Political Economy*, 87 (Pt. 2), S213–S226.

_____. (1980). "Estimation of Fixed Effects Models for Time Series of Cross-Sections with Arbitrary Intertemporal Covariance." *Journal of Econometrics*, 14, 195–202.

_____. (1988). "Economic Duration Data and Hazard Functions." *Journal of Economic Literature*, 26, 646–79.

Kim, J. and D. Pollard. (1990). "Cube Root Asymptotics." *Annals of Statistics*, 18, 191–219.

Kiviet, J.F. (1995). "On Bias Inconsistency and Efficiency in Various Estimators of Dynamic Panel Data Models." *Journal of Econometrics*, 68, 53–78.

Kiviet, J.F., and G.D.A. Phillips. (1993). "Alternative Bias Approximation with a Lagged Dependent Variables." *Econometric Theory*, 9, 62–80.

Klein, L.R. (1953). *A Textbook of Econometrics*. Evanston, IL: Row Peterson.

_____. (1988). "The Statistical Approach to Economics." *Journal of Econometrics*, 37, 7–26.

Klein, R., and R. Spady. (1993). "An Efficient Semiparametric Estimator for Binary Response Models." *Econometrica*, 61(2), 387–423.

Koenker, R. (2004). "Quantile Regression for Longitudinal Data." *Journal of Multivariate Analysis*, 91, 74–89.

Koenker, R., and G. Bassett. (1978). "Regression Quantiles." *Econometrica*, 46, 33–50.

Koenker, R., and J.A.F. Machado. (1999). "GMM Inference When the Number of Moment Conditions Is Large." *Journal of Econometrics*, 93, 327–344.

Krishnakumar, J., and E. Ronchetti. (2000). *Panel Data Econometrics: Future Directions, Papers in Honor of Professor Pietro Balestra*. Amsterdam: North-Holland.

Kruiniger, H. (2009). "GMM Estimation and Inference in Dynamic Panel Data Models with Persistent Data." *Econometric Theory*, 25, 1348–91.

Kuh, E. (1959). "The Validity of Cross Sectionally Estimated Behavior Equations in Time Series Applications." *Econometrica*, 27, 197–214.

_____. (1963). *Capital Stock Growth: A Micro-Econometric Approach*. Amsterdam: North-Holland.

Kuh, E., and J.R. Meyer. (1957). "How Extraneous Are Extraneous Estimates?" *Review of Economics and Statistics*, 39, 380–93.

Kwiatkowski, D., P.C.B. Phillips, P. Schmidt, and Y. Shin. (1992). "Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root." *Journal of Econometrics*, 54, 159–78.

Kyriazidou, E. (1997). "Estimation of a Panel Data Sample Selection Model." *Econometrica*, 65, 1335–64.

_____. (2001). "Estimation of Dynamic Panel Data Sample Selection Models." *Review of Economic Studies*, 68, 543–72.

LaLonde, R. (1986). "Evaluating the Econometric Evaluations of Training Programs." *American Economic Review*, 76, 604–20.

Lamarche, C. (2010). "Robust Penalized Quantile Regression Estimation for Panel Data." *Journal of Econometrics*, 157, 396–408.

Lancaster, T. (1984). "The Covariance Matrix of the Information Matrix Test." *Econometrica*, 52, 1051–53.

_____. (1990). *The Econometric Analysis of Transition Data*. New York: Cambridge University Press.

_____. (2001). "Some Econometrics of Scarring," in *Nonlinear Statistical Inference*, edited by C. Hsiao, K. Morimune and J.L. Powell, pp. 393–402. New York: Cambridge University Press.

Layton, L. (1978). "Unemployment Over the Work History." Ph.D. dissertation, Department of Economics, Columbia University.

Lee, L.F. (1978a). "Unionism and Wage Rates: A Simultaneous Equations Model with Qualitative and Limited Dependent Variables." *International Economic Review*, 19, 415–34.

_____. (1978b). "On the Issues of Fixed Effects vs. Random Effects Econometric Models with Panel Data." Discussion Paper 78–101, University of Minnesota.

_____. (1979). "Efficient Estimation of Dynamic Error Components Models with Panel Data." Discussion Paper No. 79–118, Center for Economic Research, University of Minnesota.

_____. (1982). "Specification Error in Multinominal Logit Models: Analysis of the Omitted Variable Bias." *Journal of Econometrics*, 20, 197–209.

_____. (1987). "Nonparametric Testing of Discrete Panel Data Models." *Journal of Econometrics*, 34, 147–78.

_____. (2002). "Consistency and Efficiency of Least Squares Estimation for Mixed Regressive, Spatial Autoregressive Models." *Econometric Theory*, 18, 252–77.

_____. (2003). "Best Spatial Two-Stage Least Squares Estimators for a Spatial Autoregressive Model with Autoregressive Disturbances." *Econometric Reviews*, 22, 307–35.

_____. (2004). "Asymptotic Distribution of Quasi-Maximum Likelihood Estimators for Spatial Autoregressive Models." *Econometrica*, 72, 1899–925.

Lee, L.F., and W.E. Griffiths. (1979). "The Prior Likelihood and Best Linear Unbiased Prediction in Stochastic Coefficient Linear Models," University of New England Working Papers in Econometrics and Applied Statistics, No. 1.

Lee, L.F., and J. Yu (2010a). "Estimation of Spatial Autoregressive Panel Data Models with Fixed Effects." *Journal of Econometrics*, 154, 165–85.

_____. (2010b). "A Spatial Dynamic Panel Data Model with Both Time and Individual Fixed Effects." *Econometric Theory*, 26, 564–97.

_____. (2010c). "Some Recent Developments in Spatial Panel Data Models." *Regional Science and Urban Economics*, 40, 255–71.

Lee, M.J. (1999). "A Root-N Consistent Semiparametric Estimator for Related Effects Binary Response Panel Data." *Econometrica*, 67, 427–33.

Levin, A., and C. Lin. (1993). "Unit Root Tests in Panel Data: Asymptotic and Finite Sample Properties." Mimeo, University of California, San Diego.

Levin, A., C. Lin, and J. Chu. (2002). "Unit Root Tests in Panel Data: Asymptotic and Finite-Sample Properties." *Journal of Econometrics*, 108, 1–24.

Lewbel, A. (1992). "Aggregation with Log Linear Models." *Review of Economic Studies*, 59, 535–43.

Lewbel, A. (1994). "Aggregation and Simple Dynamics." *American Economic Review*, 84, 905–918.

Li, M., and J.L. Tobias. (2011). "Bayesian Inference in a Correlated Random Coefficients Model: Modeling Causal Effect Heterogeneity with an Application to Heterogeneous Returns to Schooling." *Journal of Econometrics*, 162, 345–61.

Li, Q., and C. Hsiao. (1998). "Testing Serial Correlation in Semi-parametric Panel Data Models." *Journal of Econometrics*, 87, 207–37.

Li, Q., and J.S. Racine. (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton, NJ: Princeton University Press.

Li, Q., and T. Stengos. (1996). "Semi-parametric Estimation of Partially Linear Panel Data Models." *Journal of Econometrics*, 71, 389–97.

Liesenfeld, L., and J.F. Richard. (2008). "Simulation Techniques for Panels: Efficient Importance Sampling," in *The Econometrics of Panel Data*, 3rd ed., edited by L. Mátyas and P. Severstre, pp. 419–50. Berlin: Springer-Verlag.

Lillard, L.A., and Y. Weiss. (1979). "Components of Variation in Panel Earnings Data: American Scientists 1960–70." *Econometrica*, 47, 437–54.

Lillard, L.A., and R. Willis. (1978). "Dynamic Aspects of Earnings Mobility." *Econometrica*, 46, 985–1012.

Lin, C.C., and S. Ng. (2012). "Estimation of Panel Data Models with Parameter Heterogeneity When Group Membership is Unknown." *Journal of Econometric Methods*, 1, 42–55.

Lindley, D.V., and A.F.M. Smith. (1972). "Bayes Estimates for the Linear Model," and Discussion." *Journal of the Royal Statistical Society B*, 34, 1–41.

Little, R.J.A., and D.B. Rubin. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Liu, E., C. Hsiao, T. Matsumoto, and S. Chou. (2009). "Maternal Full-Time Employment and Overweight Children: Parametric, Semi-Parametric and Non-parametric Assessment." *Journal of Econometrics*, 152, 61–69.

Liu, L.M., and G.C. Tiao. (1980). "Random Coefficient First-Order Autoregressive Models." *Journal of Econometrics*, 13, 305–25.

Liu, T.C. (1960). "Underidentification, Structural Estimation, and Forecasting." *Econometrica*, 28, 855–65.

MaCurdy, T.E. (1981). "An Empirical Model of Labor Supply in a Life Cycle Setting." *Journal of Political Economy*, 89, 1059–85.

_____. (1982). "The Use of Time Series Processes to Model the Error Structure of Earnings in a Longitudinal Data Analysis." *Journal of Econometrics*, 18, 83–114.

Maddala, G.S. (1971a). "The Use of Variance Components Models in Pooling Cross Section and Time Series Data," *Econometrica*, 39, 341–58.

_____. (1971b). "The Likelihood Approach to Pooling Cross-Section and Time Series Data." *Econometrica*, 39, 939–53.

_____. (1983). *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Maddala, G.S., and T.D. Mount. (1973). "A Comparative Study of Alternative Estimators for Variance Components Models Used in Econometric Applications." *Journal of the American Statistical Association*, 68, 324–28.

Maddala, G.S., and S. Wu. (1999). "A Comparative Study of Unit Root Tests with Panel Data and a New Simple Test." *Oxford Bulletin of Economics and Statistics*, 61, 631–52.

Magnus, J.R., and H. Neudecker. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics*, revised ed. New York: John Wiley & Sons.

Mairesse, J. (1990). "Time-series and Cross-sectional Estimates on Panel Data: Why are they Different and Why Should They Be Equal?" in *Panel Data and Labor Market Studies*, edited by J. Hartog, G. Ridder, and J. Theeuwes, pp. 81–95. Amsterdam: North-Holland.

_____. (2007). "Comment on Panel Data Analysis – Advantages and Challenges." *TEST*, 16, 37–41.

Malinvaud, E. (1970). *Statistical Methods of Econometrics*, 2nd ed. Amsterdam: North-Holland.

Mankiw, N.G., D. Romer, and D. Weil. (1992). "A Contribution to the Empirics of Economic Growth." *Quarterly Journal of Economics*, 107, 407–37.

Manski, C.F. (1975). "Maximum Score Estimation of the Stochastic Utility Model of Choice." *Journal of Econometrics*, 3, 205–28.

_____. (1985). "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator." *Journal of Econometrics*, 27, 313–33.

_____. (1987). "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data." *Econometrica*, 55, 357–62.

Manski, C.F., and E. Tamer. (2002). "Inference on Regressions with Interval Data on a Regressor or Outcome." *Econometrica*, 70, 519–46.

Mao, G., and Y. Shen. (2013). "Bubbles or Fundamentals? Modeling Provincial House Prices in China Allowing for Cross-Sectional Dependence." Mimeo, National School of Development, Peking University.

Martins-Filho, C., and F. Yao. (2009). "Nonparametric Regression Estimation with General Parametric Error Covariance." *Journal of Multivariate Analysis*, 100, 309–33.

Matyás, L., and P. Sevestre. (1996). "The Econometrics of Panel Data: A Handbook of the Theory with Applications," 2nd ed. Dordrecht: Kluwer Academic.

Mazodier, P., and A. Trognon. (1978). "Heteroscedasticity and Stratification in Error Components Models." *Annales de l'INSEE*, 30–31, 451–82.

McCullah, P. (1987). *Tensor Methods in Statistics*. London: Chapman and Hall.

McCoskey, S., and C. Kao. (1998). "A Residual-Based Test of the Null of Cointegration in Panel Data." *Econometric Reviews*, 17, 57–84.

McFadden, D. (1974). "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers in Econometrics*, edited by P. Zarembka, pp. 105–42. New York: Academic Press.

_____. (1976). "Quantal Choice Analysis: A Survey." *Annals of Economic and Social Measurement*, 5, 363–90.

_____. (1984). "Econometric Analysis of Qualitative Response Models," in *Handbook of Econometrics*, Vol. II, edited by Z. Griliches and M. D. Intriligator, pp. 1395–457. Amsterdam: North-Holland.

_____. (1989). "A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration." *Econometrica*, 57, 995–1026.

McKenzie, D.J. (2004). "Asymptotic Theory for Heterogeneous Dynamic Pseudo-Panels." *Journal of Econometrics*, 120, 235–62.

Mehta, J.S., G.V.L. Narasimham, and P.A.V.B. Swamy. (1978). "Estimation of a Dynamic Demand Function for Gasoline with Different Schemes of Parameter Variation." *Journal of Econometrics*, 7, 263–79.

Merton, R.C. (1974). "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates." *Journal of Finance*, 29, 449–70.

Meyer, J.R., and E. Kuh. (1957). *The Investment Decision: An Empirical Study*. Cambridge, MA: Harvard University Press.

Miller, J.J. (1977). "Asymptotic Properties of Maximum Likelihood Estimates in the Mixed Model of the Analysis of Variance." *Annals of Statistics*, 5, 746–62.

Miller, M.H., and F. Modigliani. (1961). "Dividend Policy, Growth and the Valuation of Shares." *Journal of Business*, 34, 411–33.

Min, C.K., and A. Zellner. (1993). "Bayesian and Non-Bayesian Methods for Combining Models and Forecasts with Applications to Forecasting International Growth Rate." *Journal of Econometrics*, 56, 89–118.

Modigliani, F., and M.H. Miller. (1958). "The Cost of Capital, Corporation Finance and the Theory of Investment." *American Economic Review*, 48, 261–97.

Moffitt, R. (1993). "Identification and Estimation of Dynamic Models with a Time Series of Repeated Cross-Sections." *Journal of Econometrics*, 59, 99–123.

Moffitt, R., J. Fitzgerald, and P. Gottschalk. (1997). "Sample Selection in Panel Data: The Role of Selection on Observables." Mimeo, Johns Hopkins University.

Moon, H.R., and B. Perron. (2004). "Testing for a Unit Root in Panels with Dynamic Factors." *Journal of Econometrics*, 122, 81–126.

Mundlak, Y. (1961). "Empirical Production Function Free of Management Bias." *Journal of Farm Economics*, 43, 44–56.

_____. (1978a). "On the Pooling of Time Series and Cross Section Data." *Econometrica*, 46, 69–85.

_____. (1978b). "Models with Variable Coefficients: Integration and Extension." *Annales de l'INSEE*, 30–31, 483–509.

Nagar, A.L. (1959). "The Bias and Moment Matrix of $k$-class Estimators of the Parameters in Simultaneous Equations." *Econometrica*, 27, 575–95.

Nerlove, M. (1965). *Estimation and Identification of Cobb-Douglas Production Functions*. Chicago: Rand McNally.

_____. (1967). "Experimental Evidence on the Estimation of Dynamic Economic Relations from a Time Series of Cross-Sections." *Economic Studies Quarterly*, 18, 42–74.

_____. (1971a). "Further Evidence on the Estimation of Dynamic Economic Relations from a Time Series of Cross Sections." *Econometrica*, 39, 359–82.

_____. (1971b). "A Note on Error Components Models." *Econometrica*, 39, 383–96.

_____. (2000). "An Essay on the History of Panel Data Econometrics." Paper presented at 2000 Panel Data Conference in Geneva.

Newey, W. (1997). "Convergence Rate and Asymptotic Normality for Series Estimators." *Journal of Econometrics*, 79, 147–68.

_____. (2009). "Two Step Series Estimation of Sample Selection Models." *The Econometrics Journal*, 12, S217–S229.

Newey, W., and K. West. (1987). "A Simple Positive Semi-Definite, Heteroscedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica*, 50, 703–8.

Neyman, J. (1949). "Contribution to the Theory of the $\chi^2$ Test," in *Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probabilities*, edited by J. Neyman, pp. 230–70. Berkeley: University of California Press.

Neyman, J., and E.L. Scott. (1948). "Consistent Estimates Based on Partially Consistent Observations." *Econometrica*, 16, 1–32.

Ng, S. (2008). "A Simple Test for Nonstationarity in Mixed Panels." *Journal of Business and Economic Statistics*, 26, 113–26.

Nicholls, D.F., and B.G. Quinn. (1982). *Random Coefficient Autoregressive Models: An Introduction*. Berlin: Springer-Verlag.

Nickell, S. (1979). "Estimating the Probability of Leaving Unemployment." *Econometrica*, 47: 1249–66.

_____. (1981). "Biases in Dynamic Models with Fixed Effects." *Econometrica*, 49, 1399–416.

Nijman, T.H.E., and M. Verbeek. (1992). "Nonresponse in Panel Data: The Impact on Estimates of a Life Cycle Consumption Function." *Journal of Applied Econometrics*, 7, 243–57.

Nijman, T.H.E., M. Verbeek, and A. van Soest. (1991). "The Efficiency of Rotating Panel Designs in an Analysis of Variance Model." *Journal of Econometrics*, 49, 373–99.

Okui, R. (2009). "The Optimal Choice of Moments in Dynamic Panel Data Models." *Journal of Econometrics*, 151, 1–16.

Ord, J.K. (1975). "Estimation Methods for Models of Spatial Interaction." *Journal of the American Statistical Association*, 70, 120–26.

Pagan, A. (1980). "Some Identification and Estimation Results for Regression Models with Stochastically Varying Coefficients." *Journal of Econometrics*, 13, 341–64.

Pakes, A., and Z. Griliches. (1984). "Estimating Distributed Lags in Short Panels with an Application to the Specification of Depreciation Patterns and Capital Stock Constructs." *Review of Economic Studies*, 51, 243–62.

Pakes, A., and D. Pollard. (1989). "Simulation and the Asymptotics of Optimization Estimators." *Econometrica*, 57, 1027–1057.

Pedroni, P. (1995). "Panel Cointegration: Asymptotic and Finite Sample Properties of Pooled Time Series Tests with an Application to the PPP Hypothesis," Indiana University Working Paper in Economics, No. 95-013.

_____. (2004). "Panel Cointegration: Asymptotic and Finite Sample Properties of Pooled Time Series Tests with an Application to the PPP Hypothesis." *Econometric Theory*, 20, 597–625.

Peracchi, F. (2000). "The European Community Household Panel: A Review." Paper presented at the Panel Data Conference in Geneva.

Perron, P. (1989). "The Great Crash, the Oil Price Shock and the Unit Root Hypothesis." *Econometrica*, 57, 1361–401.

Pesaran, M.H. (2003). "On Aggregation of Linear Dynamic Models: An Application to Life-Cycle Consumption Models Under Habit Formation." *Economic Modeling*, 20, 227–435.

_____. (2004). "General Diagnostic Tests for Cross-Section Dependence in Panels."

_____. (2006). "Estimation and Inference in Large Heterogeneous Panels with Cross-Section Dependence." *Econometrica*, 74, 967–1012.

_____. (2007). "A Simple Panel Unit Root Test in the Presence of Cross-Section Dependence." *Journal of Applied Econometrics*, 22, 265–312.

_____. (2012). "On the Interpretation of Panel Unit Root Tests." *Economics Letters*, 116, 545–46.

Pesaran, M.H., and R. Smith. (1995). "Estimation of Long-Run Relationships from Dynamic Heterogeneous Panels." *Journal of Econometrics*, 68, 79–114.

Pesaran, M.H., and E. Tosetti. (2010). "Large Panels with Common Factors and Spatial Correlations." *Journal of Econometrics*, 161, 182–202.

Pesaran, M.H., and T. Yamagata. (2008). "Testing Slope Homogeneity in Large Panels." *Journal of Econometrica*, 142, 50–93.

Pesaran, M.H., and Z. Zhao. (1999). "Bias Reduction in Estimating Long-Run Relationships from Dynamic Heterogeneous Panels," in *Analysis of Panels and Limited Dependent Variables*, edited by C. Hsiao, K. Lahiri, L.F. Lee, and M.H. Pesaran, pp. 297–322. Cambridge: Cambridge University Press.

Pesaran, M.H., T. Schuermann, and S.M. Weiner. (2004). "Modelling Regional Interdependencies Using a Global Error-Correction Macroeconometrics Model." *Journal of Business and Economic Statistics*, 22, 129–62.

Pesaran, M.H., Y. Shin, and R.J. Smith. (1999). "Pooled Mean Group Estimation of Dynamic Heterogeneous Panels." *Journal of the American Statistical Association*, 94, 621–34.

_____. (2000). "Structural Analysis of Vector Error Correction Models with Exogenous I(1) Variables." *Journal of Econometrics*, 97, 293–343.

Pesaran, M.H., L.V. Smith, and T. Yamagata. (2013). "Panel Unit Root Tests in the Presence of a Multifactor Error Structure." *Journal of Econometrics*, 175, 94–115.

Pesaran, M.H., A. Ullah, and T. Yamagata. (2008). "A Bias Adjusted LM Test of Error Cross-Section Independence." *Econometrics Journal*, 11, 105–27.

Phelps, E. (1972). *Inflation Policy and Unemployment Theory: The Cost Benefit Approach to Monetary Planning*. London: Macmillan.

Phillips, P.C.B. (1986). "Understanding Spurious Regressions in Econometrics." *Journal of Econometrics*, 33, 311–40.

Phillips, P.C.B. (1991). "Optimal Inference in Cointegrated Systems." *Econometrica*, 59, 283–306.

Phillips, P.C.B., and S.N. Durlauf. (1986). "Multiple Time Series Regression with Integrated Processes." *Review of Economic Studies*, 53, 473–95.

Phillips, P.C.B., and B. E. Hansen. (1990). "Statistical Inference in Instrumental Variables Regression with I(1) Processes." *Review of Economic Studies*, 57, 99–125.

Phillips, P.C.B., and H.R. Moon. (1999). "Linear Regression Limit Theory for Nonstationary Panel Data." *Econometrica*, 67, 1057–111.

_____. (2000). "Nonstationary Panel Data Analysis: An Overview of Some Recent Developments." *Econometrics Review*, 19, 263–86.

Phillips, P.C.B. and D. Sul. (2007). "Bias in Dynamic Panel Estimation with Fixed Effects, Incidental Trends and Cross-Section Dependence." *Journal of Econometrics*, 137, 162–88.

Pinkse, J. (2000). "Asymptotic Properties of Moran and Related Tests and Testing for Spatial Correlation in Probit Models." Mimeo, Pennsylvania State University.

Powell, J.L. (1984). "Least Absolute Deviations Estimation for the Censored Regression Model." *Journal of Econometrics*, 25, 303–325.

_____. (1986). "Symmetrically Trimmed Least Squares Estimation for Tobit Models." *Econometrica*, 54, 1435–60.

Powell, J.L., J. Stock, and T. Stoker. (1989). "Semiparametric Estimation of Index Coefficients." *Econometrica*, 57, 1403–30?

Priestley, M.B. (1982). *Spectral Analysis and Time Series*, Vols. I and II. New York: Academic Press.

Prucha, I.R. (1983). "Maximum Likelihood and Instrumental Variable Estimation in Simultaneous Equation Systems with Error Components." Working Paper No. 83-6, Department of Economics, University of Maryland.

Quah, D. (1994). "Exploiting Cross-Section Variations for Unit Root Inference in Dynamic Data." *Economic Letters*, 44, 9–19.

Quandt, R.E. (1982). "Econometric Disequilibrium Models." *Econometric Reviews*, 1, 1–64.

Raj, B., and A. Ullah. (1981). *Econometrics: A Varying Coefficient Approach*. London: Croom Helm.

Rao, C.R. (1952). *Advanced Statistical Methods in Biometric Research*. New York: John Wiley & Sons.

_____. (1970). "Estimation of Heteroscedastic Variances in Linear Models." *Journal of the American Statistical Association*, 65, 161–72.

_____. (1972). "Estimation of Variance and Covariance Components in Linear Models." *Journal of the American Statistical Association*, 67, 112–15.

_____. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. New York: John Wiley & Sons.

Richard, J.F. (1996). "Simulation Techniques," in *The Econometrics of Panel Data*, 2nd ed. edited by L. Matyas and P. Sevestre, pp. 613–38. Dordrecht: Kluwer Academic.

Richard, J.F., and W. Zhang. (2007). "Efficient High-Dimensional Importance Sampling." *Journal of Econometrics*, 141, 1385–1411.

Ridder, G. (1990). "Attrition in Multi-Wave Panel Data," in Panel Data and Labor Market Studies, edited by J. Hartog, G. Ridder, and J. Theeuwes, pp. 45–79. Amsterdam: North-Holland.

_____. (1992). "An Empirical Evaluation of Some Models for Non-random Attrition in Panel Data." *Structural Change and Economic Dynamics*, 3, 337–35.

Robinson, P.M. (1988a). "Semiparametric Econometrics: A Survey." *Journal of Applied Econometrics*, 3, 35–51.

_____. (1988b). "Root-N-Consistent Semiparametric Regression." *Econometrica*, 56, 931–54.

_____. (1989). "Notes on Nonparametric and Semiparametric Estimation." Mimeo, London School of Economics.

Rosenberg, B. (1972). "The Estimation of Stationary Stochastic Regression Parameters Reexamined." *Journal of the American Statistical Association*, 67, 650–4.

———. (1973). "The Analysis of a Cross-Section of Time Series by Stochastically Convergent Parameter Regression." *Annals of Economic and Social Measurement*, 2, 39–428.

Rosenbaum, P. and D. Rubin. (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*, 70, 41–55.

Rothenberg, T.J. (1973). *Efficient Estimation with a Priori Information*. New Haven, CT: Yale University Press.

Rubin, D.B. (1976). "Inference and Missing Data." *Biometrica*, 63, 581–92.

Saikkonen, P. (1991). "Asymptotically Efficient Estimation of Cointegration Regressions." *Econometric Theory*, 7, 1–21.

Sant, D. (1977). "Generalized Least Squares Applied to Time-Varying Parameter Models." *Annals of Economic and Social Measurement*, 6, 301–14.

Sarafidis, V., and T. Wansbeek. (2012). "Cross-Sectional Dependence in Panel Data Analysis." *Econometric Reviews*, 31, 483–531.

Sarafidis, V., T. Yamagata, and D. Robertson. (2009). "A Test of Cross-Section Dependence for a Linear dynamic Panel Model with Regressors." *Journal of Econometrics*, 148, 149–61.

Sargan, J.D. (1958). "The Estimation of Economic Relationships Using Instrumental Variables." *Econometrica*, 26, 393–415.

Sargan, J.D., and A. Bhargava. (1983). "Testing for Residuals from Least Squares Regression Being Generated by Gaussian Random Walk." *Econometrica*, 51, 153–74.

Scheffé, H. (1959). *The Analysis of Variance*. New York: John Wiley & Sons.

Schmidt, P. (1984). "Simultaneous Equation Models with Fixed Effects." Mimeo, Michigan State University.

Schwarz, G. (1978). "Estimating the Dimension of a Model." *Annals of Statistics*, 6, 461–64.

Searle, S.R. (1971). *Linear Models*. New York: John Wiley & Sons.

Sevestre, P., and A. Trognon. (1982). "A Note on Autoregressive Error Component Models,"#8204, Ecole Nationale de la Statistique et de l'Administration Economique et Unite de Recherche.

Sheiner, L., B. Rosenberg, and K. Melmon. (1972). "Modeling of Individual Pharmacokinetics for Computer-Aided Drug Dosage." *Computers and Biomedical Research*, 5, 441–59.

Shen, X. (1997). "On Methods of Sieves and Penalization." *Annals of Statistics*, 25, 2555–91.

Shiller, R., and P. Perron. (1985). "Testing the Random Walk Hypothesis: Power versus Frequency of Observation." *Economic Letters*, 18, 381–86.

Sims, C. (1980). "Macroeconomic and Reality." *Econometrica*, 48, 1–48.

Sims, C., J.H. Stock, and M.W. Watson. (1990). "Inference in Linear Time Series Models with Some Unit Roots." *Econometrica*, 58(1), 113–44.

Singer, B., and S. Spilerman. (1974). "Social Mobility Models for Heterogeneous Populations," in *Sociological Methodology 1973–1974*, edited by H.L. Costner, pp. 356–401. San Francisco: Jossey-Bass.

———. (1976). "Some Methodological Issues in the Analysis of Longitudinal Surveys." *Annals of Economic and Social Measurement*, 5, 447–74.

Singh, B., A.L. Nagar, N.K. Choudhry, and B. Raj. (1976). "On the Estimation of Structural Changes: A Generalization of the Random Coefficients Regression Model." *International Economic Review*, 17, 340–61.

Small, K., and C. Hsiao. (1985). "Multinominal Logit Specification Tests." *International Economic Review*, 26, 619–27.

Smith, A.F.M. (1973). "A General Bayesian Linear Model." *Journal of the Royal Statistical Society B*, 35, 67–75.

Solon, G. (1985). "Comment on 'Benefits and Limitations of Panel Data by C. Hsiao." *Econometric Reviews*, 4, 183–86.

Stiglitz, J.E., and A. Weiss. (1981). "Credit Rationing in Markets with Imperfect Information." *American Economic Review*, 71, 393–410.

Stock, J.H. (1987). "Asymptotic Properties of Least Squares Estimators of Cointegrating Vectors." *Econometrica*, 55, 1035–56.

Stock, J., and M.W. Watson. (2008). "Heteroskedasticity-Robust Standard Errors for Fixed Effects Regression." *Econometrica*, 76, 155–74.

Stoker, T.M. (1993). "Empirical Approaches to the Problem of Aggregation Over Individuals." *Journal of Economic Literature*, 31, 1827–74.

Stone, R. (1954). *The Measurement of Consumers' Expenditure and Behavior in the United Kingdom, 1920–38*. Cambridge: Cambridge University Press.

Stroud, A.H., and D. Secrest. (1966). *Gaussian Quadrature Formulas*. Englewood Cliffs, NJ: Prentice-Hall.

Su, L., and A. Ullah. (2011). "Nonparametric and Semiparametric Panel Econometric Models: Estimation and Testing," in *Handbook of Empirical Economics and Finance*, edited by A. Ullah and D.E.A. Giles, pp. 455–97. New York: Taylor and Francis Group.

Su, L., Z. Shi, and P.C.B. Phillips. (2013). "Identifying Latent Structures in Panel Data." Mimeo, Singapore Management University.

Su, L., A. Ullah, and Y. Wang. (2010). "A Note on Nonparametric Regression Estimation with General Parametric Error Covariance," in Working Paper, School of Economics, Singapore Management University, Singapore.

Summers, L.H. (1981). "Taxation and Corporate Investment: A $q$-theory Approach." *Brookings Papers on Economic Activity*, 1, 67–127.

Swamy, P.A.V.B. (1970). "Efficient Inference in a Random Coefficient Regression Model." *Econometrica*, 38, 311–23.

_____. (1971). *Statistical Inference in Random Coefficient Regression Models*. Berlin: Springer-Verlag.

_____. (1974). "Linear Models with Random Coefficients," in *Frontiers in Econometrics*, edited by P. Zarembka, pp. 143–68. New York: Academic Press.

Swamy, P.A.V.B., and J.S. Mehta. (1973). "Bayesian Analysis of Error Components Regression Models." *Journal of the American Statistical Association*, 68, 648–58.

_____. (1977). "Estimation of Linear Models with Time and Cross-Sectionally Varying Coefficients." *Journal of the American Statistical Association*, 72, 890–98.

Swamy, P.A.V.B., and P.A. Tinsley. (1977). "Linear Prediction and Estimation Method for Regression Models with Stationary Stochastic Coefficients." Special Studies Paper No. 78, Federal Reserve Board Division of Research and Statistics, Washington, D.C.

Taub, A.J. (1979). "Prediction in the Context of the Variance-Components Model." *Journal of Econometrics*, 10, 103–7.

Taylor, W.E. (1980). "Small Sample Consideration in Estimation from Panel Data." *Journal of Econometrics*, 13, 203–23.

Temple, J. (1999). "The New Growth Evidence." *Journal of Economic Literature*, 37(1), 112–56.

Theil, H. (1954). *Linear Aggregation of Economic Relations*. Amsterdam: North-Holland.

_____. (1958). *Economic Forecasts and Policy*. Amsterdam: North-Holland.

_____. (1971). *Principles of Econometrics*. New York: John Wiley & Sons.

Theil, H., and L.B.M. Mennes. (1959). "Conception Stochastique de Coefficients Multipli-cateurs dans l'Adjustment Lineaire des Series Temporelles." *Publications de l'Institut de Statistique de l'Universite de Paris*, 8, 211–27.

Tibshirani, R.J. (1996). "Regression Shrinkage and Selection via the LASSO." *Journal of the Royal Statistical Society B*, 58, 267–88.

Tobin, J. (1950). "A Statistical Demand Function for Food in the U.S.A." *Journal of the Royal Statistical Society A*, 113, 113–41.

_____. (1958). "Estimation of Relationships for Limited Dependent Variables." *Econometrica*, 26, 24–36.

_____. (1969). "A General Equilibrium Approach to Monetary Policy." *Journal of Money, Credit and Banking*, 1, 15–29.

Trivedi, P.K. (1985). "Distributed Lags, Aggregation and Compounding: Some Econometric Implciations." *Review of Economic Studies*, 52, 19–35.

Trivedi, P., and M.K. Munkin. (2011). "Recent Developments in Cross-Section and Panel Count Models," in *Handbook of Empirical Economics and Finance*, edited by A. Ullah and D.E.A. Giles, pp. 87–132. New York: Taylor and Francis Group.

Trognon, A. (1978). "Miscellaneous Asymptotic Properties of Ordinary Least Squares and Maximum Likelihood Estimators in Dynamic Error Components Models." *Annales de L'INSEE*, 30–31, 631–57.

_____. (2000). "Panel Data Econometrics: A Successful Past and a Promising Future." Paper presented at the 2000 Panel Data Conference in Geneva.

Tsiatis, A.A. (1981). "A Large Sample Study of Cox's Regression Model." *The Annals of Statistics*, 9, 93–108.

Vella, F., and M. Verbeek. (1999). "Two-Step Estimation of Panel Data Models with Censored Endogenous Variables and Selection Bias." *Journal of Econometrics*, 90, 239–64.

Verbeek, M. (1992). "The Design of Panel Surveys and the Treatment of Missing Observations." Ph.D. dissertation, Tilburg University.

_____. (2007). "Pseudo-Panels and Repeated Cross-Sections," in *The Econometrics of Panel Data*, 3rd ed., edited by L. Matyas and P. Severstre. Berlin: Springer-Verlag, 369–84.

Verbeek, M., and T.H.E. Nijman. (1996). "Incomplete Panels and Selection Bias," in the *Econometrics of Panel Data*, 2nd ed., edited by L. Matyas and P. Sevester, pp. 449–90. Dordercht: Kluwer Academic.

Verbeek, M., and F. Vella. (2005). "Estimating Dynamic Models from Repeated Cross-Sections." *Journal of Econometrics*, 127, 83–102.

Vogelsang, T. (2012). "Heteroscedasticity, Autocorrelation and Spatial Correlation Robust Inference in Linear Panel Models with Fixed Effects." *Journal of Econometrics*, 166, 303–19.

Wachter, M.L. (1970). "Relative Wage Equations for U.S. Manufacturing Industries 1947–1967." *Review of Economics and Statistics*, 52, 405–10.

Wallace, T.D., and A. Hussain. (1969). "The Use of Error Components Models in Combining Cross-Section with Time Series Data." *Econometrica*, 37, 55–72.

Wansbeek, T.J. (2001). "GMM Estimation in Panel Data Models with Measurement Error." *Journal of Econometrics*, 104, 259–68.

———. (1982). "A Class of Decompositions of the Variance-Covariance Matrix of a Generalized Error Components Model." *Econometrica*, 50, 713–24.

Wansbeek, T.J., and P.A. Bekker. (1996). "On IV, GMM and ML in a Dynamic Panel Data Model." *Economic Letters*, 51, 145–52.

Wansbeek, T.J., and A. Kapteyn. (1978). "The Separation of Individual Variation and Systematic Change in the Analysis of Panel Data." *Annales de l'INSEE*, 30–31, 659–80.

Wansbeek, T.J., and R.H. Koning. (1989). "Measurement Error and Panel Data." *Statistica Neerlandica*, 45, 85–92.

Wansbeek, T.J., and E. Meijer. (2000). *Measurement Error and Latent Variables in Econometrics*. Amsterdam: Elsevier.

———. (2007). "Comment on Panel Data Analysis – Advantages and Challenges." TEST, 16, 33–36.

Westerlund, J. (2005). "New Simple Tests for Panel Cointegration." *Econometric Reviews*, 24, 297–316.

Westerlund, J., and R. Larsson. (2012). "Testing for a Unit Root in Random Coefficient Panel Data Model." *Journal of Econometrics*, 167, 254–73.

Westerlund, J., and J.P. Urbain. (2012). "Cross-Sectional Averages or Principal Components?," Mimeo, Maastricht University.

White, H. (1980). "A Heteroscedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity." *Econometrica*, 48, 817–38.

———. (1982). "Maximum Likelihood Estimation of Misspecified Models." *Econometrica*, 50, 1–25.

Windmeijer, F. (2008). "GMM for Panel Data Count Models," in *The Econometrics of Panel Data*, 3rd ed., edited by L. Mátyás and P. Sevestre, pp. 603–24. Berlin: Springer-Verlag.

Wooldridge, J.M. (1999). "Distribution-Free Estimation of Some Nonlinear Panel Data Models." *Journal of Econometrics*, 90, 77–98.

Wright, B.D., and G. Douglas. (1976). "Better Procedures for Sample-Free Item Analysis," Research Memorandum 20, Statistical Laboratory, Department of Education, University of Chicago.

Yatchew, A. (1998). "Nonparametric Regression Techniques in Economics." *Journal of Economic Literature*, 36, 669–721.

Yu, J., and L.F. Lee. (2010). "Estimation of Unit Root Spatial Dynamic Panel Data Models." *Econometric Theory*, 26, 1332–62.

Yu, J., R. de Jong, and L.F. Lee. (2012). "Estimation for Spatial Dynamic Panel Data with Fixed Effects: The Case of Spatial Cointegration." *Journal of Econometrics*, 167, 16–37.

Zacks, S. (1971). *The Theory of Statistical Inference*. New York: John Wiley & Sons.

Zellner, A. (1962). "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias." *Journal of the American Statistical Association*, 57, 348–68.

———. (1966). "On the Aggregation Problem: A New Approach to a Troublesome Problem," in *Economic Models, Estimation and Risk Programming: Essays in Honor of Gerhard Tintner*, edited by K. Fox, pp. 365–74. Berlin: Springer-Verlag.

———. (1970). "Estimation of Regression Relationships Containing Unobservable Variables." *International Economic Review*, 11, 441–54.

Zellner, A., and H. Theil. (1962). "Three Stage Least Squares: Simultaneous Estimation of Simultaneous Equations,"*Econometrica*, 30, 54–78.

Zellner, A., C. Hong, and C.K. Min. (1991). "Forecasting Turning Points in International Output Growth Rates Using Bayesian Exponentially Weighted Autoregression, Time Varying Parameter and Pooling Techniques." *Journal of Econometrics*, 49, 275–304.

Ziliak, J.P. (1997). "Efficient Estimation with Panel Data When Instruments are Predetermined: An Empirical Comparison of Moment-Condition Estimators." *Journal of Business and Economic Statistics*, 15, 419–31.

# Author Index


Abramowitz, M., 245
Ahn, H., 290, 312
Ahn, S. C., 58, 102–103, 125n22, 341
Ai, C., 463
Aigner, D. J., 9, 404, 456
Akaike, H., 204, 363
Akashi, K., 402
Allison, P., 466
Alvarez, J., 13, 106, 117, 133, 271, 472
Amemiya, T., 44, 97, 99, 116, 170, 193, 195, 230, 232–233, 281n1, 283, 287, 292, 422n11
Andersen, E. B., 239, 312
Anderson, T. W., 9, 13, 66, 82n2, 84, 94, 99, 184, 209–210, 338–339, 377, 398, 402, 425n14, 469
Angrist, J. D., 352n9
Anselin, L., 13, 329, 335, 345
Antweiler, W., 453
Appelbe, T. W., 8, 201
Arellano, M., xix, 10, 13, 68, 99–106, 117, 125, 133, 207n19, 243, 261n17, 271, 318, 401–402, 471–472
Ashenfelter, O., 2, 81, 456
Avery, R. B., 140n4, 141, 144

Bai, J., 13, 337–340, 342, 392, 394, 398n8, 472
Balestra, P., 10, 81, 108–111, 198, 404
Baltagi, B. H., 47n8, 140n4, 141, 144, 147, 151, 201, 335, 395, 453–454, 473
Banerjee, A., 395
Barro, R., 8
Barth, J., 179–180
Bartolucci, F., 277, 280
Bassett, G., Jr., 445
Bates, G., 250
Becketti, S., 2
Beckwith, N., 2n1
Bekker, P. A., 103
Ben-Porath, Y., 5
Berkson, J., 233
Bernard, A., 386
Bhargava, A., 92, 93n8, 95, 106–107, 108n17, 109t, 111n18, 120, 394
Binder, M., 9, 128, 374, 396
Biørn, E., 9, 404, 458, 460
Bishop, Y. M. M., 262n18
Blanchard, P., xx
Blundell, R., 97, 102n14, 117n19, 125n22, 409
Bodkin, R. G., 139, 199, 370n2
Bond, S., 99, 101, 102nn13,14, 105, 117n19, 125n22, 212, 401
Bonhomme, S., 243, 471
Borus, M. E., 2
Bover, O., 102–103, 318
Box, G. E. P., 193, 345, 379, 426
Breitung, J., 328, 395
Bresson, G., 170
Breusch, T. S., 47n8, 99, 176–177, 185, 192, 193n14, 204, 345
Browning, M., 409
Burridge, P., 344
Butler, J., 245

Caines, P. E., 82n2
Canova, F., 167

Card, D., 220
Carrasco, R., 261n17
Carrion-i-Silvestre, J. L., 394
Carro, J. M., 273
Case, A. C., 333
Chamberlain, G., 11, 13, 53, 69–75, 92n7, 95–96, 140, 145, 152–153, 155, 162, 164, 172n3, 207nn18,19, 240, 244, 246, 248n14, 257, 263, 268, 312, 457
Chan, M. W. Luke, 8
Charlier, E., 313–314, 315t–316t
Chen, S., 291–292, 298n5, 462
Chesher, A. D., 12, 177
Chetty, V. K., 415
Chiang, M. H., 73n21, 381–382
Ching, H. S., 361, 363–364
Chintagunta, P., 250, 270
Choi, I., 390, 393–395
Chou, S., 357
Chow, G. C., 170, 186n13, 187, 190
Chu, J., 9, 387
Chue, T. K., 394
Chui, C. K., 462
Coleman, J. S., 251
Collado, M. D., 407
Conley, T. G., 135, 330
Cooley, T. F., 187
Cornwell, C., 239n10
Cosslett, S. R., 274
Cox, D. R., 232, 277, 433
Crépon, B., 70, 75n26, 444
Cripps, T., 262

Damrongplasit, K., 354, 361n14
Das, S., 328
Davis, P., 453
Deaton, A., 409, 456
De Finneti, B., 201
Dehejia, R. H., 354, 357
Dhrymes, P., 419, 421
Dickey, D. A., 9, 386, 389–390, 392, 469
Dielman, T., 2n1
Dineen, C. R., 8, 201
Donald, P., 104
Douglas, G., 242n11
Driscoll, J., 331
Duan, J. C., 435–437
Duguet, E., 444
Duncan, G. M., 162
Durbin, J., 66, 214t, 345, 415
Durlauf, S. N., 8–9, 167, 386, 388, 469

Eicker, F., 313
Engle, R. F., 134, 370n2

Fan, J., 358
Fan, K. T., 224
Fazzari, S. M., 212
Ferguson, T. S., 233
Fienberg, S. E., 262n18
Fisher, N. I., 390
Fisher, R. A., 331
Fitzgerald, J., 297
Florens, J. P., 434
Fougére, D., 434
Frankel, J. A., 386
Freeman, R. B., 5
Friedman, M., 205, 470
Frisch, R., 342
Fujiki, H., 8, 218–219, 468
Fuller, W. A., 9, 97, 386, 389–390, 392, 469

Galvao, A. F., 447
Gardiol, L., 39n5
Geweke, J., 448
Gijbels, I., 358
Girma, S., 410
Goldberger, A. S., 152
Goodman, L. A., 252
Goodrich, R. L., 82n2
Gorseline, D. E., 162
Gottschalk, S., 297
Gourieroux, C., 299n6, 448, 452
Granger, C. W. J., 8, 134, 370n2, 468
Grassetti, L., 399
Greenaway-McGrevy, R., 339n6
Griffith, D. A., 329
Griffiths, W. E., 175, 185
Griliches, Z., 9, 137, 140, 152–155, 162, 164, 219, 292, 296, 420, 422, 425, 429, 443, 456–457, 481
Groen, J. J. J., 385, 397
Gronau, R., 287
Grunfeld, Y., 168, 175, 219
Gurmu, S., 444

Hadri, K., 387
Hahn, J., 13, 84, 104, 117, 123–124, 128, 133, 271, 273
Hajivassilou, V., 448
Hall, B., 292, 443
Hall, P., 331
Han, A., 434
Han, C., 114, 339n6
Hansen, B., 101, 348, 381
Härdle, W., 289
Harris, R. D. F., 444
Hartley, H. O., 182
Hausman, J. A., 9, 11, 14, 48, 57–58, 61, 81, 104–105, 121, 149, 204, 235, 236n4, 292, 295–297, 314, 408, 434, 443, 456–457, 473
Hayashi, F., 213
Heckman, J. J., 13–14, 220–221, 227, 242, 243t, 245–246, 251, 254–255, 256t, 261, 263–265, 267t, 268t, 283–284, 287–288, 292, 294, 299, 311, 352n9, 354, 360n13, 409, 473
Henderson, C. J., Jr., 63
Hendricks, W., 193
Hildreth, C., 184
Hirano, K., 298, 473
Hoch, I., 32–33, 34t
Hoffman, B., 331
Ho-Illman, K. F., 464
Holland, P. W., 262n18
Holly, A., 39n5
Hong, C., 186,
Hong, Y., 66n17, 463
Honoré, B., 13, 207n19, 256, 258–260, 269–270, 274, 276–277, 299, 304–306, 309–310, 317n9, 318, 360
Hood, Wm. C., 139, 397
Horowitz, J. L., 248–249
Hotellings, H., 472
Hotz, 360n13
Houck, J. P., 184
Hsiao, C., xx, 3, 8–9, 13–14, 30n6, 47, 50, 66, 76, 84, 92t, 93–94, 97, 99, 112, 114–115, 117–118, 119t, 120t, 122–128, 139, 170, 174, 184n10, 186, 197, 199, 201–220, 223–224, 227, 235, 284, 328, 349–350, 354, 357, 360–370, 374, 377, 398–402, 412, 415, 448, 452, 458, 460, 464, 468
Hu, L., 323–324
Hubbard, R. G., 212
Hurvich, C. M., 363
Hurwicz, L., 169
Hussain, A., 63–64, 198
Hyslop, D., 251

Ichimura, H., 354
Im, K. S., 9, 389–392, 394
Imbens, G. W., 298
Inoue, A., 410
Intriligator, M. D., 139, 370n2
Izan, H. Y., 417n6

Jang, S. J., 466
Janz, N., 3
Jayet, H., 329
Jenkins, G. M., 193, 379, 426
Jeong, K. J., 415
Johansen, S., 370, 383, 396
Johnson, P., 167
Johnston, J., 26n3
Jones, C., 386
Jorgenson, D. W., 212
Judge, G., 170, 174
Judson, R. A., 103
Jung, B., 453
Juster, T., 2

Kalman, R. E., 188
Kao, C., 66n17, 135, 381–382, 395, 460
Kapetanios, G., 342
Kapoor, M., 334–336
Kapteyn, A., 9, 39n5, 335, 453–454, 458
Karlin, S., 253
Kato, K., 447
Keane, M. P., 299n6, 448
Kelejian, H. H., 170, 185, 333–334, 345
Kiefer, N. M., 67, 81, 274
Kiviet, J. F., 103, 210
Kleibergen, F., 385, 396–397
Klein, L. R., 169, 205
Klein, R., 235, 289
Klette, T. J., 460
Koenker, R., 103, 193, 445, 447
Koning, R. H., 9, 457–458
Koopmans, T. C., 139, 397
Kraay, A., 331
Kraft, A., 179

Kraft, J., 179
Krishnakumar, J., xx, 460
Kruiniger, H., 114, 118
Kuersteiner, G., 13, 84, 104
Kuh, E., 10, 13, 18, 24–26, 27t, 29t, 167–168, 198, 213, 412–413
Kunitomo, N., 402
Kwiatkowski, D., 387
Kyriazidou, P. E., 13, 250, 256, 258–260, 269–270, 277, 299, 311–317, 324–326, 360, 408

Labeaga, J., 318
Lahiri, K., xx
LaLonde, R., 354, 357
Lamarche, C., 447
Lancaster, T., 12, 177, 292n3
Larsson, R., 387, 390–391, 396
Layton, L., 251
Lee, J., 394
Lee, L. F., xx, 13, 175, 235, 333–334, 336
Lee, M. J., 249–250, 312
Lee, Y. H., 341
Le Gallo, J., 329
Levin, A., 9, 387–390
Lewbel, A., 8, 217, 468
Li, M., 220
Li, Q., 66n17, 201n16, 224, 461, 463
Liang, Z., 224
Liesenfeld, L., 449
Lillard, L., 65–66
Lin, C., 9, 171, 387–390
Lindley, D. V., 175, 209
Little, R. J. A., 297
Liu, L. M., 357, 369

Machado, J. A. F., 103
MaCurdy, T. E., 7, 67, 99
Maddala, G. S., 41, 44, 47, 80, 230, 281n1, 292, 390, 393, 412, 414–416
Magnus, J. R., 373n3, 384, 458
Mairesse, J., 26n5, 70, 75n26, 473
Malinvaud, E., 6, 73, 418n8, 419
Mankiw, N. G., 8
Manski, C. F., 13, 235, 247–249, 256, 259, 276, 312
Mao, G., 350, 351t
Martins-Filho, C., 461
Matsumoto, T., 357
Matyas, L., xx
Mazodier, P., 39n5
McCoskey, S., 135, 395
McFadden, D., 230, 234–235, 240, 448, 451
McKenzie, D. J., 410
Medoff, J. L., 5
Meghir, C., 212, 409
Mehta, J. S., 168n1, 186
Meijer, E., 4, 460
Melenberg, B., 314
Melmon, K., 2n1
Mennes, L. B. M., 169
Meyer, J. R., 26, 412–413
Min, C. K., 186, 205
Mizon, G. E., 99
Moffitt, R., 245, 297, 407, 409–410
Monfort, A., 299n6, 448, 452
Montes-Rojas, G. V., 447
Moon, H. R., 9, 13, 58, 123–124, 128, 131–132, 134, 206n17, 392–393, 472
Morimune, K., xx
Mouchart, M., 434
Mount, T. D., 44, 47n8
Mountain, D. C., 8, 464
Mundlak, Y., 11, 32, 50–54, 56, 70, 80–81, 84, 92n7, 172n3, 178, 226–227, 404n1
Munkin, M. K., 444

Nagar, A. L., 103
Nantell, T., 2n1
Narasimham, G. V. L., 168n1
Nerlove, M., xix, 6, 10, 32, 40, 63, 77, 79, 81, 84, 86, 94, 108, 110t, 111, 198, 242
Neudecker, H., 373n3, 384, 458
Newey, W., 104, 224, 271, 289, 299, 331, 363, 389, 462
Neyman, J., 55, 65, 112, 233, 236, 238, 241, 250, 429n15, 471
Ng, S., 171, 337–339, 392, 398n8
Nickell, S., 84
Nielson, A. C., 266
Nigro, V., 277, 280
Nijman, T. H. E., 297, 299, 404

Okui, R., 103
Ord, J. K., 329, 332
Owen, A. L., 103

Pagan, A. R., 176–177, 185, 187–188, 192, 193n14, 204, 345
Pakes, A., 418, 420, 422, 425, 429, 448
Pedroni, P., 395
Peracchi, F., 3
Perktold, J., 250, 270
Perron, P., 392–393
Pesaran, M. H., xx, 8–9, 13, 93, 97, 112, 114–115, 118t–120t, 122t, 126, 128, 135, 170, 174, 178, 206n17, 208–210, 211t, 217, 220, 223, 262n18, 328nn1,2, 339, 342–343, 345–346, 349–350, 370, 372, 374, 377–378, 386, 389, 392–395
Petersen, B., C., 212
Phelps, E., 261–262
Phillips, G. D. A., 192
Phillips, P. C. B., 9, 13, 114, 131–133, 171, 206n17, 210, 328, 370, 377, 381, 386, 388, 395, 469, 472
Picks, A., 349
Poirier, D. J., 193
Pollard, D., 448
Powell, J. L., xx, 235, 284–285, 287, 290–291, 302n7, 312
Prescott, E. C., 187
Priestley, M. B., 331
Prucha, I. R., 142n6, 333–334, 345

Quah, D., 167
Quandt, R. E., 296n4

Racine, J. S., 461
Raj, B., 170
Rao, C. R., 43, 71–72, 75n27, 173n4, 176, 182, 184
Richard, J. F., 299n6, 448–449
Ridder, G., 297–299, 473
Robb, R., 354, 409
Robertson, D., 346–348
Robinson, P. M., 289, 291, 299
Romer, D., 8
Ronchetti, E., xx
Rose, A. K., 386
Rosenbaum, B., 227, 354–355, 357
Rosenberg, B., 2n1, 187
Rothenberg, T. J., 74
Rubin, D. B., 227, 297–298, 354–355, 357, 398, 402

Saikkonen, P., 382
Sala-i-Martin, X., 8
Sarafidis, V., 346–348
Saranadasa, H., 472
Sargan, J. D., 92, 93n8, 95, 105–107, 108n17, 109t, 111n18, 120, 346, 348, 394
Schmidt, P., 99, 102, 125n22, 138n2, 239n10, 341
Schmierer, D. A., 220, 227
Schnell, J. F., 460
Schuermann, T., 378
Schwarz, G., 204, 363
Scott, E. L., 55, 65, 112, 236, 238, 241, 471
Searle, S. R., 18
Secrest, D., 245
Sevestre, P., xx, 86n4, 97, 111n18
Sheiner, L., 2n1
Shen, X., 462
Shen, Y., 8, 216t, 218–219, 350, 351t, 360, 468
Shi, Z., 171
Shin, Y., 9, 206n17, 370, 372, 378, 389–392, 394
Silverstein, J. W., 472
Sims, C., 369–370
Singer, B., 252, 292n3
Singh, B., 193
Small, K., 235
Smith, A. F. M., 174–175, 209
Smith, L. V., 393
Smith, R., 206n17, 217, 344, 370, 372, 378
Smith, R. J., 97
Solon, G., 2, 456
Song, S., 453, 473
Spady, R., 235, 289
Spilerman, S., 252
Stegun, J., 245
Stengos, T., 461
Stephan, S. W., 64, 185

Stock, J. H., 69, 235, 370
Stoker, T. M., 217, 235
Stone, R., 412
Stroud, A. H., 245
Su, L., 171, 461, 463
Sul, D., 328, 339n6
Summers, L. H., 213
Sun, B. H., 20–26, 50, 204, 436t
Swamy, P. A. V. B., 168, 172–175, 177–178, 180, 186, 193, 198, 210n20, 420

Tahmiscioglu, A. K., 30n6, 93, 97, 112, 114–115, 118, 119t–120t, 123–124, 126–128, 174, 208–213, 214t–216t, 223, 328, 377
Tamer, E., 274, 276
Tarling, R., 262
Taylor, G., 458
Taylor, H., 253
Taylor, W. E., 44, 61, 149
Temple, J., 8
Theil, H., 37n3, 43–44, 52, 59n15, 74, 95n10, 104, 122, 141, 160n12, 169, 175n6, 180, 217, 405
Thornberry, T. P., 466
Tibshirani, R. J., 171
Tieslau, M., 394
Tinsley, P. A., 193
Tiao, G. C., 53
Tobias, J. L., 220
Tobin, J., 213, 217, 281, 412, 416
Todd, P., 354
Tosetti, E., 135, 328nn1,2
Trivedi, P. K., 217, 444
Trognon, A., xix, 39n5, 64, 86, 94, 97, 111n18
Tsai, C. L., 363
Tsiatis, A. A., 434
Tsui, K. Y., 8

Ullah, A., 170, 461, 463
Urzua, S. S., 220, 227

van Soest, A., 314, 315t–316t, 404
Vella, F., 299, 411
Verbeek, M., 297, 299, 404, 411
Vogelsang, T., 224, 331
Vytlacil, E. J., 220, 227, 352n9

Wachter, M. L., 193
Wahba, S., 354, 357
Wallace, T. D., 63–64, 198
Walras, Léon, 369
Wan, S., 361, 363–364
Wang, B., 360
Wang, K. Q., 448, 452
Wang, L. Q., 452
Wang, T., 436–437, 439t, 441t
Wang, Y., 461
Wansbeek, T., 4, 9, 39n5, 103, 201, 335, 344, 453–454, 457–458, 460
Watson, M. W., 69, 345, 370
Weeks, G., 360
Weil, D., 8
Weiner, S. M., 378
Weiss, Y., 65
West, K., 224, 331, 363, 389
Westerlund, J., 390–391
White, H., 72n20, 177, 313, 463
Willis, R., 65–66, 251, 265
Windmeijer, F., 444
Wise, D., 14, 236n4, 295–297, 473
Wolfowitz, J., 274
Wooldridge, J. M., 299
Wright, R., 2n1, 242n11
Wu, S., 390, 393

Xie, W., 224, 227

Yamagata, T., 342, 346–348, 393
Yao, F., 461
Yu, J., 334, 336

Zacks, S., 239n9
Zellner, A., 152, 169–171, 175, 180, 186, 197, 205, 328, 377–378, 384, 386
Zhang, J., 84, 93, 117, 128, 401
Zhang, W., 448
Zhang, Y. T., 224
Zhao, X., 354, 444
Zhao, Z., 206n17
Zhou, Q., 399–400, 402
Ziliak, J. P., 103, 117

# Subject Index


AICC criterion, 363–368
Aitken estimators, 37n3, 148, 179, 185, 196
Almon lag, 419
analysis of covariance (ANCOVA): *F* statistic and, 21, 23; homogeneity and, 17–26; Kuh and, 24–26; linear regression models and, 17–26, 35, 37; parameter constancy and, 193; parameter number and, 18; random-coefficient model and, 177, 179, 186; residual sum of squares (RSS) and, 19–21, 22t; time series and, 24–26; within-group estimates and, 19
analysis of variance (ANOVA): fixed-coefficient model and, 182; linear regression models and, 17, 35; time and, 182
arbitrary error structure, 69–75
asymptotics: count data model and, 442–443; cross-sectionally dependent panel data and, 328, 333, 336–337, 340–345, 349; diagonal path limits and, 131; discrete data and, 233, 240, 242, 246n13, 249–250, 259, 266t, 271–273; duration model and, 434; dynamic models and, 81–86, 94, 96–99, 101, 103, 105–107, 111n18, 112–113, 115–121, 124, 126–135; dynamic systems and, 374–377, 381–382, 385–386, 388–392, 395–402; incomplete panel data and, 410, 414–415, 428–429; joint limits and, 131; multidimensional, 13, 469, 472; nonparametric panel data and, 462–463; normality and, 75–77, 82n2, 288; panel data advantages and, 9, 469; panel data issues and, 13, 472; panel quantile regression and, 447; sample truncation and, 283, 287–289, 295, 298–299, 305, 310, 313, 323, 326; sequential limits and, 130; simple regression and, 44, 47, 52, 55, 57–58, 65–68, 72–77; simulation methods and, 451–452; static simultaneous-equations models and, 136, 139, 143, 146–147, 150–152, 183n2; variable coefficient models and, 121, 174–178, 185–186, 193, 196, 201n16, 208, 223–225
attrition: bias and, 292–296; Gary income-maintenance project and, 296–298; Hausman-Wise two-period model of, 295, 297–298; incomplete panel data and, 408; panel data issues and, 13, 469, 472–473; probability and, 292–296; sample selection and, 292–298, 300t, 312, 472–473
augmented Dickey-Fuller (ADF) test, 389–390, 392, 395
autocorellation: cross-sectionally dependent panel data and, 331, 363; dynamic models and, 86, 94; heteroscedasticity and, 68–69; incomplete panel data and, 417n6; measurement errors and, 460; serially correlated errors and, 65–68; simple regression models and, 34, 64–69; variable coefficient models and, 214t, 217, 224

autoregression: cross-sectionally dependent panel data and, 329–332, 334, 336, 345, 363; discrete data and, 251; dynamic models and, 82n2, 86, 104, 118; dynamic systems and, 369–378, 390; heterogeneity and, 377–378; heteroscedastic autoregression (HAC) covariance matrix and, 331, 363; homogeneity and, 370–377; incomplete panel data and, 417n6, 418, 422, 425–426; panel vector models and, 370–378; sample selection and, 324; simple regression and, 66; simulation methods and, 450; variable coefficient models and, 187, 207n18; vector autoregressive (VAR) models and, 369–378, 386, 393n7, 397
autoregressive moving average (ARMA), 187
average treatment effect (ATE), 352–353

Bank of America, 438
baseline hazard function, 433
Bayesian methods, xix; incomplete panel data and, 412, 415n5; mode estimators and, 174; panel data issues and, 471; variable coefficient models and, 174, 198–201, 204–205, 208–212
best linear unbiased estimators (BLUE): simple regression models and, 36, 38–39, 41, 52, 60, 63; variable coefficient models and, 183–185, 195
bias: adjusted estimators and, 270–273; attrition and, 292–296; best linear unbiased estimators (BLUE) and, 36, 38–39, 41, 52, 60, 63, 183–185, 195; control groups and, 356; cross-sectionally dependent panel data and, 337, 340, 342, 344–346, 349, 354, 356, 360; discrete data and, 230, 242–243, 254–255, 264, 270–273, 280; dynamic models and, 81–86, 97, 99, 103–104, 108, 109t, 111–112, 117–119, 123, 128, 133, 270–273; dynamic systems and, 381, 388–389, 400–402; fixed-effects inference and, 56, 467; homogeneity and, 24; incomplete panel data and, 412, 417, 419; linear regression models and, 24, 33, 36, 38, 41, 50–52, 55–58, 61; ordinary least squares (OLS) estimators and, 85–86, 97; panel data advantages and, 7–8, 464, 467–468; panel data issues and, 11–14, 471–472; panel quantile regression and, 447; probability and, 292–296; random effects and, 56, 85–86, 467; reducing, 270, 467–468; sample truncation and, 282, 290, 292–297, 312, 314; selection, 14, 292–296, 312, 354, 360; simulation methods and, 450, 452, 455–457; state dependence and, 270–273; static simultaneous-equations models and, 136–137, 143, 152; variable coefficient models and, 173–175, 178, 180, 183, 184n10, 200, 210, 211t, 215, 220, 223
bounding parameters, 274–276
Box-Jenkins method, 67, 193, 363, 379
British Household Panel Survey (BHPS), 2
Brownian motion, 134, 394, 396
Bureau of the Census, 403–404

cell-mean correction, 19, 27t–28t
Center for Research in Security Prices (CRSP), 437
central limit theory, 135
Chamberlain-pi approach, 69–75
chi-square distribution, 9, 233, 262n18, 268t, 377, 391
Cobb-Douglas production function, 32
cointegrated panel models: common trends and, 379–380; estimation and, 381–386; heterogeneity and, 383–386; homogeneity and, 381–383; maximum-likelihood estimators (MLE) and, 382–384; minimum-distance estimators (MDE) and, 382–383; properties of, 379–380
cointegration tests: likelihood approach and, 395–397; residual-based, 394–395
common correlated effects heterogeneous model (CCEMG), 350, 351t
common correlated effects model (CCEP), 350, 351t
common trends, 379–380
Community Innovation Surveys (CIS), 3
conditional inference, 48–56, 372, 470

control groups: average difference and, 353; bias and, 356; confounding variables and, 353; cross-sectionally dependent panel data and, 353–357, 360–369; difference-in-difference method and, 360; frequency weight and, 357; mean difference and, 353; population sources and, 354; predicting counterfactuals and, 361–369; propensity score method and, 355–357, 360; sample selection and, 296; treatment groups and, 353–357, 360–361
correlated random-coefficient models: conventional fixed-effects estimators and, 223–224; group mean estimators and, 223; identification with cross-sectionally dependent panel data and, 221–222; introduction to, 220; mean effects estimation and, 223–227; panel pooled maximum-likelihood estimators and, 224–226; semiparametric estimates and, 227
count data model: asymptotics and, 442–443; density and, 442; dependent variables and, 444; exogenous variables and, 441; incidental parameters and, 443; individuals and, 441–443; infinity and, 442; log-likelihood function and, 441–443; maximum-likelihood estimators (MLE) and, 442–443; negative binomial distribution and, 442; Poisson process and, 439–440, 443–444; probability and, 438–439, 441t, 444; statistics and, 443
counterfactuals: control group information and, 361–369; cross-sectionally dependent panel data and, 361–369, 469; panel data advantages and, 469; prediction of, 361–369
covariance (CV) estimators: cross-sectionally dependent panel data and, 331; dynamic models and, 80–84, 92, 123, 132; fixed effects and, 47–56; heteroscedasticity and, 68–69; incomplete panel data and, 411; least-squares estimators and, 43; random effects and, 40–41; simple regression models and, 37–44, 47–57, 59–65, 67–69; static simultaneous-equations models and, 164
Cowles Commission, 139
Cramer-Rao bound, 57
cross-sectionally dependent panel data, xxi; adjustment methods and, 354–359; AICC criterion and, 363–368; arbitrary labeling and, 327; asymptotics and, 328, 333, 336–337, 340–345, 349; autocorellation and, 331, 363; autoregression and, 329–332, 334, 336, 345, 363; bias and, 337, 340, 342, 344–346, 349, 354, 356, 360; change inferences and, 5; common correlated effects heterogeneous model (CCEMG) model and, 350, 351t; common correlated effects model (CCEP) and, 350, 351t; control groups and, 353–357, 360–369; counterfactual prediction and, 361–369, 469; covariance (CV) estimators and, 331; degrees of freedom and, 345, 348–349; density and, 356; dependent variables and, 329, 348–350; difference-in-difference method and, 360–361; discrete-response models and, 230–235; dummy variable and, 343, 352; economic distance and, 331; efficiency and, 327; error-component model and, 334, 345; error terms and, 329, 336; example of, 350–351; exogenous variables and, 336; factor approach and, 327, 337–342; feasible generalized least-squares estimators (FGLS) and, 328, 335, 342; fixed constants (FE) and, 334–335; fixed effects and, 334, 340, 360; generalized method of moments (GMM) and, 341, 346, 348n8; group mean augmented approach and, 342–344; heterogeneity and, 350, 360; heteroscedasticity and, 331, 340–341, 363; homogeneity and, 350, 356; independence testing and, 344–351; individuals and, 327–331, 334–336, 339–340, 352–353, 357, 359–364; infinity and, 331, 333–334, 336, 363; instrumental

cross-sectionally dependent panel data (*cont.*)
variable (IV) estimators and, 333; issues of, 327–329; Lagrangians and, 344–345, 349; least-squares estimation and, 328–329, 333, 339–345, 364; linear model and, 344–348; local average treatment effect (LATE) and, 352n9; log-likelihood function and, 332–336; marginal treatment effect (MTE) and, 352n9; matching observables and, 355–357; matrices and, 327–331, 334, 336–340, 344, 350, 362; maximum-likelihood estimators (MLE) and, 354, 374–375, 383, 398, 400, 402; Monte Carlo studies and, 342, 346, 348, 350; neighbors and, 329–331, 333; nonparametric panel data and, 331, 355, 357, 359–368; normalization and, 329, 332, 337–338, 340; null hypotheses and, 344; omitted variables and, 327; ordinary least squares (OLS) estimation and, 364, 368; panel data issues and, 471–472; parametric models and, 331, 354, 357, 359–360; probability and, 339, 355, 357, 361; probit models and, 349; program evaluation and, 352–368; propensity score method and, 355–357, 360; random-coefficient models and, 221–222; random effects and, 335; regression discontinuity design and, 357–359; regression models and, 328–329, 339, 341, 350, 357–358; seemingly unrelated regression method (SUR) and, 328; simple regression models and, 328; spatial approach and, 329–337, 344; standard two-way effects models and, 327; statistics and, 327–328, 344–346, 348–350, 354–355, 364, 368; stochastics and, 347–348; SYR test and, 346–348; test of cross-sectional independence and, 344–351; time series and, 331, 339–340, 344–351, 362; time-specific variables and, 327, 331, 337; time-variant coefficients and, 180–186; Tobit models and, 348–349; treatment effects and, 352–368; variable coefficient models and, 170–186, 221–222; vector autoregressive (VAR) models and, 377–378
cross-sectionally independent panel data, 350, 377; unit root tests and, 387–391
cumulative distribution function: discrete data and, 245, 249, 253; duration model and, 430–431
Current Population Survey and Social Security Administration (CPS-SSA), 323–324

degrees of freedom: cross-sectionally dependent panel data and, 345, 348–349; dynamic models and, 105–107, 118; dynamic systems and, 386, 390, 396; homogeneity and, 18, 21, 23–25, 27t, 29t; incomplete panel data and, 414, 421, 428–429; linear regression models and, 18, 21, 23–25, 27t, 29t, 54–55, 57–58, 73, 77; panel data advantages and, 4, 464; panel data issues and, 470; sample truncation and, 314, 317; variable coefficient models and, 168, 174, 177–179, 186, 193, 209, 215, 223
density: conditional, 9–10, 54–55, 177–179, 191, 209, 239, 274, 294, 356, 448–449, 452; count data model and, 442; cross-sectionally dependent panel data and, 356; discrete data and, 231–232, 239, 245–246, 259–260, 274; dynamic models and, 107; incomplete panel data and, 404; normal density function and, 231, 246, 259, 289, 293; panel data advantages and, 9–10; panel data issues and, 10; panel quantile regression and, 445; sample truncation and, 282–284, 289–291, 296, 298, 302, 305, 312, 326; simple regression and, 39, 54–56; simulation methods and, 448–452; variable coefficient models and, 177–179, 191, 205, 209, 216
dependent variables: count data model and, 444; cross-sectionally dependent panel data and, 329, 348–350; discrete data and, 230, 233, 236, 250, 256, 259, 262, 270; dynamic models and, 80–81, 85–86, 97, 111–112, 115; dynamic systems and, 369, 376–377; incomplete panel data and, 412–414,

418; limited model and, 348–350; measurement errors and, 458; model of limited dependent variables and, 281; multilevel structures and, 453; nonparametric panel data and, 463; panel data advantages and, 467–468; panel data issues and, 14, 472; sample truncation and, 281–282, 284–285, 287, 290, 292, 295, 317, 325; simple regression and, 34, 39, 49t, 67; static simultaneous-equations models and, 136, 138, 144, 161, 163; variable coefficient models and, 175–176, 190, 215t, 216t
difference-in-difference method, 360–361
discrete data, xix; asymptotics and, 233, 240, 242, 246n13, 249–250, 259, 266t, 271–273; autoregression and, 251; bias and, 230, 242–243, 254–255, 264, 270–273, 280; bounding parameters and, 274–276; conditions for existence of consistent estimators and, 238–242; cumulative distribution function and, 245, 249, 253; density and, 231–232, 239, 245–246, 259–260, 274; dependent variables and, 230, 233, 236, 250, 256, 259, 262, 270; discrete-response models and, 230–235; distributed-lag models and, 263; dummy variable and, 239, 268; dynamic models and, 250–270; efficiency and, 233, 242, 245; equilibrium and, 252–254; error-component model and, 265; error terms and, 231, 235–236, 243–245, 251–252, 262–263; exogenous variables and, 242, 252, 254–255, 258, 265, 267, 270, 277, 471n1; fixed constants (FE) and, 251, 254, 270; fixed effects and, 236–243, 254–255, 256t, 270–272; Gaussian quadrature and, 245, 248; generalized method of moments (GMM) and, 246, 261n17; Hermite integration formula and, 245; heterogeneity and, 230, 235–246, 251–252, 261–271, 278; homogeneity and, 236, 253; incidental parameters and, 236–240, 244, 247–248, 271; independence of irrelevant alternatives and, 235; individuals and, 230–239, 242–243, 248–252, 254–255, 260–263, 265, 268–271, 277n21; infinity and, 238–240, 243, 246, 253–255, 259; initial conditions and, 252–255, 256t, 261; least-squares estimation and, 232–233, 239; linear-probability model and, 231–232, 238, 247; logit models and, 231–232, 234, 237–239, 241–242, 256, 259, 268, 271, 276–280, 471n1; log-likelihood function and, 237, 240–241, 243–244, 258, 263, 265, 266t, 272, 280; Markov process and, 251–253, 255, 256t, 265; matrices and, 232–233, 240–242, 244, 246, 250–255, 264, 266, 272; maximum-likelihood estimators (MLE) and, 233, 236–240, 245, 255, 263–264, 278; maximum score estimators and, 247–249; minimum-distance estimators and, 246; Monte Carlo studies and, 242, 254–255, 256t, 270, 273, 280; Newton-Raphson method and, 233, 280; normalization and, 235–236, 241, 246–250, 252, 270, 279; null hypotheses and, 255, 262n18; omitted variables and, 235, 243; parametric models and, 230–231, 235–250, 255; probability and, 231–232, 234–236, 238–257, 261–265, 268–275; probit models and, 231, 234, 236n4, 241–244, 254–255, 261n17, 264, 268t; random-coefficient models and, 236n4; random effects and, 242–246, 253–256, 261n17, 265, 268–270; regression models and, 232, 237, 239, 244, 252; root-$N$ consistent estimators and, 249–250; semiparametric models and, 230, 235, 246–250, 255; state dependence and, 230, 252, 261–280; static models and, 230, 235–250, 255, 265–266; statistics and, 230, 233, 235–236, 239, 255, 261, 262n18, 265, 266t, 268t, 269–280; stochastics and, 251–254, 278; structural parameters and, 238–239, 242, 248, 254n16, 271, 278; Taylor series and, 273; time series and, 243, 257, 269–270; variance-components models and, 266; vectors and, 230–233, 247, 254n16

discrete-response models, 230–235
distance to default (DTD), 437
distributed-lag models: common assumptions and, 419–420; discrete data and, 263; estimation and, 428–429; exogenous variables and, 421–425; incomplete panel data and, 16, 80n1, 418–429; lag coefficients and, 6, 418–429; panel data advantages and, 6; prior structure identification and, 421–425; short panels and, 418–429; testing and, 428–429
dummy variable: cross-sectionally dependent panel data and, 343, 352; discrete data and, 239, 268; dynamic models and, 81, 110–111; incomplete panel data and, 410; least-squares estimation and, 34–39; nonparametric panel data and, 463; panel data advantages and, 5, 467; sample truncation and, 297; simple regression models and, 34–39, 48; static simultaneous-equations models and, 142; variable coefficient models and, 168
duration model: asymptotics and, 434; baseline hazard function and, 433; cumulative distribution function and, 430–431; distance to default (DTD) and, 437; exit intensity and, 435–436; exogenous variables and, 433; failure time and, 433; hazard function and, 431–435; heterogeneity and, 433–434; individuals and, 430–435, 437; least-squares estimation and, 432–433; matrices and, 432–434; measurement errors and, 430; normalization and, 434; parametric models and, 434; probability and, 430–431, 433, 435, 438; regression models and, 432; statistics and, 430; survival function and, 431
dynamic models: arbitrary serial correlations and, 121–122; asymptotics and, 81–86, 94, 96–99, 101, 103, 105–107, 111n18, 112–113, 115–121, 124, 126–135; autocorellation and, 86, 94; autoregression and, 82n2, 86, 104, 118; bias and, 81–86, 97, 99, 103–104, 108, 109t, 111–112, 117–119, 123, 128, 133, 270–273; bounding parameters and, 274–276; censored, 317–324; conditional approach and, 255–261; covariance (CV) estimators and, 80–84, 92, 123, 132; density and, 107; dependent variables and, 80–81, 85–86, 97, 111–112, 115; diagonal path limits and, 131; discrete data and, 250–270; dummy variable and, 81, 110–111; efficiency and, 80, 103–104, 121, 128; endogenous variables and, 111; error-component model and, 85, 94, 96, 106, 108, 109t; error terms and, 81, 93, 107, 111, 118; examples of, 108–111, 264–270; exogenous variables and, 81–89, 93–94, 105, 107–108, 112, 121–122, 123n21; fixed constants (FE) and, 80, 88, 95, 123–124; fixed effects and, 80–83, 111–121, 132; generalized method of moments (GMM) and, 81, 89, 99–106, 112, 116–122, 125–128; general model and, 250–252; heterogeneity and, 111, 134, 261–264; homogeneity and, 92, 120–121; incidental parameters and, 81, 92, 97n12, 112, 120; independent variables and, 135; individuals and, 80–93, 97–98, 105, 108, 110–112, 117–129; infinity and, 81–83, 86, 91, 93–94, 97, 99, 111, 121, 124, 128, 130–132; initial conditions and, 81, 82n2, 85, 89–91, 98–99, 106–107, 117n20; instrumental variable (IV) estimators and, 81, 89, 98–99, 117, 125, 128; joint limits and, 131; least-squares estimation and, 81, 82n2, 85, 95–96, 104, 111, 115, 122–123, 133–134; log-likelihood function and, 106–107, 127–128; Markov process and, 251–253, 255, 256t, 265; matrices and, 94–97, 100–102, 104, 106–107, 109t, 112–113, 115–116, 126–130, 134; maximum-likelihood estimators (MLE) and, 81, 89–96, 99, 106–107, 117n20; mean square error and, 93, 103, 118, 120t; measurement errors and, 87; minimum-distance estimators (MDE) and, 95–96, 114–121,

129–130; Monte Carlo studies and, 84, 86, 94, 107, 117–118, 128; multicollinearity and, 91, 111; normalization and, 117; null hypotheses and, 106; omitted variables and, 97; ordinary least squares (OLS) estimation and, 85–86, 97, 99, 110–111; orthogonality and, 100, 105, 112, 125; probability and, 82n2, 85, 123, 131–132; random-coefficient models and, 206–212; random effects and, 80–81, 84–108, 109t, 120–121; regression models and, 84, 97, 111, 130, 133–135; sample selection and, 324–326; sequential limits and, 130; spatial approach and, 336–337; state dependence and, 261–264; static models and, 80–82, 100, 112; statistics and, 82n2, 84, 86n5, 105–108, 117, 121, 124, 130; stochastics and, 82n2, 87–88, 92, 94–95, 104, 108; Taylor series and, 129; time series and, 81, 83, 91, 108, 110, 119, 124, 130, 133–135; time-specific variables and, 80–81, 122–129; Tobit models and, 324–326; transformed likelihood approach and, 112–115; variance-components models and, 107; vectors and, 80, 87, 93, 94n9, 97, 100, 112, 121, 124, 131

dynamic systems: asymptotics and, 374–377, 381–382, 385–386, 388–392, 395–402; autoregression and, 369–378, 390; bias and, 381, 388–389, 400–402; change inferences and, 5; cointegrated panel models and, 379–386; cointegrating system and, 383–386; cointegration tests and, 394–397; common trends and, 379–380; degrees of freedom and, 386, 390, 396; dependent variables and, 369, 376–377; engodenous variables and, 369; equilibrium and, 380, 393n7; error terms and, 397–399; exogenous variables and, 369; fixed constants (FE) and, 372, 400–401; fixed effects and, 372, 382; generalized method of moments (GMM) and, 373–374, 377, 382, 401–402; heterogeneity and, 370, 377–378, 383–386, 390–392, 395–396; heteroscedasticity and, 382n5, 387–388; homogeneity and, 370–377, 386, 389–390, 395–396, 399; incidental parameters and, 372, 375; individuals and, 369–373, 377, 380–383, 386–390, 393–399, 402; infinity and, 392, 400–402; instrumental variable (IV) estimators and, 373, 401–402; joint dependent variables and, 369; Lagrangians and, 391; least-squares estimation and, 377–378, 381, 387–389, 394–395; likelihood approach and, 386, 395–396, 398–401; log-likelihood function and, 384, 399–400; matrices and, 370–385, 396–402; maximum-likelihood estimators (MLE) and, 386, 395–396, 398–401; minimum distance estimators (MDE) and, 375–377, 382; Monte Carlo studies and, 377, 382, 389–390, 394–395, 402; nonparametric panel data and, 390; normalization and, 380, 382, 385; null hypotheses and, 385–387, 390, 392; orthogonality and, 373, 401; parametric models and, 390; pooling and, 383, 389; probability and, 372, 394; random-coefficient models and, 392; random effects and, 372, 399; regression models and, 377–378, 382, 384, 386–387, 389–392, 395; simultaneous equations models and, 397–402; statistics and, 369, 377, 385–397; time series and, 16, 369–372, 377–379, 383, 386–387, 390, 392–395; time-specific variables and, 393; unit root tests and, 386–394; vector autoregressive (VAR) models and, 369–378, 380, 386, 393n7, 397; vectors and, 369–386, 393, 396–399

efficiency: cross-sectionally dependent panel data and, 327; discrete data and, 233, 242, 245; dynamic models and, 80, 103–104, 121, 128; generalized least-squares estimators and, 147; panel data advantages and, 4, 8; panel data issues and, 470, 473; simple

efficiency (*cont.*)
regression and, 32–33, 44, 50, 52, 54–55, 75n25; simulation methods and, 452; static simultaneous-equations models and, 136, 140, 143, 153; variable coefficient models and, 197, 224
eigenvectors, 148, 151, 334n4, 338–340
endogenous variables: dynamic models and, 111; dynamic systems and, 369; sample truncation and, 296; simple regression and, 73; static simultaneous-equations models and, 137, 151, 155, 159
English Longitudinal Study of Aging (ELSA), 4
equations: analysis of covariance (ANCOVA), 18–21, 23; Cobb-Douglas production function, 32; count data model, 439–444; cross-sectionally dependent panel data, 328–350, 352–363; discrete data, 230–235, 237–255, 257–263, 267, 269, 271–279; distribution-lag model, 6; duration model, 430–438; dynamic models, 80, 82–110, 112–116, 118, 121–134; dynamic simultaneous, 397–402; dynamic systems, 370–402; homogeneity, 17–21, 23, 25t; incomplete panel data, 404–411, 413–416, 418–429; least-square estimation, 14; least-squares dummy variable, 34–39; linear regression models, 11, 17–21, 23, 25t, 32, 34–47, 50–55, 58–79; measurement errors, 455–460; multilevel structures, 453–455; nonparametric panel data, 461–463; panel quantile regression, 445–447; sample truncation, 281–283, 285–295, 298–299, 301–302, 304–307, 309–312, 317, 321–326; simple regression model, 6–7, 17–21, 23, 25t, 32, 34–37, 50–55, 58–79; simulation methods, 448–452; slope coefficients, 15; static simultaneous-equations models, 136–166; Tobit model, 9; variable coefficient models, 168–202, 205–209, 212–213, 217–218, 222–229
equilibrium: discrete data and, 252–254; dynamic systems and, 380, 393n7; incomplete panel data and, 413
error-component model: cross-sectionally dependent panel data and, 334, 345; discrete data and, 265; dynamic models and, 85, 94, 96, 106, 108, 109t; incomplete panel data and, 404; multilevel structures and, 430, 453, 455; nonparametric panel data and, 461; sample truncation and, 292; simple regression models and, 39n5, 40, 64, 75; static simultaneous-equations models and, 139, 143–144, 147, 151–153; variable coefficient models and, 168n1, 174, 198, 200, 201n16
error-component two-stage least-squares estimators (EC2SLS), 147–148, 152, 154
error sum of squares (ESS), 219t
error terms: cross-sectionally dependent panel data and, 329, 336; discrete data and, 231, 235–236, 243–245, 251–252, 262–263; dynamic models and, 81, 93, 107, 111, 118; dynamic systems and, 397–399; homogeneity and, 18; incomplete panel data and, 404, 411, 414, 416, 420; linear regression models and, 18, 35, 50, 59, 63, 65–67, 69; mean square error and, 47; panel data advantages and, 6, 467; panel data issues and, 11, 14; sample truncation and, 281–282, 284–285, 287, 290, 292, 298, 302, 313–314, 317, 325; static simultaneous-equations models and, 139, 144, 152; variable coefficient models and, 183–185, 206. *See also* bias
Euclidean norm, 132, 247
Euclidean space, 330
European Community Household Panel (ECHP), 3
Eurostat, 2–3
exit intensity, 435–436
exogenous variables: count data model and, 441; cross-sectionally dependent panel data and, 336; discrete data and, 242, 252, 254–255, 258, 265, 267, 270,

277; distributed lag models and, 421–425; duration model and, 433; dynamic models and, 81–89, 93–94, 105, 107–108, 112, 121–122, 123n21; dynamic systems and, 369; homogeneity and, 17–18; incomplete panel data and, 418, 421–425; linear regression models and, 17–18, 34–35, 53; nonparametric panel data and, 461; panel data advantages and, 6, 467; panel data issues and, 11, 14, 471n1; prior structure identification and, 421–425; sample truncation and, 281, 287, 325; static simultaneous-equations models and, 137, 144, 153, 155, 161; variable coefficient models and, 169, 177, 187, 193–196, 198, 206

factor approach, 327, 337–342
failure time, 433
feasible general least-squares estimators (FGLS), 44, 65, 127–128, 225–226, 328, 335, 342
fixed-coefficient models: BLUEs and, 183; complete heterogeneity and, 170–171; heterogeneity and, 170–172; mixed fixed- and random-coefficient model and, 196–206; random-coefficient models and, 223–224; variable coefficient models and, 170–172, 177–178, 182–183, 196–206, 216–217
fixed constants (FE): cross-sectionally dependent panel data and, 334–335; discrete data and, 251, 254, 270; dynamic models and, 80, 88, 95, 123–124; dynamic systems and, 372, 400–401; incomplete panel data and, 407; linear regression models and, 24, 33, 38–39, 41, 50, 59, 62, 66; nonparametric panel data and, 461; panel data issues and, 13; simple regression models and, 33, 38–39, 41, 50, 59, 62, 66; variable coefficient models and, 170, 193, 210, 212, 223
fixed-effects inference, 56, 467
fixed-effects models: censored regression and, 306–311; conditional inference and, 48–56; conditions for existence of consistent estimators and, 238–242; cross-sectionally dependent panel data and, 334, 340, 360; discrete data and, 236–243, 254–255, 256t, 270–272; dynamic models and, 80–83, 111–121, 132; dynamic systems and, 372, 382; generalized method of moments (GMM) and, 116–119; incomplete panel data and, 408; individual correlations and, 52–56; linear regression models and, 34–39, 43–44, 47–59, 67; maximum-likelihood estimators (MLE) and, 236–238; minimum distance estimators (MDE) and, 114–121, 129–130; misspecification tests and, 56–58; Mundlak's formulation and, 50–54, 56; pairwise trimmed least-squares estimators and, 299–311; panel data advantages and, 467; panel data issues and, 10–11, 13, 470; panel quantile regression and, 447; random effects and, 47–56; sample truncation and, 299–314, 315t; simple regression models and, 47–56; specification issues and, 119–121; static simultaneous-equations models and, 138n2, 142; transformed likelihood approach and, 112–115; truncated regression and, 301–306; unconditional (marginal) inference and, 48–56; variable coefficient models and, 182, 223–224, 227
Frisch-Waugh FGLS approach, 342
Frobenius norm, 171
$F$ statistic, 21, 23, 214t, 215t
fuzzy regression discontinuity (FRD), 357–358

Gary income-maintenance project, 296–298
Gaussian quadrature, 245, 248
generalized least-squares (GLS) estimators: cross-sectionally dependent panel data and, 328, 332, 335–336, 342; dynamic models and, 89, 96–97, 104, 106, 110t, 111, 120, 124, 126–128; incomplete panel data

generalized least-squares (GLS) (*cont.*) and, 405–408; multilevel structures and, 454; sample truncation and, 295, 297; simple regression and, 41–44, 47n8, 52–68, 75; static simultaneous-equations models and, 142–144, 151–152, 162; variable coefficient models and, 171–175, 178, 184–185, 190, 195–196, 200–201, 225–226
generalized method of moments (GMM): cross-sectionally dependent panel data and, 341, 346, 348n8; discrete data and, 246, 261n17; dynamic models and, 81, 89, 99–106, 112, 116–122, 125–128; dynamic systems and, 373–374, 377, 382, 401–402; fixed-effects models and, 116–119; incomplete panel data and, 407; maximum-likelihood estimators (MLE) and, 116–119; measurement errors and, 460; sample truncation and, 323–326; simple regression models and, 70, 75n26; simulation methods and, 451–452; vector autoregressive (VAR) models and, 373–374
German Social Economics Panel (GSOEP), 2
Gibbs sampler, 209–210
global vector autoregressive (GVAR) models, 378, 386
group mean augmented approach, 342–344
Grunfeld investment function, 168, 175

Hausman type test statistic, 57–58, 105, 121, 408
Hausman-Wise (HW) model, 295, 297–298
hazard function, 431–435
Health and Retirement Study (HRS), 3
Heckman two-step estimators, 238–244, 288, 299
Hermite integration formula, 245
heterogeneity, 16; autoregression and, 377–378; cointegrating system and, 383–386; common correlated effects heterogeneous model (CCEMG) and, 350, 351t; complete, 170–171; cross-sectionally dependent panel data and, 350, 360; discrete data and, 230, 235–246, 251–252, 261–271, 278; duration model and, 433–434; dynamic models and, 111, 134; dynamic systems and, 370, 377–378, 383–386, 390–392, 395–396; fixed-coefficient models and, 170–171; fixed-effects models and, 236–242; group, 171–172; homogeneity and, 20–24, 26n4; intercepts and, 11, 20–23; linear regression models and, 31, 50, 53–54, 56; panel data advantages and, 5–6, 8, 464–465, 468; panel data issues and, 10–13, 469–470; parametric approach and, 235–246; random-effects models and, 242–246; sample truncation and, 298; simulation methods and, 448; state dependence and, 261–264; static models and, 235–246; static simultaneous-equations models and, 136; time and, 10–13; unobserved, 10–15, 31, 111, 136, 167–169, 202, 252, 270–271, 278, 298, 360, 434, 448, 465, 469–470; unobserved across individuals and over time and, 10–13; variable coefficient models and, 167–171, 178–180, 186, 196–197, 202, 204, 212, 214t, 215, 217–219; variable-intercept models and, 15, 31–32; vector autoregressive (VAR) models and, 377–378
heteroscedastic autoregression (HAC) covariance matrix, 331, 363
heteroscedasticity: autocorellation and, 68–69; covariance (CV) estimators and, 68–69; cross-sectionally dependent panel data and, 331, 340–341, 363; dynamic systems and, 382n5, 387–388; homogeneity and, 24; linear regression models and, 24, 34, 39n5, 64–70, 73n22; sample truncation and, 284, 313; serially correlated errors and, 65–68; simple regression models and, 34, 39n5, 64–70, 73n22; static simultaneous-equations models and, 145, 147, 149; variable coefficient models and, 170, 176, 179, 184–185, 224
Hildreth-Houck estimators, 184n10

homogeneity: autoregression and, 370–377; bias and, 24; cointegrated panel models and, 381–383; common correlated effects model (CCEP) and, 350, 351t; cross-sectionally dependent panel data and, 350, 356; degrees of freedom and, 18, 21, 23–25, 27t, 29t; discrete data and, 236, 253; dynamic models and, 92, 120–121; dynamic systems and, 370–377, 386, 389–390, 395–396, 399; error terms and, 18; exogenous variables and, 17–18; fixed constants (FE) and, 24; *F* statistic and, 21, 23; heterogeneity and, 20–24, 26n4; heteroscedasticity and, 24; incomplete panel data and, 409; individuals and, 17–19, 22t, 25–26; least-squares estimation and, 17, 19–20, 24; linear regression model tests and, 17–30; nonparametric panel data and, 461; null hypotheses and, 24; panel data advantages and, 4, 8, 465, 468; panel data issues and, 11; probability and, 17; residual sum of squares (RSS) and, 19–21, 22t; simple regression and, 31–32; slope and, 11, 20–24, 26; statistics and, 21, 23–24, 26; time series and, 24–26; variable coefficient models and, 177–179, 202, 213, 215t, 217; vector autoregressive (VAR) models and, 370–377; within-group estimators and, 19

idempotent (covariance) transformation matrix, 37, 40, 77, 141
identity matrix, 36, 141–143, 148, 380, 385
importance sampling, 450
incidental parameters: count data model and, 443; discrete data and, 236–240, 244, 247–248, 271; dynamic models and, 81, 92, 97n12, 112, 120; dynamic systems and, 372, 375; linear regression models and, 55–56, 65; multidimensional statistics and, 13; panel data issues and, 11, 13, 470–471; sample truncation and, 302, 312
income-schooling model, 136–137, 154, 466
incomplete panel data: asymptotics and, 410, 414–415, 428–429; attrition and, 408; autocorellation and, 417n6; autoregression and, 417n6, 418, 422, 425–426; Bayesian methods and, 412, 415n5; bias and, 412, 417, 419; covariance (CV) estimators and, 411; degrees of freedom and, 414, 421, 428–429; density and, 404; dependent variables and, 412–414, 418; distributed-lag models and, 16, 80n1, 418–429; dummy variable and, 410; equilibrium and, 413; error-component model and, 404; error terms and, 404, 411, 414, 416, 420; exogenous variables and, 418, 421–425; fixed constants (FE) and, 407; fixed effects and, 408; generalized method of moments (GMM) and, 407; homogeneity and, 409; independent variables and, 413–414; individuals and, 403–414, 417–420; infinity and, 409, 420; initial conditions and, 407, 426; instrumental variable (IV) estimators and, 425; lag coefficients and, 418–429; least-squares estimation and, 405, 415; matrices and, 405, 408, 410, 413–415, 421–422, 426–428; maximum-likelihood estimators (MLE) and, 406–407, 412–417; mean square error and, 410, 420; minimum-distance estimators and, 428; null hypotheses and, 414, 428; omitted variables and, 412–413, 417, 419; pooling of single cross-sectional data and, 411–418; probability and, 403; pseudopanels and, 16, 408–411; randomly missing data and, 403–408; regression models and, 412–413, 416, 423, 426–427; repeated cross-sectional data and, 408–411; rotating data and, 403–408; single time series data and, 411–418; statistics and, 404, 408–410, 415–416, 429; stochastics and, 421–422; time series and, 402–403, 405, 407, 411–419; vectors and, 404–405, 407, 409, 413–414, 420–422, 427–428
independence of irrelevant alternatives, 235

independence tests: example of, 350–351; limited dependent-variable model and, 348–350; linear model and, 344–348
independent variables: dynamic models and, 135; incomplete panel data and, 413–414
individuals, 1–3, 15; count data model and, 441–443; cross-sectionally dependent panel data and, 327–331, 334–336, 339–340, 352–353, 357, 359–364; discrete data and, 230–239, 242–243, 248–252, 254–255, 260–263, 265, 268–271, 277n21; duration model and, 430–435, 437; dynamic models and, 80–93, 97–98, 105, 108, 110–112, 117–129; dynamic systems and, 369–374, 377, 380–383, 386–390, 393–399, 402; homogeneity and, 17–19, 22t, 25–26; incomplete panel data and, 403–414, 417–420; linear regression models and, 17–19, 22t, 25–26, 31–40, 44, 48–70, 73; measurement errors and, 455–459; multilevel structures and, 453; nonparametric panel data and, 462; panel data advantages and, 4–10, 464–469; panel data issues and, 10–14, 469–473; panel quantile regression and, 447; sample truncation and, 290, 292, 295–301, 311–312, 317–318, 325; static simultaneous-equations models and, 136–138, 144–145, 149–153; unobserved heterogeneity and, 10–13; variable coefficient models and, 167–180, 184–186, 193, 197–201, 204, 207–208, 214t, 215, 218, 220, 223, 227
infinity: count data model and, 442; cross-sectionally dependent panel data and, 331, 333–334, 336, 363; discrete data and, 238–240, 243, 246, 253–255, 259; dynamic models and, 81–83, 86, 91, 93–94, 97, 99, 111, 121, 124, 128, 130–132; dynamic systems and, 392, 400–402; incomplete panel data and, 409, 420; measurement errors and, 456; sample truncation and, 299, 323; simple regression and, 38, 41, 47, 52, 58, 60–61, 63, 65, 67, 69–71; static simultaneous-equations models and, 138n2, 142, 145, 149; variable coefficient models and, 174, 176–178, 185, 196, 209
initial conditions: discrete data and, 252–255, 256t, 261; dynamic models and, 81, 82n2, 85, 89–91, 98–99, 106–107, 117n20, 252–255, 256t, 261; incomplete panel data and, 407, 426; maximum-likelihood estimators (MLE) and, 106
instrumental variable (IV) estimators: cross-sectionally dependent panel data and, 333; dynamic models and, 81, 89, 98–99, 117, 125, 128; dynamic systems and, 373, 401–402; incomplete panel data and, 425; measurement errors and, 456; panel data advantages and, 468; sample truncation and, 290–291, 314; static simultaneous-equations models and, 149, 155–159; variable coefficient models and, 172n3, 208, 221–223
Investing in Children and their Societies (ICS), 3
investment function, 24, 168, 175
investment ratio, 8

jackknife estimators, 271
Jacobian matrices, 54, 154–155, 422, 427
joint dependent variables: dynamic systems and, 369; panel data advantages and, 468; sample selection and, 295; static simultaneous-equations models and, 138, 144, 161, 163. *See also* endogenous variables
joint generalized least-squares estimators, 140–144
joint probability, 239, 252–253, 288, 295, 307

Kalman filter, 188–191
kernel estimates: cross-sectionally dependent panel data and, 358; discrete data and, 259, 277; dynamic systems and, 389; nonparametric panel data and, 461; sample truncation and, 288–290, 312–313, 325–326; simple

regression and, 69; variable coefficient models and, 227
kernel weighted generalized method of moments (KGMM), 325–326
Koyck lag, 419, 426
Kronecker product, 59n15, 61, 124, 141n5, 385, 454

labor, 1, 3; dynamic models and, 230, 235–236, 250, 264–266; homogeneity and, 32, 33t, 34t; incomplete panel data and, 404; panel data and, 4–5, 7; sample truncation and, 287, 296–297; variable coefficient models and, 179–180
lag coefficients: incomplete panel data and, 418–429; panel data advantages and, 6
Lagrangians: cross-sectionally dependent panel data and, 344–345, 349; dynamic systems and, 390–391; multiplier tests and, 162n14, 176–177, 185, 192, 345, 349, 390–391; variable coefficient models and, 162n14, 176–177, 185, 192
Lasso (Least Absolute Shrinkage and Selection Operator), 171
least absolute deviation (LAD) estimation: censored data and, 306–311; Honoré and, 299; panel quantile regression and, 446; regression models and, 301–311; sample truncation and, 287, 299–311; truncated regression and, 301–306
least-squares dummy variable (LSDV), 34–39, 43, 48, 82, 97, 100, 110t, 137, 239n10
least-squares estimators: covariance (CV) estimators and, 43; cross-sectionally dependent panel data and, 328–329, 333, 339–345, 364; discrete data and, 232–233, 239; dummy variable approach and, 34–39; duration model and, 432–433; dynamic models and, 81, 82n2, 85, 95–96, 104, 111, 115, 122–123, 133–134; dynamic systems and, 377–378, 381, 387–389, 394–395; efficiency and, 147; error-component two-stage (EC2SLS), 147–148, 152, 154; generalized, 37n3, 41–44, 66, 81, 96–97, 140–144, 147, 162, 171, 295, 300t, 378, 438, 454, 463; generalized (GLS), 41–44, 47n8, 52–68 (*see also* generalized least-squares (GLS) estimators); homogeneity and, 17, 19–20, 24; incomplete panel data and, 405, 415; joint general, 140–144; linear regression models and, 17, 19–20, 24, 32, 36–38, 41, 44, 50, 61–67, 71, 74–75; measurement errors and, 455–456; multilevel structures and, 454; OLS, 14 (*see also* ordinary least squares (OLS) estimators); pairwise trimmed, 299–311; panel data advantages and, 7, 9, 461, 463, 469; panel data issues and, 11, 14; penalized, 171; pooled, 11, 17, 210, 224–227, 328, 389; random effects and, 96–106; residual sum of squares (RSS) and, 19–21, 22t, 172; sample truncation and, 282–287, 290, 295, 299, 300t, 311, 313–314; simulation methods and, 451; static simultaneous-equations models and, 136–137, 139–151, 153, 162, 164; three-stage (3SLS), 74–75, 95–96, 122, 149–152; two-stage (2SLS), 104, 145, 147–148, 152, 154, 333; variable coefficient models and, 168, 171, 173–176, 184–185, 192, 195, 208, 210, 225–227; vector autoregressive (VAR) models and, 374–375
Lehman Brothers, 438
limited dependent-variable model, 348–350
limited information principle, 145, 147, 149, 151, 398, 400, 402
linear-probability model, 231–232, 238, 247
linear regression models: analysis of covariance (ANCOVA) and, 17–26, 35, 37; analysis of variance (ANOVA) and, 17, 35; arbitrary error structure and, 69–75; asymptotics and, 44, 47, 52, 55, 57–58, 65–68, 72–77; autocorellation and, 34, 64–69; basic assumptions of, 31; best linear unbiased estimators (BLUE) and, 36, 38–39, 41, 52, 60, 63;

linear regression models (*cont.*)
bias and, 24, 33, 36, 38, 41, 50–52, 55–58, 61; Chamberlain approach and, 69–75; Cobb-Douglas production function and, 32; conditional inference and, 48–56; consistency and, 75–77; covariance (CV) estimators and, 37–38, 41, 43–44, 47, 52, 54, 57, 59–65, 68–69; degrees of freedom and, 18, 21, 23–25, 27t, 29t, 54–55, 57–58, 73, 77; density and, 39, 54–56; dependent variables and, 34, 39, 49t, 67; dummy variable and, 34–39, 48; efficiency and, 32–33, 44, 50, 52, 54–55, 75n25; endogenous variables and, 73; error-component model and, 39n5, 40, 64, 75; error terms and, 18, 35, 50, 59, 63, 65–67, 69; exogenous variables and, 17–18, 34–35, 53; fixed constants (FE) and, 24, 33, 38–39, 41, 50, 59, 62, 66; fixed-effects models and, 34–39, 43–44, 47–59, 67; generalized method of moments (GMM) and, 70, 75n26; heterogeneity and, 31, 50, 53–54, 56; heteroscedasticity and, 24, 34, 39n5, 64–70, 73n22; homogeneity tests and, 17–30; incidental parameters and, 55–56, 65; individuals and, 17–19, 22t, 25–26, 31–40, 44, 48–70, 73; individual-specific variables and, 58–61; individual/time effects and, 61–64; least-squares estimation and, 17, 19–20, 24, 32, 36–38, 41, 44, 50, 61–67, 71, 74–75; matrices and, 36–44, 47, 52, 54, 57, 59, 62–63, 65–70, 72–77; maximum-likelihood estimators (MLE) and, 45–47; mean square error and, 47n8, 71; minimum-distance estimators and, 72–77; misspecification tests and, 56–58; Monte Carlo studies and, 47n8; multicollinearity and, 59; Mundlak's formulation and, 50–54, 56; normalization and, 74; null hypotheses and, 57–58; omitted variables and, 31–35, 39, 48, 65; ordinary least squares (OLS) estimation and, 36–37, 43, 59–61, 63; parametric models and, 55, 74; probability and, 47, 58, 75–77; random effects and, 33, 39–58; regression discontinuity design and, 357–359; static models and, 52, 55; statistics and, 21, 23–24, 26, 55, 58; stochastics and, 33, 35; structural parameters and, 55; Taylor series and, 77; three-component model and, 39, 77–79; time series and, 24–26, 36, 67, 70, 73; time-specific variables and, 32–34, 58–64; unconditional (marginal) inference and, 48–56; variance-components models and, 39–47, 52, 66; vectors and, 34–36, 39–40, 58, 61, 70–71, 73, 75–79; within-group estimators and, 37, 43–44, 54
liquidity: homogeneity and, 26; variable coefficient models and, 169, 212–217
local average treatment effect (LATE), 352n9
logit models: discrete data and, 231–232, 234, 237–239, 241–242, 256, 259, 268, 271, 276–280, 471n1; panel data issues and, 471n1; pooled, 268–269
log-likelihood function: count data model and, 441–443; cross-sectionally dependent panel data and, 332–336; discrete data and, 237, 240–241, 243–244, 258, 263, 265, 266t, 272, 280; dynamic models and, 106–107, 127–128; dynamic systems and, 384, 399–400; sample truncation and, 295; simulation methods and, 451; static simultaneous-equations models and, 159–161; variable coefficient models and, 176n7, 191, 205

managerial-differences variable, 152
Manheim Innovation Panel (MIP), 3
Manheim Innovation Panel-Service Sector (MIP-S), 3
marginal treatment effect (MTE), 352n9
Markov Chain Monte Carlo method, 209–210, 251–253, 255, 256t, 265
matrices: asymptotic covariance, 74–75, 101, 115–116, 126, 128–130, 224, 240, 250, 272, 283, 305, 310, 313, 323, 336, 375–376, 434, 451; cross-sectionally

dependent panel data and, 327–331, 334, 336–340, 344, 350, 362; definite, 100, 106; discrete data and, 232–233, 240–242, 244, 246, 250–255, 264, 266, 272; duration model and, 432–434; dynamic models and, 94–97, 100–102, 104, 106–107, 109t, 112–113, 115–116, 126–130, 134; dynamic systems and, 370–385, 396–402; heteroscedastic autoregression (HAC) covariance, 331, 363; idempotent (covariance) transformation, 37, 40, 77, 141; identity, 36, 141–143, 148, 380, 385; incomplete panel data and, 405, 408, 410, 413–415, 421–422, 426–428; information, 177, 240–241, 242n11, 255, 266, 283, 471; Jacobian, 54, 154–155, 422, 427; measurement errors and, 457–459; multilevel structures and, 453–455; nonpositive semidefinite, 116; orthogonal, 165; panel data issues and, 471; partitioned, 52, 142; positive definite, 289, 323, 328, 362, 451; positive semidefinite, 43, 116, 246n13, 377; sample truncation and, 283, 288–289, 305, 310, 313, 323, 325; simple regression and, 36–44, 47, 52, 54, 57, 59, 62–63, 65–70, 72–77; simulation methods and, 450–451; sparse, 330, 333; static simultaneous-equations models and, 138–155, 158–160, 165; stochastic, 325; transformation, 37, 40, 44, 59, 62, 67, 334n4, 459; variable coefficient models and, 170–175, 177, 182–184, 187–195, 197, 201–203, 206–209, 219t, 223–227; variance-covariance, 38, 40, 47, 54n11, 57, 62–77, 95–96, 107, 139–140, 146–147, 150–151, 154–155, 173, 183–184, 194–195, 201, 219t, 242n11, 244, 246n13, 251, 283, 376–377, 405, 410, 414–415, 428, 453

maximum-likelihood estimators (MLE): cointegrated panel models and, 382–384; count data model and, 442–443; cross-sectionally dependent panel data and, 354, 374–375, 383, 398, 400, 402; discrete data and, 233, 236–240, 245, 255, 263–264, 278; dynamic models and, 81, 89–96, 99, 106–107, 117n20; dynamic systems and, 386, 395–396, 398–401; fixed-effects models and, 116–119, 236–238; generalized method of moments (GMM) and, 116–119; hazard function and, 431–435; incomplete panel data and, 406–407, 412–417; initial conditions and, 106; limited information, 117n20, 400, 402; linear regression models and, 45–47; log-likelihood function and, 106–107, 127–128, 159–161, 176n7, 191, 205, 237, 240–241, 243–244, 258, 263, 265, 266t, 272, 280, 295, 332–336, 384, 399–400, 441–443, 451; Monte Carlo studies and, 94; multilevel structures and, 455; panel data advantages and, 9, 469; panel limited information (PLIML), 400–402; pooling and, 413–415; quasi, 124, 335–337, 401; random-coefficient models and, 224–226; sample truncation and, 283, 294, 297, 300t, 312; simulated (SMLE), 451–452; spatial dynamic models and, 336–337; spatial error model and, 334–335; spatial lag model and, 333; static simultaneous-equations models and, 152, 155, 159–162; time series and, 413–415; transformed, 112–115, 374–375; triangular system and, 159–162; variable coefficient models and, 176–177, 184, 188, 191–192

maximum score estimators, 247–249, 256, 259–260, 312

mean square error: dynamic models and, 93, 103, 118, 120t; incomplete panel data and, 410, 420; root (RMSE), 118–119, 120t, 128, 203t, 204t; simple regression models and, 47n8, 71; variable coefficient models and, 190, 201n16, 203, 204t

measurement errors, 16; arbitrary error structure and, 69–75; autocorellation and, 460; Chamberlain approach and, 69–75; correlation and, 456–460; dependent variables and, 458; duration

measurement errors (*cont.*)
model and, 430; dynamic models and, 87; generalized method of moments (GMM) and, 460; individuals and, 455–459; infinity and, 456; instrumental variable (IV) estimators and, 456; least squares and, 455–456; matrices and, 457–459; mean square error and, 47 (*see also* mean square error); misspecification tests and, 56–58; ordinary least squares (OLS) estimation and, 455; orthogonality and, 459; panel data advantages and, 9, 467–468; probability and, 456; regression models and, 456, 460; static simultaneous-equations models and, 153; vectors and, 379–386, 396, 458–460
Merrill Lynch, 438
minimum-distance estimators (MDE): cointegrated panel models and, 382–383; discrete data and, 246; dynamic models and, 95–96, 114–121, 129–130; dynamic systems and, 375–377; fixed-effects models and, 114–121, 129–130; incomplete panel data and, 428; limited-information, 147; simple regression models and, 72–77; static simultaneous-equations models and, 145, 147, 149–151; vector autoregressive (VAR) models and, 375–377
missing at random (MAR) model, 297–298
misspecification tests, 56–58
mixed fixed- and random-coefficient model: Bayes solution and, 198–201; formulation of, 196–198; individual parameter estimates and, 201–202; model selection and, 204–206; pooled parameter estimates and, 201–202; prediction comparison and, 202–204
model of limited dependent variables, 281
Modigliani-Miller theory, 212
Monte Carlo studies: cross-sectionally dependent panel data and, 342, 346, 348, 350; discrete data and, 242, 254–255, 256t, 270, 273, 280; dynamic models and, 84, 86, 94, 107, 117–118, 128; dynamic systems and, 377, 382, 389–390, 394–395, 402; Gibbs sampler and, 209–210; importance sampling and, 450; Markov Chain, 209–210, 251–253, 255, 256t, 265; maximum-likelihood estimators (MLE) and, 94; panel quantile regression and, 447; simple regression models and, 47n8; simulation methods and, 450; variable coefficient models and, 206, 209–210, 212, 227
multicollinearity: dynamic models and, 91, 111; panel data advantages and, 6, 464; simple regression and, 59; variable coefficient models and, 201, 213–215, 223, 226
multilevel structures: dependent variables and, 453; error-component model and, 430, 453, 455; individuals and, 453; least-squares estimation and, 454; matrices and, 453–455; maximum-likelihood estimators (MLE) and, 455; multiway error components model and, 453; vectors and, 453
Mundlak-Chamberlain approach, 172n3
Mundlak's formulation, 50–54, 56

National Data Collection Units (NDU), 2
National Longitudinal Surveys (NLS), 1–2, 13
negative binomial distribution, 442
neighbors, 329–331, 333
New Cronos, 3
Newey-West heteroscedasticity-autocorrelation consistent formula, 224
Newton-Raphson iterative procedure, 46, 94, 114, 233, 280, 283
non-government organizations (NGOs), 3
nonparametric panel data: asymptotics and, 462–463; counterfactuals prediction and, 361–363; cross-sectionally dependent panel data and, 331, 355, 357, 359–368; dependent variables and, 463; difference-in-difference method and, 360–361; dummy variable and, 463;

dynamic systems and, 390; error-component model and, 461; exogenous variables and, 461; fixed constants (FE) and, 461; homogeneity and, 461; individuals and, 462; normalization and, 461; parametric models and, 461; sample truncation and, 288, 298, 326; time series and, 462; vectors and, 463
Nordica, 267
normalization: cross-sectionally dependent panel data and, 329, 332, 337–338, 340; discrete data and, 235–236, 241, 246–250, 252, 270, 279; duration model and, 434; dynamic models and, 117; dynamic systems and, 380, 382, 385; nonparametric panel data and, 461; sample truncation and, 287, 293; simple regression and, 74; static simultaneous-equations models and, 154–155, 161, 162n14, 164; variable coefficient models and, 199
null hypotheses: cross-sectionally dependent panel data and, 344; discrete data and, 255, 262n18; dynamic models and, 106; dynamic systems and, 385–387, 390, 392; homogeneity and, 24; incomplete panel data and, 414, 428; simple regression and, 57–58; variable coefficient models and, 176–177, 186, 192–193, 212, 214

omitted variables: average vs. individual behavior and, 235; covariations of, 412–413; cross-sectionally dependent panel data and, 327; discrete data and, 235, 243; dynamic models and, 97; incomplete panel data and, 412–413, 417, 419; linear regression models and, 31–35, 39, 48, 65; panel data advantages and, 6, 8, 466–467; static simultaneous-equations models and, 136, 152–153; time series and, 412–413; variable coefficient models and, 167, 198
ordinary least squares (OLS) estimators: asymptotics and, 97; bias and, 85–86, 97; cross-sectionally dependent panel data and, 364, 368; dynamic models and, 85–86, 97, 99, 110–111; linear regression models and, 36–37, 43, 59–61, 63; measurement errors and, 455; random effects and, 85–86; sample truncation and, 316t; two-step GLS estimator and, 111n18; variable coefficient models and, 211t
Organisation for Economic Co-operation and Development (OECD), 3, 364
orthogonality: cross-sectionally dependent panel data and, 355; dynamic models and, 100, 105, 112, 125; dynamic systems and, 373, 401; measurement errors and, 459; panel data issues and, 471; sample selection and, 322–323

panel analysis of nonstationarity in idiosyncratic and common components (PANIC), 398n8
panel data: accuracy and, 4; capacity and, 4; cross-sectionally dependent panel data and, 327–368 (*see also* cross-sectionally dependent panel data); incidental parameters and, 470–471; increasing availability of, xvii; issues with, 10–14, 469–473; omitted variables and, 6, 8, 31–35 (*see also* omitted variables); statistics and, 3 (*see also* statistics); treatment effects and, 359–368
panel data advantages: aggregation levels and, 468–469; asymptotics and, 9, 469; bias and, 7–8, 464, 467–468; computation simplification and, 469; degrees of freedom and, 4, 464; density and, 9–10; dependent variables and, 467–468; dummy variable and, 5; efficiency and, 4, 8; error terms and, 6, 467; exogenous variables and, 6, 467; fixed effects and, 467; homogeneity and, 4, 8, 465, 468; identification/discrimination of competing hypotheses and, 465–467; impact of observables selection and, 469; individuals and, 4–10, 464–469; instrumental variable (IV) estimators

panel data advantages (*cont.*)
and, 468; joint dependent variables and, 468; least-squares estimation and, 7, 9, 461, 463, 469; maximum-likelihood estimators (MLE) and, 9, 469; measurement errors and, 9, 467, 467–468; multicollinearity and, 6, 464; parametric models and, 467; prediction accuracy and, 468–469; random effects and, 467; regression models and, 6–8, 467–468; specification problem and, 467; statistics and, 467, 469; time series and, 4, 6–9, 464–469; time-specific variables and, 467–468; vectors and, 6
panel data issues: asymptotics and, 13, 472; attrition and, 13, 469, 472–473; Bayesian methods and, 471; bias and, 11–14, 471–472; cross-sectionally dependent panel data and, 471–472; degrees of freedom and, 470; density and, 10; dependent variables and, 14, 472; efficiency and, 470, 473; error terms and, 11, 14; exogenous variables and, 11, 14; fixed constants (FE) and, 13; fixed effects and, 10–11, 13, 470; homogeneity and, 11; incidental parameters and, 11, 13; individuals and, 10–14, 469–473; least-squares estimation and, 11, 14; logit models and, 471n1; matrices and, 471; multidimensional asymptotics and, 472; orthogonality and, 471; parametric models and, 10; probability and, 470; random effects and, 10, 470–471; regression models and, 11, 14; sample attrition and, 13–14, 469, 472–473; state dependence and, 13; statistics and, 471–473; structural parameters and, 10, 471; time series and, 13, 469–473; unobserved heterogeneity and, 469–471; vectors and, 10
panel least variance ratio estimators (PLVAR), 402
panel limited information maximum-likelihood estimators (PLIML), 400–402
panel quantile regression, xvii; asymptotics and, 447; bias and, 447; density and, 445; fixed effects and, 447; individuals and, 447; probability and, 445; vectors and, 446
Panel Study of Income Dynamics (PSID), 1–2
panel vector autoregressive (PVAR) models: dynamic systems and, 370–378, 380, 382, 395–396; heterogenous, 377–378; homogenous, 370–377; likelihood approach and, 395–396
parametric models: cross-sectionally dependent panel data and, 331, 354, 357, 359–360; discrete data and, 230–231, 235–250, 255; duration model and, 434; dynamic systems and, 390; nonparametric panel data and, 461–462; panel data advantages and, 467; panel data issues and, 10; sample truncation and, 288–291, 298–299, 311–313, 326; simple regression and, 55, 74; variable coefficient models and, 227
Penn-World tables, 130
Poisson process, 439–440, 443–444
pooled regression models, 11, 19–20, 390
pooling: dynamic systems and, 383, 389; homogeneity and, 18n1, 24; incomplete panel data and, 16, 411–418; panel data advantages and, 8, 464; panel data issues and, 11; panel quantile regression and, 447; time series and, 411–418; variable coefficient models and, 169, 197, 203, 215, 217
predicted error sum of squares (PES), 219t
Primary School Deworming Project (PDSP), 3
probability: attrition and, 292–296; conditional, 239–241, 252, 256, 262–263, 295, 431, 433; count data model and, 438–439, 441t, 444; cross-sectionally dependent panel data and, 339, 355, 357, 361; discrete data and, 231–232, 234–236, 238–257, 261–265, 268–275; duration model

and, 430–431, 433, 435, 438; dynamic models and, 82n2, 85, 123, 131–132; dynamic systems and, 372, 394; homogeneity and, 17; incomplete panel data and, 403; joint, 239, 252–253, 288, 295, 307; measurement errors and, 456; panel data advantages and, 5; panel data issues and, 10, 470; panel quantile regression and, 445; sample truncation and, 284, 288–297, 302, 307, 321, 325; selection bias and, 292–296; simple regression and, 47, 58, 75–77; simulation methods and, 448; static simultaneous-equations models and, 137, 148; variable coefficient models and, 169, 198–199, 205, 209
probit models: cross-sectionally dependent panel data and, 349; discrete data and, 231, 234, 236n4, 241–244, 254–255, 261n17, 264, 268t; sample truncation and, 281; simulation methods and, 449
propensity score method, 227, 355–357, 360
pseudopanels, xix, 16, 408–411

quasi maximum-likelihood estimators (MLE), 124, 335–337, 401

random-coefficient models: analysis of covariance (ANCOVA) and, 177, 179, 186; Bayes solution and, 198–201; correlated, 220–227; cross-sectionally dependent panel data and, 221–222; description of, 172–173; discrete data and, 236n4; dynamic, 206–212, 392; estimation and, 173–175; example for, 179–180; fixed coefficients and, 178–179, 201–206, 223–224; Grunfeld investment function and, 175; maximum-likelihood estimators (MLE) and, 224–226; mixed fixed- and random-coefficient model and, 196–206; parameter reduction and, 169; predicting individual coefficients and, 175; sample truncation and, 298n5; simple regression and, 70; simulation methods and, 448; Swamy formulation and, 178, 180; testing for coefficient variation and, 175–178; variable coefficient models and, 169–170, 172–180, 183–186, 196–212, 215–227
random-effects inference, 56, 467
random effects models: covariance (CV) estimators and, 40–41; cross-sectionally dependent panel data and, 335; discrete data and, 242–246, 253–256, 261n17, 265, 268–270; dynamic models and, 80–81, 84–108, 109t, 120–121; dynamic systems and, 372, 399; estimation of, 89–106; estimation of variance-components models and, 39–47; fixed effects and, 47–56; generalized least-squares estimation and, 37n3, 41–44, 66; individual correlations and, 52–56; initial conditions testing and, 106–107; instrumental-variable estimators and, 98–99; least-squares estimators and, 96–106; linear regression models and, 33, 39–58; maximum-likelihood estimators (MLE) and, 45–47, 89–96; misspecification tests and, 56–58; model formulation and, 86–89; Mundlak's formulation and, 50–54, 56; OLS estimators bias and, 85–86; panel data advantages and, 467; panel data issues and, 10, 470–471; sample truncation and, 299, 314–317; simple regression models and, 47–56; simulation methods and, 107–108, 448; specification issues and, 119–121; unconditional (marginal) inference and, 48–56; variable coefficient models and, 215; variance-components models and, 39–47
randomly missing data, 403–408
regression discontinuity (RD) design, 357–359
regression models: augmented Dickey-Fuller (ADF), 389–390, 392; cell-mean corrected, 19, 27t–28t; censored, 306–311; cross-sectionally dependent panel data and, 328–329, 339, 341, 350, 357–358; discrete data and, 232, 237, 239, 244, 252; duration

regression models (*cont.*)
model and, 432; dynamic models and, 84, 97, 111, 130, 133–135; dynamic systems and, 377–378, 382, 384, 386–387, 389–392, 395; incomplete panel data and, 412–413, 416, 423, 426–427; least-squares estimators and, 7 (*see also* least-squares estimators); linear, 15 (*see also* linear regression models); measurement errors and, 456, 460; OLS, 85–86 (*see also* ordinary least squares (OLS) estimators); panel data advantages and, 6–8, 467–468; panel data issues and, 11, 14; panel quantile, 16, 445–447; pooled, 11, 19–20, 390; residual sum of squares (RSS) and, 19–21, 22t, 172; sample truncation and, 281, 285, 289–290, 296, 299, 306, 309, 311, 313, 324n10, 326; seemingly unrelated regression method (SUR) and, 140n4, 143, 170–171, 197, 328, 377–378, 384, 386; static, 84; static simultaneous-equations models and, 140n4, 143, 146; truncated, 301–306; $T$-variate, 70; variable coefficient models and, 167–176, 185, 187, 189, 192–193, 196–197, 202t, 208, 213, 214t–216t, 220, 225–226
repeated cross-sectional data, 408–411
research and development (R&D), 3
Research Center for Rural Development, 2
residual sum of squares (RSS), 19–21, 22t, 172
return-to-normality model, 187
root mean square error (RMSE), 118–119, 120t, 128, 203t, 204t
rotating data, 403–408

sample selection, xix; attrition and, 13–14, 292–298, 300t, 312, 469, 472–473; autoregression and, 324; control groups and, 296; dynamic models and, 324–326; endogenous issues and, 287–288; Hausman-Wise two-period model and, 295, 297–298; Heckman two-step estimators and, 238–244, 288, 299; housing expenditure example and, 313–317; joint dependent variables and, 295; latent response function and, 281; missing at random (MAR) model and, 297–298; nonrandomly missing data and, 292–298; orthogonality and, 322–323; Robinson approach and, 289, 291; self selection and, 292; semiparametric two-step estimators and, 311–313; single index, 290; Tobit models and, 298–299, 324–326
sample truncation: asymptotics and, 283, 287–289, 295, 298–299, 305, 310, 313, 323, 326; bias and, 282, 290, 292–297, 312, 314; censored data and, 281–287, 298–301, 306–311, 317–318, 323–324; censored regression and, 306–311; conditional expectation and, 282; data points and, 281–284; degrees of freedom and, 314, 317; density and, 282–284, 289–291, 296, 298, 302, 305, 312, 326; dependent variables and, 281–282, 284–285, 287, 290, 292, 295, 317, 325; dummy variable and, 297; endogenous issues and, 287–288; endogenous variables and, 296; error-component model and, 292; error terms and, 281–282, 284–285, 287, 290, 292, 298, 302, 313–314, 317, 325; exogenous variables and, 281, 287, 325; fixed effects and, 299–314, 315t; generalized method of moments (GMM) and, 323–326; Heckman two-step estimators and, 238–244, 288, 299; heterogeneity and, 298; heteroscedasticity and, 284, 313; housing expenditure example and, 313–317; incidental parameters and, 302, 312; individuals and, 290, 292, 295–301, 311–312, 317–318, 325; infinity and, 299, 323; instrumental variable (IV) estimators and, 290–291, 314; latent response function and, 281; least absolute deviation (LAD) estimators and, 287, 299–311; least-squares estimation and, 282–287, 290, 295, 299, 300t, 311, 313–314; log-likelihood function and, 295; matrices and, 283, 288–289, 305, 310,

313, 323, 325; maximum-likelihood estimators (MLE) and, 283, 294, 297, 300t, 312; model of limited dependent variables and, 281; nonparametric panel data and, 288, 298, 326; nonrandomly missing data and, 292–298; normalization and, 287, 293; ordinary least squares (OLS) estimation and, 316t; parametric models and, 288–291, 298–299, 311–313, 326; probability and, 284, 288–297, 302, 307, 321, 325; probit models and, 281; random-coefficient models and, 298n5; random effects and, 299, 314–317; regression models and, 281, 285, 289–290, 296, 299, 306, 309, 311, 313, 324n10, 326; Robinson approach and, 289, 291; semiparametric models and, 288–291, 311–313; static models and, 317; statistics and, 281, 292, 297, 314, 317, 323; stochastics and, 287, 325; structural models and, 292, 297, 300t; structural parameters and, 293, 295, 297; Tobit models and, 281, 288, 298–299, 317–326; truncated regression and, 301–306; unimodality and, 286–287; variable-intercept models and, 311; variance-components models and, 295; vectors and, 281, 287, 289–290, 293, 306, 310, 313–314, 324–325
seemingly unrelated regression method (SUR), 140n4, 143, 170–171, 197, 328, 377–378, 384, 386
semiparametric models: discrete data and, 230, 235, 246–250, 255; duration model and, 434; maximum score estimators and, 247–249; nonparametric panel data and, 461; panel data issues and, 471; root-*N* consistent estimators and, 249–250; sample truncation and, 288–291, 311–313; static models and, 246–250; variable coefficient models and, 227
sequential limit theory, 130–131
sharp regression discontinuity (SRD), 357–358
simple regression models: Aitken estimators and, 37n3; arbitrary error structure and, 69–75; asymptotics and, 44, 47, 52, 55, 57–58, 65–68, 72–77; autocorellation and, 34, 64–69; autoregression and, 66; basic assumptions of, 31; best linear unbiased estimators (BLUE) and, 36, 38–39, 41, 52, 60, 63; Chamberlain approach and, 69–75; Cobb-Douglas production function and, 32; conditional inference and, 48–56; consistency and, 75–77; covariance (CV) estimators and, 37, 40–41, 67–68; cross-sectionally dependent panel data and, 328; density and, 39, 54–56; dependent variables and, 34, 39, 49t, 67; dummy variable and, 34–39, 48; dynamic models and, 80–135 (*see also* dynamic models); efficiency and, 32–33, 44, 50, 52, 54–55, 75n25; endogenous variables and, 73; error-component model and, 39n5, 40, 64, 75; fixed constants (FE) and, 33, 38–39, 41, 50, 59, 62, 66; fixed-effects models and, 34–39, 43–44, 47–59, 67; generalized least-squares estimation and, 37n3, 41–44, 66; generalized method of moments (GMM) and, 70, 75n26; heteroscedasticity and, 34, 39n5, 64–70, 73n22; homogeneity and, 31–32; individual correlations and, 52–56; individual-specific variables and, 58–64; individual/time effects and, 61–64; infinity and, 38, 41, 47, 52, 58, 60–61, 63, 65, 67, 69–71; matrices and, 36–44, 47, 52, 54, 57, 59, 62–63, 65–70, 72–77; maximum likelihood estimation and, 45–47; mean square error and, 47n8, 71; minimum-distance estimators and, 72–77; misspecification tests and, 56–58; Monte Carlo studies and, 47n8; Mundlak's formulation and, 50–54, 56; multicollinearity and, 59; normalization and, 74; null hypotheses and, 57–58; panel data advantages and, 6; parametric models and, 55, 74;

simple regression models (*cont.*)
probability and, 47, 58, 75–77; random-coefficient models and, 70; random effects models and, 33, 33, 39–58; serially correlated errors and, 65–68; static models and, 52, 55; stochastics and, 33, 35; three-component model and, 39, 77–79; time effects and, 61–64; time-specific variables and, 32–34, 58–64; unconditional (marginal) inference and, 48–56; variable coefficient models and, 173; variance-components models and, 39–47, 52, 66; vectors and, 34–36, 39–40, 58, 61, 70–71, 73, 75–79; within-group estimators and, 37, 43–44, 54
simulated generalized method of moments (SGMM) estimators, 451–452
simulated maximum-likelihood estimators (SMLE), 451–452
simulated method of moments (SMM) estimators and, 451–452; Tobit models and, 448–449, 451; vectors and, 451
simulation methods: asymptotics and, 451–452; autoregression and, 450; bias and, 450, 452, 455–457; density and, 448–452; efficiency and, 452; generalized method of moments (GMM) and, 451–452; heterogeneity and, 448; importance sampling and, 450; least-squares estimation and, 451; log-likelihood function and, 451; matrices and, 450–451; maximum-likelihood estimators (MLE) and, 451; Monte Carlo studies and, 450; probability and, 448; probit models and, 449; random-coefficient models and, 448; random effects and, 107–108, 448
Social Economic Panel (PSELL) (Luxembourg), 2
*Social Science Citation Index*, xix
Socio-Economic Panel (SEP), 2
sparse elements, 330, 333
spatial approach: cross-sectionally dependent panel data and, 329–335, 337, 344; dynamic models and, 336–337; economic distance and, 331; error model and, 332–335; independence testing and, 344–351; individual-specific effects and, 334–336; lag model and, 333, 335–336; neighbors and, 329–331, 330, 333; sparse elements and, 330; SYR test and, 346–348
spatial autoregressive form, 329–330
spatial lag operator, 329
spatial moving average, 329
specification problem, 467
state dependence: approximate method and, 276–280; bias-adjusted estimators and, 270–273; bounding parameters and, 274–276; discrete data and, 230, 252, 261–280; heterogeneity and, 261–264; panel data issues and, 13
static models, 15; discrete data and, 230, 235–250, 255, 265–266; dynamic models and, 80–82, 100, 112; heterogeneity and, 235–246; linear, 52, 55, 80–81, 100, 138n2, 239n10; maximum score estimators and, 247–249; parametric, 235–246; root-$N$ consistent estimators and, 249–250; sample truncation and, 317; semiparametric, 230, 246–250; simple regression and, 52, 55
static simultaneous-equations models: Aitken estimators and, 148; asymptotics and, 136, 139, 143, 146–147, 150–152, 183n2; bias and, 136–137, 143, 152; covariance (CV) estimators and, 164; dependent variables and, 136, 138, 144, 161, 163; dummy variable and, 142; efficiency and, 136, 140, 143, 153; endogenous variables and, 137, 151, 155, 159; error-component model and, 139, 143–144, 147, 151–153; error terms and, 139, 144, 152; estimation of structural equations and, 144–152; exogenous variables and, 137, 144, 153, 155, 161; fixed effects and, 138n2, 142; heterogeneity and, 136; heteroscedasticity and, 145, 147, 149; identification and, 153–155; income-schooling model and, 136–137, 154, 466; individuals and,

136–138, 144–145, 149–153; infinity and, 138n2, 142, 145, 149; instrumental variable (IV) estimators and, 149, 155–159; joint dependent variables and, 138, 144, 161, 163; least-squares estimation and, 136–137, 139–151, 153, 162, 164; log-likelihood function and, 159–161; matrices and, 138–155, 158–160, 165; maximum-likelihood estimators (MLE) and, 152, 155, 159–162; measurement errors and, 153; minimum-distance estimators and, 145, 147, 149–151; normalization and, 154–155, 161, 162n14, 164; omitted variables and, 136, 152–153; probability and, 137, 148; regression models and, 140n4, 143, 146; structural models and, 139, 144–152; structural parameters and, 152; time series and, 138; time-specific variables and, 138; variance-components models and, 144, 149, 152–153; vectors and, 138–141, 143–144, 148, 151, 154, 160–161, 162n14, 165, 327–329, 334n4, 336–340, 343, 356–357, 362

Statistical Office of the European Communities, 2

statistics: Bureau of Labor Statistics and, 1; chi-square distribution and, 9, 233, 262n18, 268t, 377, 391; count data model and, 443; cross-sectionally dependent panel data and, 327–328, 344–346, 348–350, 354–355, 364, 368; discrete data and, 230, 233, 235–236, 239, 255, 261, 262n18, 265, 266t, 268t, 269, 270–280; duration model and, 430; dynamic models and, 82n2, 84, 86n5, 105–108, 117, 121, 124, 130; dynamic systems and, 369, 377, 385–397; European Community Household Panel (ECHP) and, 3; formal level of, 24; foundational theorems of, 10; $F$ statistic and, 21, 23, 214t, 215t; Hausman type test, 57–58, 105, 121, 408; homogeneity and, 21, 23–24, 26; incidental parameters and, 11 (*see also* incidental parameters); incomplete panel data and, 404, 408–410, 415–416, 429; inference and, 9, 117, 173, 236, 327, 369, 397, 430, 469, 473; Lagrangian multiplier tests and, 162n14, 176–177, 185, 192, 345, 349, 391; least-squares estimators and, 82n2 (*see also* least-squares estimators); linear regression models and, 21, 23–24, 26, 55, 58; MLE and, 9 (*see also* maximum-likelihood estimators (MLE); multidimensional, 13; panel data advantages and, 467, 469; panel data issues and, 471–473; random effects and, 10 (*see also* random effects models); regression models and, 24 (*see also* regression models); sample truncation and, 281, 292, 297, 314, 317, 323; specification problem and, 467; structural parameters and, 10 (*see also* structural parameters); test interpretation and, 24; $t$-statistic and, 84, 124, 202t, 368, 388–390, 394, 410; unobserved heterogeneity and, 10; variable coefficient models and, 173, 176–179, 184, 193, 202t, 204, 214, 215t, 217–218; Wald type tests and, 9, 377, 385–386

stochastics: cross-sectionally dependent panel data and, 347–348; discrete data and, 251–254, 278; dynamic models and, 82n2, 87–88, 92, 94–95, 104, 108; incomplete panel data and, 421–422; sample truncation and, 287, 325; simple regression and, 33, 35; variable coefficient models and, 169, 193, 195, 197

structural models: estimation of complete system and, 149–152; estimation of single equation in, 144–149; sample truncation and, 292, 297, 300t; static simultaneous-equations models and, 139, 144–152; triangular system and, 152–164

structural parameters: discrete data and, 238–239, 242, 248, 254n16, 271, 278; linear regression models and, 55; panel data issues and, 10, 471; sample truncation and, 293, 295, 297; static simultaneous-equations models and, 152

Survey of Health, Aging and Retirement in Europe (SHARE), 4
Survey of Income and Program Participation, 404
survival function, 431
symmetry: cross-sectionally dependent panel data and, 328–329, 332; discrete data and, 232, 249; dynamic systems and, 390; sample truncation and, 284–291, 302, 304, 307, 318, 321–323; simple regression and, 77; static simultaneous-equations models and, 141, 155, 164; variable coefficient models and, 212, 224

Taylor series, 77, 129, 273
three-stage least-squares (3SLS) estimators, 74–75, 95–96, 122, 149–152
time-evolving coefficients: Kalman filter predictions and, 188–191; maximum-likelihood estimators (MLE) and, 191–192; model of, 186–188; parameter constancy tests and, 192–193; variable coefficient models and, 186–193
time series: analysis of covariance (ANCOVA) and, 24–26; cross-sectionally dependent panel data and, 331, 339–340, 344–351, 362; discrete data and, 243, 257, 269–270; duration model and, 430–438; Durbin-Watson/Box-Pierce tests and, 345; dynamic models and, 81, 83, 91, 108, 110, 119, 124, 130, 133–135; dynamic systems and, 16, 369–372, 377–379, 383, 386–387, 390, 392–395; homogeneity and, 24–26; incomplete panel data and, 402–403, 405, 407, 411–419; independence test and, 344–351; linear regression models and, 24–26, 36, 67, 70, 73; maximum-likelihood estimators (MLE) and, 413–415; nonparametric panel data and, 462; panel data advantages and, 4, 6–9, 464–469; panel data issues and, 13, 469–473; pooling and, 411–418; static simultaneous-equations models and, 138; variable coefficient models and, 170–171, 184, 193–194, 201, 205, 207, 218–219, 223, 227
time-specific variables: cross-sectionally dependent panel data and, 327, 331, 337; dynamic models and, 80–81, 122–129; dynamic systems and, 393; panel data advantages and, 467–468; simple regression models and, 32–34, 58–64; static simultaneous-equations models and, 138; variable coefficient models and, 167
time-variant coefficients, 180–186
Tobin's $q$, 213, 217
Tobit models: cross-sectionally dependent panel data and, 348–349; dynamic, 317–326; panel data advantages and, 9; random individual effects and, 298–299; sample selection and, 298–299, 324–326; sample truncation and, 281, 288, 298–299, 317–324; simulation methods and, 448–449, 451; type II, 287, 298–299
transformed likelihood approach, 112–115
treatment effects: adjustment methods and, 354–359; counterfactuals prediction and, 361–363; cross-sectionally dependent panel data and, 352–368; definition of, 352–354; difference-in-difference method and, 360–361; example of, 363–368; panel data approach and, 359–368; regression discontinuity design and, 357–359
treatment groups, 353–357, 360–361
treatment on the treated (TT) effect, 352–353
triangular system: estimation and, 155–162; example of, 162–164; identification and, 153–155; maximum-likelihood estimators (MLE) and, 159–162; static simultaneous-equations models and, 152–164
$t$-statistic, 84, 124, 202t, 368, 388–390, 394, 410
two-stage least-squares estimators (2SLS), 104, 145, 147–148, 152, 154, 333

unconditional inference, 48–56, 178
unit root tests: augmented Dickey-Fuller (ADF) and, 389–390, 392; cross-sectionally correlated data and, 392–394; cross-sectionally independent data and, 387–392; dynamic systems and, 386–394; Lagrangian multiplier (LM) and, 391; Sargan-Bharvaga (SB) test statistic and, 394

U.S. Department of Labor, 1

variable coefficient models: aggregate vs. disaggregate analysis and, 217–220; Aitken estimators and, 179, 185, 196; asymptotics and, 121, 174–178, 185–186, 193, 196, 201n16, 208, 223–225; autocorellation and, 214t, 217, 224; autoregression and, 187, 207n18; Bayesian methods and, 174, 198–201, 204–205, 208–212; best linear unbiased estimators (BLUE) and, 183–185, 195; bias and, 173–175, 178, 180, 183, 184n10, 200, 210, 211t, 215, 220, 223; combination of two normal distributions and, 228–229; cross-sectionally dependent panel data and, 170–186, 221–222; degrees of freedom and, 168, 174, 177–179, 186, 193, 209, 215, 223; density and, 177–179, 191, 205, 209, 216; dependent variables and, 175–176, 190, 215t, 216t; dummy variable and, 168; dynamic, 206–212; efficiency and, 197, 224; error-component model and, 168n1, 174, 198, 200, 201n16; error terms and, 183–185, 206; exogenous variables and, 169, 177, 187, 193–196, 198, 206; firm investment expenditure and, 212–217; fixed-coefficient models and, 170–172, 177–178, 182–183, 196–206, 216–217; fixed constants (FE) and, 170, 193, 210, 212, 223; fixed effects and, 182, 223–224, 227; Frobenius norm and, 171; Gibbs sampler and, 209–210; Grunfeld investment function and, 175; heterogeneity and, 167–171, 178–180, 186, 196–197, 202, 204, 212, 214t, 215, 217–219; heteroscedasticity and, 170, 176, 179, 184–185, 224; Hildreth-Houck estimators and, 184n10; homogeneity and, 177–179, 202, 213, 215t, 217; individuals and, 167–180, 184–186, 193, 197–201, 204, 207–208, 214t, 215, 218, 220, 223, 227; infinity and, 174, 176–178, 185, 196, 209; instrumental variable (IV) estimators and, 172n3, 208, 221–223; Kalman filter and, 188–191; Lagrangians and, 162n14, 176–177, 185, 192; least-squares estimation and, 168, 171, 173–176, 184–185, 192, 195, 208, 210, 225–227; liquidity and, 169, 212–217; log-likelihood function and, 176n7, 191, 205; matrices and, 170–175, 177, 182–184, 187–195, 197, 201–203, 206–209, 219t, 223–227; maximum-likelihood estimators (MLE) and, 176–177, 184, 188, 191–192; mean square error and, 190, 201n16, 203, 204t; mixed fixed- and random- coefficient model and, 196–206; Monte Carlo studies and, 206, 209–210, 212, 227; multicollinearity and, 201, 213–215, 223, 226; normalization and, 199; null hypotheses and, 176–177, 186, 192–193, 212, 214; omitted variables and, 167, 198; ordinary least squares (OLS) estimation and, 211t; parameter variation assumptions and, 169; parametric models and, 227; parsimonious regression and, 167; pooling and, 169, 197, 203, 215, 217; probability and, 169, 198–199, 205, 209; random-coefficient models and, 169–170, 172–180, 183–186, 196–212, 215–227, 220–227; random effects and, 215; regression models and, 167–176, 185, 187, 189, 192–193, 196–197, 202t, 208, 213, 214t–216t, 220, 225–226; simple regression models and, 173 (*see also* simple regression models); statistics and, 173, 176–179, 184, 193, 202t, 204, 214, 215t, 217–218; stochastics and, 169, 193, 195, 197; time-evolving

variable coefficient models (*cont.*)
coefficients and, 186–193; time series and, 170–171, 184, 193–194, 201, 205, 207, 218–219, 223, 227; time-specific variables and, 167; time-variant coefficients and, 180–186; Tobin's *q* and, 213, 217; variable-intercept models and, 167–169, 178, 180, 198, 214, 215t, 217; variance-components models and, 184, 208; vectors and, 168–171, 177–179, 184, 186–187, 191–194, 196–198, 203, 206, 209, 222, 225–226
variable-intercept models: assumption of, 31–32; heterogeneity and, 15, 31–32; sample truncation and, 311; variable coefficient models and, 167–169, 178, 180, 198, 214, 215t, 217
variance-components models: discrete data and, 266; dynamic models and, 107; random effects and, 39–47; sample truncation and, 295; simple regression models and, 39–47, 52, 66; static simultaneous-equations models and, 144, 149, 152–153; variable coefficient models and, 184, 208
vector autoregressive (VAR) models: cross-sectionally dependent panel data and, 377–378; cross-sectionally independent data and, 377; dynamic systems and, 369–378, 380, 386, 393n7, 397–402; general method of moments (GMM) estimation and, 373–374; global, 378, 386; heterogeneity and, 377–378; homogeneity and, 370–377; maximum-likelihood estimators (MLE) and, 374–375; minimum-distance estimators (MDE) and, 375–377; model formulation and, 370–372; MonteCarlo studies and, 377; simultaneous equations and, 397–402
vectors, 15; autoregressive models and, 379–388; cointegrated models and, 379–386; cross-sectionally dependent panel data and, 327–329, 334n4, 336–340, 343, 356–357, 362; discrete data and, 230–233, 247, 254n16; dynamic models and, 80, 87, 93, 94n9, 97, 100, 112, 121, 124, 131; dynamic systems and, 369–386, 393, 396–399; eigenvectors, 148, 151, 334n4, 338–340; error correction and, 379–386, 396; homogeneity and, 17–18; incomplete panel data and, 404–405, 407, 409, 413–414, 420–422, 427–428; measurement errors and, 458–460; multilevel structures and, 453; nonparametric panel data and, 463; panel data advantages and, 6; panel data issues and, 10; panel quantile regression and, 446; sample truncation and, 281, 287, 289–290, 293, 306, 310, 313–314, 324–325; simple regression and, 34–36, 39–40, 58, 61, 70–71, 73, 75–79; simulation methods and, 451; static simultaneous-equations models and, 138–141, 143–144, 148, 151, 154, 160–161, 162n14, 165; variable coefficient models and, 168–171, 177–179, 184, 186–187, 191–194, 196–198, 203, 206, 209, 222, 225–226

Wald type tests, 9, 377, 385–386
within-group estimators: homogeneity and, 19; linear regression models and, 37, 43–44, 54
World Bank, 3

Yoplait, 267

Other titles in the series (*continued from page iii*)

Donald P. Jacobs, Ehud Kalai, and Morton I. Kamien, Editors, *Frontiers of research in economic theory: The Nancy L. Schwartz Memorial Lectures*, 1983–1997, 9780521632225, 9780521635387

A. Colin Cameron and Pravin K. Trivedi, *Regression analysis of count data*, 9780521632010, 9780521635677

Steinar Strom, Editor, *Econometrics and economic theory in the 20th century: The Ragnar Frisch Centennial Symposium*, 9780521633239, 9780521633659

Eric Ghysels, Norman R. Swanson, and Mark Watson, Editors, *Essays in econometrics: Collected papers of Clive W. J. Granger (Volume I)*, 9780521772976, 9780521774963

Eric Ghysels, Norman R. Swanson, and Mark Watson, Editors, *Essays in econometrics: Collected papers of Clive W. J. Granger (Volume II)*, 9780521792073, 9780521796491

Cheng Hsiao, *Analysis of panel data*, second edition, 9780521818551, 9780521522717

Mathias Dewatripont, Lars Peter Hansen, and Stephen J. Turnovsky, Editors, *Advances in economics and econometrics – Eighth World Congress (Volume I)*, 9780521818728, 9780521524117

Mathias Dewatripont, Lars Peter Hansen, and Stephen J. Turnovsky, Editors, *Advances in economics and econometrics – Eighth World Congress (Volume II)*, 9780521818735, 9780521524124

Mathias Dewatripont, Lars Peter Hansen, and Stephen J. Turnovsky, Editors, *Advances in economics and econometrics – Eighth World Congress (Volume III)*, 9780521818742, 9780521524131

Roger Koenker, *Quantile regression*, 9780521845731, 9780521608275

Charles Blackorby, Walter Bossert, and David Donaldson, *Population issues in social choice theory, welfare economics, and ethics*, 9780521825511, 9780521532587

John E. Roemer, *Democracy, education, and equality*, 9780521846653, 9780521609135

Richard Blundell, Whitney K. Newey, and Thorsten Persson, *Advances in economics and econometrics – Ninth World Congress (Volume I)*, 9780521871525, 9780521692083

Richard Blundell, Whitney K. Newey, and Thorsten Persson, *Advances in economics and econometrics – Ninth World Congress (Volume I)*, 9780521871532, 9780521692090

Richard Blundell, Whitney K. Newey, and Thorsten Persson, *Advances in economics and econometrics – Ninth World Congress (Volume I)*, 9780521871549, 9780521692106

Fernando Vega-Redondo, *Complex social networks*, 9780521857406, 9780521674096

Itzhak Gilboa, *Theory of decision under uncertainty*, 9780521517324, 9780521741231

Krislert Samphantharak and Robert M. Townsend, *Households as corporate firms: an analysis of household finance using integrated household surveys and corporate financial accounting*, 9780521195829, 9780521124164

Rakesh Vohra, *Mechanism design: A linear programming approach*, 9781107004368, 9780521179461

Daron Acemoglu, Manuel Arellano, Eddie Dekel, *Advances in economics and econometrics – Tenth World Congress (Volume I)*, 9781107016040/9781107638105

Daron Acemoglu, Manuel Arellano, Eddie Dekel, *Advances in economics and econometrics – Tenth World Congress (Volume II)*, 9781107016057/9781107674165

Daron Acemoglu, Manuel Arellano, Eddie Dekel, *Advances in economics and econometrics – Tenth World Congress (Volume III)*, 9781107016064/9781107627314

Andrew Harvey, *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial Time Series*, 9781107034723/ 9781107630024

A. Colin Cameron and Pravin K. Trivedi, *Regression analysis of count data, second edition* 9781107014169, 9781107667273