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. 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
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
Contents

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
xii Contents

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
Contents

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
## Contents

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  
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
   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 non-linear 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 assistance.
typing of various drafts of the monograph; R. Matzkin and two referees; and
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 Fig-
ures 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...
Preface to the Second Edition


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.