Analysis of Panel Data

Second Edition

CHENG HSIAO

University of Southern California

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2003

Printed in the United Kingdom at the University Press, Cambridge

Typeface Times Roman 10/12 pt. *System* LaTeX 2ε [TB]

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

Library of Congress Cataloging in Publication Data

Hsiao, Cheng, 1943–
Analysis of panel data / Cheng Hsiao. – 2nd ed.
p. cm. – (Econometric Society monographs ; no. 34)
Includes bibliographical references and index.
ISBN 0-521-81855-9 – ISBN 0-521-52271-4 (pb.)
1. Econometrics. 2. Panel analysis. 3. Analysis of variance. I. Title. II. Series.
HB139 .H75 2002
330′.01′5195 – dc21 2002023348

ISBN 0 521 81855 9 hardback
ISBN 0 521 52271 4 paperback

Contents

Preface to the Second Edition *page* xiii
Preface to the First Edition xv

Chapter 1. Introduction 1
1.1 Advantages of Panel Data 1
1.2 Issues Involved in Utilizing Panel Data 8
1.2.1 Heterogeneity Bias 8
1.2.2 Selectivity Bias 9
1.3 Outline of the Monograph 11

Chapter 2. Analysis of Covariance 14
2.1 Introduction 14
2.2 Analysis of Covariance 15
2.3 An Example 21

Chapter 3. Simple Regression with Variable Intercepts 27
3.1 Introduction 27
3.2 Fixed-Effects Models: Least-Squares Dummy-Variable Approach 30
3.3 Random-Effects Models: Estimation of Variance-Components Models 34
3.3.1 Covariance Estimation 35
3.3.2 Generalized-Least-Squares Estimation 35
3.3.3 Maximum Likelihood Estimation 39
3.4 Fixed Effects or Random Effects 41
3.4.1 An Example 41
3.4.2 Conditional Inference or Unconditional (Marginal) Inference 43
3.4.2.a Mundlak's Formulation 44
3.4.2.b Conditional and Unconditional Inferences in the Presence or Absence of Correlation between Individual Effects and Attributes 46

3.5 Tests for Misspecification 49
3.6 Models with Specific Variables and Both Individual- and Time-Specific Effects 51
3.6.1 Estimation of Models with Individual-Specific Variables 51
3.6.2 Estimation of Models with Both Individual and Time Effects 53
3.7 Heteroscedasticity 55
3.8 Models with Serially Correlated Errors 57
3.9 Models with Arbitrary Error Structure – Chamberlain π Approach 60
Appendix 3A: Consistency and Asymptotic Normality of the Minimum-Distance Estimator 65
Appendix 3B: Characteristic Vectors and the Inverse of the Variance–Covariance Matrix of a Three-Component Model 67

Chapter 4. Dynamic Models with Variable Intercepts 69
4.1 Introduction 69
4.2 The Covariance Estimator 71
4.3 Random-Effects Models 73
4.3.1 Bias in the OLS Estimator 73
4.3.2 Model Formulation 75
4.3.3 Estimation of Random-Effects Models 78
4.3.3.a Maximum Likelihood Estimator 78
4.3.3.b Generalized-Least-Squares Estimator 84
4.3.3.c Instrumental-Variable Estimator 85
4.3.3.d Generalized Method of Moments Estimator 86
4.3.4 Testing Some Maintained Hypotheses on Initial Conditions 90
4.3.5 Simulation Evidence 91
4.4 An Example 92
4.5 Fixed-Effects Models 95
4.5.1 Transformed Likelihood Approach 96
4.5.2 Minimum-Distance Estimator 98
4.5.3 Relations between the Likelihood-Based Estimator and the Generalized Method of Moments Estimator (GMM) 99
4.5.4 Random- versus Fixed-Effects Specification 101
4.6 Estimation of Dynamic Models with Arbitary Correlations in the Residuals 103
4.7 Fixed-Effects Vector Autoregressive Models 105
4.7.1 Model Formulation 105
4.7.2 Generalized Method of Moments (GMM) Estimation 107

4.7.3 (Transformed) Maximum Likelihood Estimator 109
4.7.4 Minimum-Distance Estimator 109
Appendix 4A: Derivation of the Asymptotic Covariance Matrix of the Feasible MDE 111

Chapter 5. Simultaneous-Equations Models 113
5.1 Introduction 113
5.2 Joint Generalized-Least-Squares Estimation Technique 116
5.3 Estimation of Structural Equations 119
5.3.1 Estimation of a Single Equation in the Structural Model 119
5.3.2 Estimation of the Complete Structural System 124
5.4 Triangular System 127
5.4.1 Identification 127
5.4.2 Estimation 129
5.4.2.a Instrumental-Variable Method 130
5.4.2.b Maximum-Likelihood Method 133
5.4.3 An Example 136
Appendix 5A 138

Chapter 6. Variable-Coefficient Models 141
6.1 Introduction 141
6.2 Coefficients That Vary over Cross-Sectional Units 143
6.2.1 Fixed-Coefficient Model 144
6.2.2 Random-Coefficient Model 144
6.2.2.a The Model 144
6.2.2.b Estimation 145
6.2.2.c Predicting Individual Coefficients 147
6.2.2.d Testing for Coefficient Variation 147
6.2.2.e Fixed or Random Coefficients 149
6.2.2.f An Example 150
6.3 Coefficients That Vary over Time and Cross-Sectional Units 151
6.3.1 The Model 151
6.3.2 Fixed-Coefficient Model 153
6.3.3 Random-Coefficient Model 153
6.4 Coefficients That Evolve over Time 156
6.4.1 The Model 156
6.4.2 Predicting $\boldsymbol{\beta}_t$ by the Kalman Filter 158
6.4.3 Maximum Likelihood Estimation 161
6.4.4 Tests for Parameter Constancy 162
6.5 Coefficients That Are Functions of Other Exogenous Variables 163
6.6 A Mixed Fixed- and Random-Coefficients Model 165
6.6.1 Model Formulation 165
6.6.2 A Bayes Solution 168
6.6.3 An Example 170

6.6.4 Random or Fixed Parameters 172
6.6.4.a An Example 172
6.6.4.b Model Selection 173
6.7 Dynamic Random-Coefficient Models 175
6.8 An Example – Liquidity Constraints and Firm Investment Expenditure 180
Appendix 6A: Combination of Two Normal Distributions 185

Chapter 7. Discrete Data 188
7.1 Introduction 188
7.2 Some Discrete-Response Models 188
7.3 Parametric Approach to Static Models with Heterogeneity 193
7.3.1 Fixed-Effects Models 194
7.3.1.a Maximum Likelihood Estimator 194
7.3.1.b Conditions for the Existence of a Consistent Estimator 195
7.3.1.c Some Monte Carlo Evidence 198
7.3.2 Random-Effects Models 199
7.4 Semiparametric Approach to Static Models 202
7.4.1 Maximum Score Estimator 203
7.4.2 A Root-N Consistent Semiparametric Estimator 205
7.5 Dynamic Models 206
7.5.1 The General Model 206
7.5.2 Initial Conditions 208
7.5.3 A Conditional Approach 211
7.5.4 State Dependence versus Heterogeneity 216
7.5.5 Two Examples 218
7.5.5.a Female Employment 218
7.5.5.b Household Brand Choices 221

Chapter 8. Truncated and Censored Data 225
8.1 Introduction 225
8.2 An Example – Nonrandomly Missing Data 234
8.2.1 Introduction 234
8.2.2 A Probability Model of Attrition and Selection Bias 235
8.2.3 Attrition in the Gary Income-Maintenance Experiment 238
8.3 Tobit Models with Random Individual Effects 240
8.4 Fixed-Effects Estimator 243
8.4.1 Pairwise Trimmed Least-Squares and Least-Absolute-Deviation Estimators for Truncated and Censored Regressions 243
8.4.1.a Truncated Regression 243
8.4.1.b Censored Regressions 249
8.4.2 A Semiparametric Two-Step Estimator for the Endogenously Determined Sample Selection Model 253

8.5 An Example: Housing Expenditure 255
8.6 Dynamic Tobit Models 259
8.6.1 Dynamic Censored Models 259
8.6.2 Dynamic Sample Selection Models 265

Chapter 9. Incomplete Panel Data 268
9.1 Estimating Distributed Lags in Short Panels 268
9.1.1 Introduction 268
9.1.2 Common Assumptions 270
9.1.3 Identification Using Prior Structure of the Process of the Exogenous Variable 271
9.1.4 Identification Using Prior Structure of the Lag Coefficients 275
9.1.5 Estimation and Testing 277
9.2 Rotating or Randomly Missing Data 279
9.3 Pseudopanels (or Repeated Cross-Sectional Data) 283
9.4 Pooling of a Single Cross-Sectional and a Single Time-Series Data Set 285
9.4.1 Introduction 285
9.4.2 The Likelihood Approach to Pooling Cross-Sectional and Time-Series Data 287
9.4.3 An Example 288

Chapter 10. Miscellaneous Topics 291
10.1 Simulation Methods 291
10.2 Panels with Large N and T 295
10.3 Unit-Root Tests 298
10.4 Data with Multilevel Structures 302
10.5 Errors of Measurement 304
10.6 Modeling Cross-Sectional Dependence 309

Chapter 11. A Summary View 311
11.1 Introduction 311
11.2 Benefits and Limitations of Panel Data 311
11.2.1 Increasing Degrees of Freedom and Lessening the Problem of Multicollinearity 311
11.2.2 Identification and Discrimination between Competing Hypotheses 312
11.2.3 Reducing Estimation Bias 313
11.2.3.a Omitted-Variable Bias 313
11.2.3.b Bias Induced by the Dynamic Structure of a Model 315
11.2.3.c Simultaneity Bias 316
11.2.3.d Bias Induced by Measurement Errors 316

11.2.4 Providing Micro Foundations for Aggregate Data Analysis 316
11.3 Efficiency of the Estimates 317

Notes 319
References 331
Author Index 353
Subject Index 359

CHAPTER 1

Introduction

1.1 ADVANTAGES OF PANEL DATA

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. For instance, in the U.S., 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 began in the mid-1960s. It contains five separate longitudinal data bases covering distinct segments of the labor force: men whose ages were 45 to 59 in 1966, young men 14 to 24 in 1966, women 30 to 44 in 1967, young women 14 to 24 in 1968, and youth of both sexes 14 to 21 in 1979. In 1986, the NLS expanded to include surveys of the children born to women who participated in the National Longitudinal Survey of Youth 1979. The list of variables surveyed is running into the thousands, with the emphasis on the supply side of the labor market. Table 1.1 summarizes the NLS survey groups, the sizes of the original samples, the span of years each group has been interviewed, and the current interview status of each group (for detail, see *NLS Handbook 2000*, U.S. Department of Labor, Bureau of Labor Statistics).

The PSID began with collection of annual economic information from a representative national sample of about 6,000 families and 15,000 individuals in 1968 and has continued to the present. The data set contains over 5,000 variables, including employment, income, and human-capital variables, as well as information on housing, travel to work, and mobility. 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 (2000); 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 Social Economic Panel (PSELL),

Table 1.1. *The NLS: Survey groups, sample sizes, interview years, and survey status*

Survey group	Age cohort	Birth year cohort	Original sample	Initial year/ latest year	Number of surveys	Number at last interview	Status
Older men	45–59	4/2/07–4/1/21	5,020	1966/1990	13	2,092[1]	Ended
Mature women	30–44	4/2/23–4/1/37	5,083	1967/1999	19	2,466[2]	Continuing
Young men	14–24	4/2/42–4/1/52	5,225	1966/1981	12	3,398	Ended
Young women	14–24	1944–1954	5,159	1968/1999	20	2,900[2]	Continuing
NLSY79	14–21	1957–1964	12,686[3]	1979/1998	18	8,399	Continuing
NLSY79 children	birth–14	—	—[4]	1986/1998	7	4,924	Continuing
NLSY79 young adults	15–22	—	—[4]	1994/1998	3	2,143	Continuing
NLSY97	12–16	1980–1984	8,984	1997/1999	3	8,386	Continuing

[1] Interviews in 1990 were also conducted with 2,206 widows or other next-of-kin of deceased respondents.

[2] Preliminary numbers.

[3] After dropping the military (in 1985) and economically disadvantaged non-Black, non-Hispanic oversamples (in 1991), the sample contains 9,964 respondents eligible for interview.

[4] The sizes of the NLSY79 children and young adult samples are dependent on the number of children born to female NLSY79 respondents, which is increasing over time.

Source: NLS Handbook, 2000, U.S. Department of Labor, Bureau of Labor Statistics.

the British Household Panel Survey (BHPS), etc. Starting in 1994, the National Data Collection Units (NDUs) 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 longitudinal surveys. For instance, the Manheim Innovation Panel (MIP) and the Manheim Innovation Panel – Service Sector (MIP-S) contain annual surveys of innovative activities (product innovations, expenditure on innovations, expenditure on R&D, factors hampering innovations, the stock of capital, wages, and skill structures of employees, etc.) of German firms with at least five employees in manufacturing and service sectors, started in 1993 and 1995, respectively. The survey methodology is closely related to the recommendations on innovation surveys manifested in the *OSLO Manual* of the 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 detail, see Janz et al. (2001)). Similarly, the European Community Household Panel (ECHP) is to represent the population of the European Union (EU) at the household and individual levels. The ECHP contains information on demographics, labor-force behavior, income, health, education and training, housing, migration, etc. The ECHP now covers 14 of the 15 countries, the exception being Sweden (Peracchi (2000)). Detailed statistics from the ECHP are published in Eurostat's reference data base New Cronos in three domains, namely health, housing, and income and living conditions (ILC).[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 of special importance to obtain original survey data to answer many significant and important questions. The World Bank has 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)).

Panel data sets for economic research possess several major advantages over conventional cross-sectional or time-series data sets (e.g., Hsiao (1985a, 1995, 2000)). Panel data usually give the researcher 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. More importantly, 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, consider the following example taken from Ben-Porath (1973): Suppose that a cross-sectional sample of married women is found to have an average yearly labor-force participation rate of 50 percent. At one extreme this might be interpreted as implying that each woman in a homogeneous population has a 50 percent chance of being in the labor force in any given year, while at the other extreme it might imply that 50 percent of the women in a heterogeneous population always work and 50 percent never work. In the first case, 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. In the second case, there is no turnover, and current information about work status is a perfect predictor of future work status. To discriminate between these two models, we need to utilize individual labor-force histories to estimate the probability of participation in different subintervals of the life cycle. This is possible only if we have sequential observations for a number of individuals.

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 or 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 mainly due to differences between union and nonunion firms or workers prior to unionism or postunion 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.

Whereas microdynamic and macrodynamic effects typically cannot be estimated using a cross-sectional data set, a single time-series data set usually cannot provide precise estimates of dynamic coefficients either. For instance, consider the estimation of a distributed-lag model:

$$y_t = \sum_{\tau=0}^{h} \beta_\tau x_{t-\tau} + u_t, \qquad t = 1, \ldots, T, \tag{1.1.1}$$

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; this allows 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.

Another example is that measurement errors can lead to unidentification of a model in the usual circumstance. However, the availability of multiple observations for a given individual or at a given time may allow a researcher to identify an otherwise unidentified model (e.g., Biørn (1992); Griliches and Hausman (1986); Hsiao (1991b); Hsiao and Taylor (1991); Wansbeek and Koning (1989)).

Besides the advantage that panel data allow us to construct and test more complicated behavioral models than purely cross-sectional or time-series data, the use of panel data also 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 the presence of omitted (mismeasured or unobserved) variables that are correlated 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:

$$y_{it} = \alpha^* + \boldsymbol{\beta}'\mathbf{x}_{it} + \boldsymbol{\rho}'\mathbf{z}_{it} + u_{it}, \qquad \begin{array}{l} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{array} \tag{1.1.2}$$

where $\mathbf{x}_{it}$ and $\mathbf{z}_{it}$ are $k_1 \times 1$ and $k_2 \times 1$ vectors of exogenous variables; α^*, $\boldsymbol{\beta}$, and $\boldsymbol{\rho}$ are 1×1, $k_1 \times 1$, 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 σ_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 α^*, $\boldsymbol{\beta}$, 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 effect of $\mathbf{z}$. 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

$$y_{it} - y_{i,t-1} = \boldsymbol{\beta}'(\mathbf{x}_{it} - \mathbf{x}_{i,t-1}) + (u_{it} - u_{i,t-1}), \qquad \begin{array}{l} i = 1, \ldots, N, \\ t = 2, \ldots, T. \end{array} \tag{1.1.3}$$

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

$$y_{it} - \bar{y}_t = \boldsymbol{\beta}'(\mathbf{x}_{it} - \bar{\mathbf{x}}_t) + (u_{it} - \bar{u}_t), \qquad \begin{array}{l} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{array} \tag{1.1.4}$$

where $\bar{y}_t = (1/N)\sum_{i=1}^N y_{it}$, $\bar{\mathbf{x}}_t = (1/N)\sum_{i=1}^N \mathbf{x}_{it}$, and $\bar{u}_t = (1/N)\sum_{i=1}^N u_{it}$. Least-squares regression of (1.1.3) or (1.1.4) 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 are correlated with $\mathbf{x}$ but are uncorrelated 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 (1.1.2), where y is the logarithm of hours worked, 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 is x_{it} correlated with z_i, but every economic variable that could act as an instrument for x_{it} (such as education) is also correlated 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 (1.1.2).

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); Mankiew, Romer, and Weil (1992)). Thus, a growth-rate regression model typically includes investment ratio, initial income, and measures of policy outcomes like 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. Since 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} is correlated with the regressors and the resulting cross-sectional parameter 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 (1.1.3).

Panel data involve two dimensions: a cross-sectional dimension N, and a time-series dimension T. We would expect that the computation of panel data estimators would be more complicated than the analysis of cross-section data alone (where $T = 1$) or time series data alone (where $N = 1$). However, in certain cases the availability of panel data can actually simplify the computation and inference. For instance, consider a dynamic Tobit model of the form

$$y_{it}^* = \gamma y_{i,t-1}^* + \beta x_{it} + \epsilon_{it} \tag{1.1.5}$$

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 γ and β 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 ϵ_{it} (Arellano, Bover, and Labeager (1999)).

Another example is the time-series analysis of nonstationary data. The large-sample approximation of the distributions of the least-squares or maximum likelihood estimators when $T \to \infty$ are no longer normally distributed if the data are nonstationary (e.g., Dickey and Fuller (1979, 1981); Phillips and Durlauf (1986)). Hence, the behavior of the usual test statistics will often have to be inferred through computer simulations. But if panel data are available, and observations among cross-sectional units are independent, then 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 (2000); Levin and Lin (1993); Pesaran, Shin, and Smith (1999), Phillips and Moon (1999, 2000); Quah (1994)).

Panel data also provide the possibility of generating more accurate predictions for individual outcomes than time-series data alone. 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, in addition to the information on that individual's behavior. Thus, a more accurate description of an individual's behavior can be obtained by pooling the data (e.g., Hsiao and Mountain (1994); Hsiao and Tahmiscioglu (1997); Hsiao et al. (1989); Hsiao, Applebe, and Dineen (1993)).

1.2 ISSUES INVOLVED IN UTILIZING PANEL DATA

1.2.1 Heterogeneity Bias

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, since the purpose of modeling is not to mimic the reality but is 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 $P(y \mid \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is an m-dimensional real vector, identical for all individuals at all times, may not be a realistic one. Ignoring the individual or time-specific effects that exist among cross-sectional or time-series units but are not captured by the included explanatory variables can lead to parameter heterogeneity in the model specification. Ignoring such heterogeneity could lead to inconsistent or meaningless estimates of interesting parameters. For example, consider a simple model postulated as

$$y_{it} = \alpha_i^* + \beta_i x_{it} + u_{it}, \qquad \begin{aligned} i &= 1, \ldots, N, \\ t &= 1, \ldots, T, \end{aligned} \tag{1.2.1}$$

where x is a scalar exogenous variable ($k_1 = 1$) and u_{it} is the error term with mean zero and constant variance σ_u^2. The parameters α_i^* and β_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:

$$y_{it} = \alpha^* + \beta x_{it} + u_{it}, \qquad \begin{aligned} i &= 1, \ldots, N, \\ t &= 1, \ldots, T. \end{aligned} \tag{1.2.2}$$

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 due to the assumption that $\alpha_i^* \neq \alpha_j^*$ and $\beta_i = \beta_j$. In these graphs, the broken-line ellipses 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 (1.2.2) 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 (1.2.2) 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 to nonsensical results because it would represent an average of coefficients that differ greatly across individuals. Nor does the case of Figure 1.5 make any sense in pooling, 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 pooling the data under homogeneity assumption 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).

1.2.2 Selectivity Bias

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. For example, 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)).

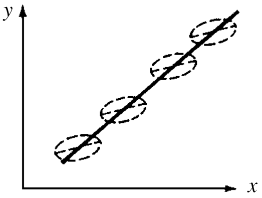

Fig. 1.1

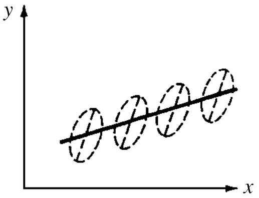

Fig. 1.2

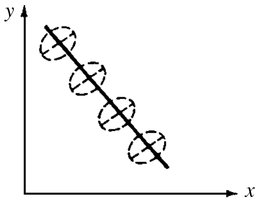

Fig. 1.3

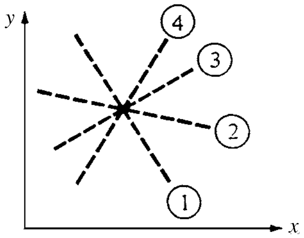

Fig. 1.4

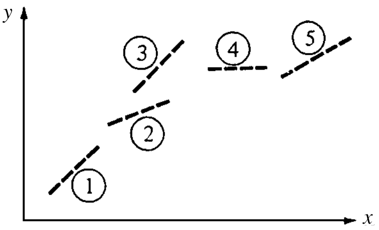

Fig. 1.5

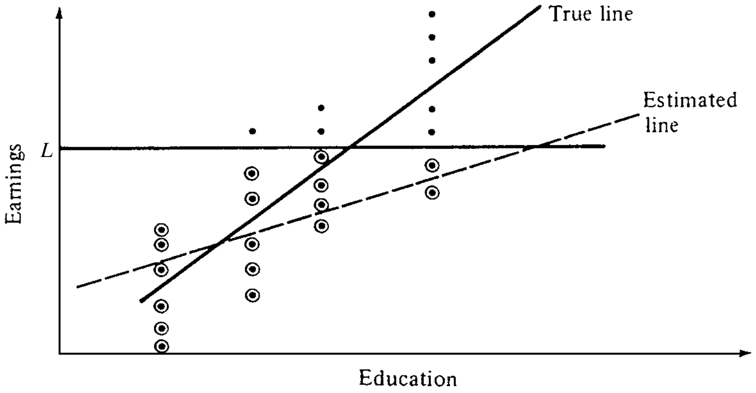

Fig. 1.6

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-squares 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

$$y_i = \boldsymbol{\beta}'\mathbf{x}_i + u_i, \qquad i = 1, \ldots, N, \tag{1.2.3}$$

where the disturbance term u_i is independently distributed with mean zero and variance σ_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:

$$\begin{aligned} y_i &= \boldsymbol{\beta}'\mathbf{x}_i + u_i \le L, \quad \text{included}, \\ y_i &= \boldsymbol{\beta}'\mathbf{x}_i + u_i > L, \quad \text{excluded}. \end{aligned} \tag{1.2.4}$$

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.

These examples demonstrate that despite the advantages panel data may possess, they are subject to their own potential experimental problems. It is only by taking proper account of selectivity and heterogeneity biases in the panel data that one can have confidence in the results obtained. The focus of this monograph will be on controlling for the effect of unobserved individual and/or time characteristics to draw proper inference about the characteristics of the population.

1.3 OUTLINE OF THE MONOGRAPH

Because the source of sample variation critically affects the formulation and estimation of many economic models, we shall first briefly review the classic analysis of covariance procedures in Chapter 2. We shall 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:

$$y_{it} = \alpha_i^* + \sum_{k=1}^{K} \beta_k x_{kit} + u_{it}, \qquad \begin{aligned} i &= 1, \ldots, N, \\ t &= 1, \ldots, T. \end{aligned} \tag{1.3.1}$$

2. Slope coefficients are constant, and the intercept varies over individuals and time:

$$y_{it} = \alpha_{it}^* + \sum_{k=1}^{K} \beta_k x_{kit} + u_{it}, \qquad \begin{array}{l} i = 1, \ldots, N, \\ t = 1, \ldots, T. \end{array} \tag{1.3.2}$$

3. All coefficients vary over individuals:

$$y_{it} = \alpha_{i}^* + \sum_{k=1}^{K} \beta_{ki} x_{kit} + u_{it}, \qquad \begin{array}{l} i = 1, \ldots, N, \\ t = 1, \ldots, T. \end{array} \tag{1.3.3}$$

4. All coefficients vary over time and individuals:

$$y_{it} = \alpha_{it}^* + \sum_{k=1}^{K} \beta_{kit} x_{kit} + u_{it}, \qquad \begin{array}{l} i = 1, \ldots, N, \\ t = 1, \ldots, T. \end{array} \tag{1.3.4}$$

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

Models with constant slopes and variable intercepts [such as (1.3.1) and (1.3.2)] are most widely used when analyzing panel data because they provide simple yet reasonably general alternatives to the assumption that parameters take values common to all agents at all times. We shall consequently devote the majority of this monograph to this type of model. Static models with variable intercepts will be discussed in Chapter 3, dynamic models in Chapter 4, simultaneous-equations models in Chapter 5, and discrete-data and sample selection models in Chapters 7 and 8, respectively. Basic issues in variable-coefficient models for linear models [such as (1.3.3) and (1.3.4)] will be discussed in Chapter 6. Chapter 9 discusses issues of incomplete panel models, such as estimating distributed-lag models in short panels, rotating samples, pooling of a series of independent cross sections (pseudopanels), and the pooling of data on a single cross section and a single time series. Miscellaneous topics such as simulation methods, measurement errors, panels with large N and large T, unit-root tests, cross-sectional dependence, and multilevel panels will be discussed in Chapter 10. A summary view of the issues involved in utilizing panel data will be presented in Chapter 11.

The challenge of panel data analysis has been, and will continue to be, the best way to formulate statistical models for inference motivated and shaped by substantive problems compatible with our understanding of the processes generating the data. The goal of this monograph is to summarize previous work in such a way as to provide the reader with the basic tools for analyzing and drawing inferences from panel data. Analyses of several important and advanced