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 developed and developing countries. In the United States, two of the longest-running panel data sets are the National Longitudinal Surveys of Labor Market Experience (NLS) and the Panel Study of Income Dynamics (PSID).

The NLS are a set of surveys sponsored by the Bureau of Labor Statistics (BLS). These surveys have gathered information at multiple points in time on several groups of men and women regarding their labor market experiences and other significant events in life such as schooling, employment, marriage, fertility, training, child care, health, and drug and alcohol use. The original four cohorts were men aged 45–59 in 1966, young men aged 14–24 in 1966, women aged 30–44 in 1967, and young women aged 14–24 in 1968. Table 1.1 summarizes the size of each group and the span of years each group has been interviewed for these original samples, as well as for 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 14–22 years old. These individuals were interviewed annually through 1994 and are currently interviewed on a biennial basis (NLS79). In 1986, the NLS began surveying children born to women who participated in the National Longitudinal Survey of Youth 1979 (NLS79 Children and Young Adult). In addition to all the mothers, information from the NLS79, the child survey includes additional demographic and development information. For children aged 10 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 12–16 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, as well as data on the youths, family and community background. It documents the transition from school to work and from adolescence to adulthood. NLS data and documentation are available online at the NLS Product Availability Center at www.nlsinfo.org. Questions about NLS data can be answered by contacting usersvc@chrr.osu.edu.

The PSID began in 1968 with the collection of annual economic information from a representative national sample of about 6,000 families and 15,000 individuals and their descendants, and this has continued to the present. The PSID gathers data on the family

- ? F > < ? 2 5 6 9 C C @ B + ! ! F F F 4 2 = 3 A : 5 8 6 , 2 = 3 A : 5 8 6 ,
9 C C @ B + ! ! F F F 4 2 = 3 A : 5 8 6 # " # (! * ( ) # " " * " ' ((& ' " " $

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).

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 over 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, several other panel data sets 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 – including the Netherlands Socio-Economic Panel (SEP), the German Social Economics Panel (GSOEP), Luxembourg's Social Economic Panel (PSELL), and the British Household Panel Survey (BHPS). 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 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 like 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. The survey methodology is closely related to the recommendations on innovation surveys manifested in the Oslo Manual of the OECD and Eurostat, thereby yielding internationally comparable data on innovation activities of German firms. The 1993 and 1997 surveys also became 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) represents 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, and migration. With the

[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).

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 database New Cronos in three domains, namely health, housing, and ILC (income and living conditions).[2]

In Australia, the Household, Income and Labour Dynamics in Australia (HILDA) Survey, administered by the Melbourne Institute of Applied Economic and Social Research at the University of Melbourne, under the support of the Australian government's Department of Social Services, has followed more than 17,000 Australians annually starting in 2001. It collects information on many aspects of life in Australia, including household and family relationships, income and employment, and health and education. Data from the HILDA Survey are available to researchers through the National Centre for Longitudinal Data Dataverse (Australian Government Department of Social Services).

Panel data have also become increasingly available in developing countries. These countries may not have a long tradition of statistical collection. Therefore, it is of special importance 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 nongovernment organization (NGO), Investing in Children and their Societies (ICS), Africa, collaborated with the Kenya Ministry of Health to carry out a Primary School Deworming Project (PDSP). The project took place in the Busia district, a poor and densely settled farming region in western Kenya. The 75 project schools included nearly all rural primary schools in this area, with over 30,000 enrolled pupils between ages of 6–18 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; www.rand.org/labor/aging/dataprod/, http://hrsonline.isr.umich.edu/), followed by the English Longitudinal Study of Ageing (ELSA; www.ifs.org.uk/elsa/), and the Survey of Health, Ageing and Retirement (SHARE; www.share-project.org/), which 28 European countries and Israel have joined. Other countries are 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 lots of information about people aged (approximately) 50 years 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, due to the highly synchronized questionnaires across a large number of countries, it becomes possible to study the role of institutional factors, like pension systems, retirement laws, and social

[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: https://ec.europa.eu/eurostat/web/microdata/european-community-household-panel.

security plans, on labor force participation and retirement (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 the following:

1.2.1 More Accurate Inference of Model Parameters

Empirical researchers often encounter shortages of degree of freedom and multicollinearity. To narrow the gap between the requirement of a model and the information in the sample, ad hoc restrictions are often imposed (e.g., Almon 1965; Cagan 1956; Hsiao et al. 1995). Panel data usually contain a large number of data points, increasing the degrees of freedom and reducing the collinearity among explanatory variables; hence, the data are more capable of improving the efficiency of econometric estimates.

1.2.2 Greater Capacity for Constructing More Realistic Behavioral Hypotheses

By blending inter-individual differences with intra-individual 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 alone. For instance, a typical assumption for the analysis using cross-sectional data is that individuals with the same conditional variables, $\boldsymbol{x}$, have the same expected value, $E(y_i|\boldsymbol{x}_i = \boldsymbol{a}) = E(y_j|\boldsymbol{x}_j = \boldsymbol{a})$. Under this assumption, if a cross-sectional sample of married women is found to have an average yearly labor force participation rate of 50%, it would imply that each woman in a homogeneous population has a 50% 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 frequent job turnover would be expected with an average job duration of two years. However, cross-sectional samples may not be randomly drawn from a common population. They may be drawn from a heterogeneous population as illustrated in the article by Ben-Porath (1973), in which 50% of the women came from the population that always work and 50% came from the population that never work. 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 enables onei to distinguish between these two groups. The sequential observations for a number of individuals allow 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 also seen 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 mainly due 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 inter-individual 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 worker's quality 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.

1.2.3 Uncovering Dynamic Relationships

Because of institutional or technological rigidities or inertia in human behavior, "economic behavior is inherently dynamic" (Nerlove 2002). 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:

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

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 could 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 inter-individual 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 (e.g., Pakes and Griliches 1984).

1.2.4 Controlling the Impact of Omitted Variables (or unobserved Individual or Time Heterogeneity)

Economists typically are interested in identifying causal relationships of a few variables they consider important. However, factors affecting the outcomes of a variable are numerous. A common and very useful approach is to consider the aggregate of the impacts of all omitted variables that affect the outcomes as the error terms that follow certain probability laws. In a regression model, such an error term is typically assumed to be uncorrelated with the included explanatory variables. However, if the error is correlated with the regressors, then regressing the dependent variable on the regressors could be biased and lead to 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 are correlated with explanatory variables. Panel data, providing information on both the intertemporal dynamics and the individuality of the entities being investigated, are better able to control the effects of missing or unobserved variables. For instance, consider a simple regression model:

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

where $\boldsymbol{x}_{it}$ and 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 independent of $\boldsymbol{x}_{it}$ and z_{it} and 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 $\boldsymbol{x}_{it}$ and z_{it} yields unbiased and consistent estimators of α^*, $\boldsymbol{\beta}$, and $\boldsymbol{\rho}$. Now suppose that z_{it} values are unobservable, and the covariances between $\boldsymbol{x}_{it}$ and z_{it} are nonzero. Then the least squares regression coefficients of y_{it} on $\boldsymbol{x}_{it}$ are biased. However, if repeated observations for a group of individuals are available, they may allow us to get rid of the effect of z. For example, if $z_{it} = z_i$ for all t (i.e., 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

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

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

$$y_{it} - \bar{y}_t = \boldsymbol{\beta}'(\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_t) + (u_{it} - \bar{u}_t), \qquad \begin{array}{l} i = 1, \ldots, N, \\ t = 1, \ldots, T, \end{array} \tag{1.2.4}$$

where $\bar{y}_t = (1/N)\sum_{i=1}^N y_{it}, \bar{\boldsymbol{x}}_t = (1/N)\sum_{i=1}^N \boldsymbol{x}_{it}$ and $\bar{u}_t = (1/N)\sum_{i=1}^N u_{it}$. 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 ($z_{it} = z_i$), or a single time series data set ($N = 1$) for the latter case ($z_{it} = z_t$), such transformations cannot be performed. We cannot get consistent estimates of $\boldsymbol{\beta}$ unless there exist instruments that are correlated with $\boldsymbol{x}$ but are uncorrelated with 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, $\boldsymbol{x}$ 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 $\boldsymbol{x}_{it}$ correlated with z_i, but every economic variable that could act as an instrument for $\boldsymbol{x}_{it}$ (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 (1.2.2).

[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.

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. Since a country that is less efficient is also more likely to have a lower investment rate or school enrollment, one can easily imagine that z_{i0} is correlated with the regressors and that 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).

1.2.5 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).

1.2.6 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 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 micro-equations (e.g., 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.

1.2.7 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:

Analysis of Nonstationary Time Series. 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, 81; 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).

Measurement Errors. Measurement errors can lead to under-identification of an econometric model (e.g., Aigner et al. 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).

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 that value is 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 would be 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 $y_{i,t-1}$. 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 Labeaga 1999).

1.3 CHALLENGES TO PANEL DATA ANALYSIS

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, since 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 the 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 realistic.

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 $\boldsymbol{x}_{it}$ or interact with $\boldsymbol{x}_{it}$. A linear regression model for y_{it} to take account of both possibilities with a single explanatory variable x_{it} could be postulated as

$$y_{it} = \alpha_i^* + \beta_i x_{it} + u_{it}, \qquad \begin{array}{l} i = 1, \dots, N, \\ t = 1, \dots, T, \end{array} \tag{1.3.1}$$

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:

$$y_{it} = \alpha^* + \beta x_{it} + v_{it}, \qquad \begin{array}{l} i = 1, \dots, N, \\ t = 1, \dots, T. \end{array} \tag{1.3.2}$$

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

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) due to the fact that $\alpha_i^* \neq \alpha_j^*$ 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 to nonsensical results because it 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.

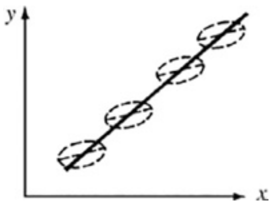

Figure 1.1. Homogeneous slope, heterogeneous intercept, a

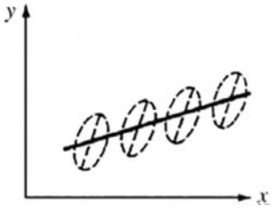

Figure 1.2. Homogeneous slope, heterogeneous intercept, b

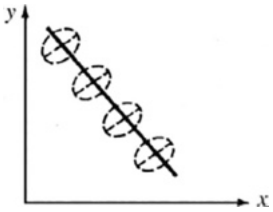

Figure 1.3. Homogeneous slope, heterogeneous intercept, c

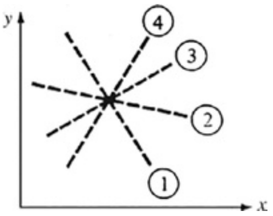

Figure 1.4. Heterogeneous slope and intercept, a

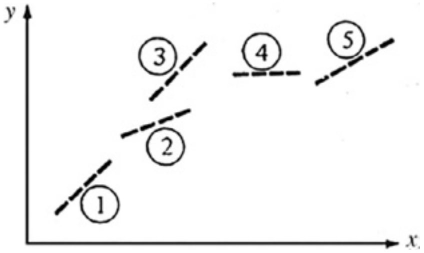

Figure 1.5. Heterogeneous slope and intercept, b

These are some of the likely biases when unobserved heterogeneities among cross-sectional units are ignored. More elaborate patterns than those depicted here are, of course, likely to occur (e.g., Chesher and Lancaster 1983; Kuh 1963). If the conditional density of y_{it} given $\boldsymbol{x}_{it}$ varies across i and over t, $f_{it}(y_{it} \mid \boldsymbol{x}_{it})$, 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 unobserved heterogeneity across individuals and over time that are not captured by the included conditional variables, $\boldsymbol{x}$.

A popular approach to control the unobserved heterogeneity is to let the parameters characterizing the conditional distribution of y_{it} given $\boldsymbol{x}_{it}$ to vary across i and over t, $f(y_{it} \mid \boldsymbol{x}_{it}, \boldsymbol{\theta}_{it})$. 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 $\boldsymbol{x}_{it}, \boldsymbol{\theta}_{it}$ 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. The individual-time varying $\boldsymbol{\gamma}_{it}$ can take a variety of forms. When $\boldsymbol{\gamma}_{it}$ are treated as random variables, it is called the *random-effects* model (e.g., Balestra and Nerlove 1966). When $\boldsymbol{\gamma}_{it}$ is treated as a fixed unknown constant, it is called the *fixed-effects* model (e.g., Kuh 1963). For $\boldsymbol{\gamma}_{it}$ to be identifiable by panel data, often specific assumptions are made. The focus of panel data analysis is on how to control the impact of unobserved heterogeneity to obtain valid inference on the common effects, $\boldsymbol{\beta}$.

1.3.2 Incidental Parameters and Multidimensional Statistics

When either N or T are fixed, standard one-dimensional asymptotics can be applied to derive large sample results. 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. To derive consistent estimators of the structural parameters $\boldsymbol{\beta}$, one either has to explore the specific structure of a model to find appropriate transformation to get rid of incidental parameters or to reparametrize the model so the information matrix of the incidental parameters are uncorrelated with the structural parameters (e.g., Lancaster 2001).

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 N and

T are of similar magnitude or increase at the same rate, Phillips and Moon (2000) show that naively applying one-dimensional asymptotics, followed by expanding the sample size in another dimension, could lead to misleading inferences. The multidimensional asymptotics are quite complex. Moreover, multidimensional asymptotic results may shed little insight on finite samples of high dimensional data. For example, Bai and Saranadasa (1996) showed that when testing the difference of means of two high dimensional populations, Dempster's (1958) non-exact test is more powerful than Hotelling's (1931) T^2-test, even though the latter is well defined.

1.3.3 Modeling Cross-Sectional Dependence

When cross-sectional units are correlated, it raises the issue of how to model cross-sectional dependence and what the relationship is between sample and population. Contrary to time series observations, there is no natural ordering of how events are evolved across cross-sectional units. Moreover, the traditional assumption that members in the sample and members in the population are independent may not hold. For instance, in the network literature (e.g., Brock and Durlauf 2001, 2007; Graham 2016), the probability of forming the link between units i and j is a function of the numbers of neighbors or cumulative number of links that i and j have in common in period t. If the members in the sample are not independent of members in the population, the measures of the number of neighbors or cumulative number of links based on the sample create measurement error problems (e.g., Aigner et al. 1984).

1.3.4 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 features 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 was stopped, only 2,092 individuals remained. 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 1976, 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 squares estimates. We assume that in the population, the relationship between earnings (y) and exogenous variables ($\boldsymbol{x}$), including education, intelligence, and so forth, is of the form

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

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 having earnings of less than L, the selection criterion for families considered for inclusion in the experiment can be stated as follows:

- ? F > < ? 2 5 6 5 C C @ B + ! ! F F F 4 2 = 3 A 6 5 > 3 B G < E 2 > 4 2 A 0 C 2 % 6 1 > G E $ ' A $ ' : C 2 C B D # 3) ; + 6 % ' C + ' C & C 9 6 , 2 = 3 A : 5 8 6 ,
9 C C @ B + ! ! F F F 4 2 = 3 A : 5 8 6 C C @ B + ! ! 4 5 A 6 ! C A 8 = B " # " # (! * ( ) # " " * " ' ((& ' " " $

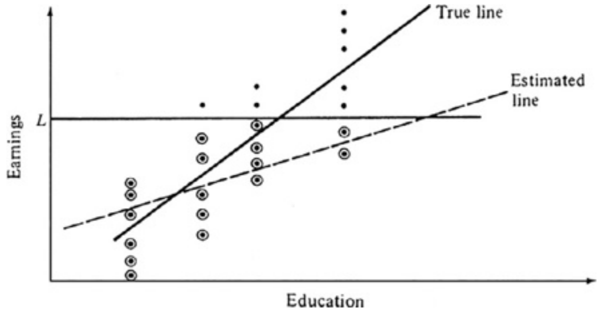

Figure 1.6. Earning and Education

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

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 focus of this revision remains on identifying causal relations of a few variables that an economist or econometrician considers important. The monograph takes a pedagogical approach. We hope to convey the intuitive ideas behind the methods discussed rather than providing rigorous derivations. We examine the material taking a simple (or specific)-to-more complex manner (or general approach). We first consider controlling the impact of unobserved heterogeneity to draw inferences about certain characteristics of the population that are of interest to an investigator by assuming that the error of an equation, v_{it}, can be decomposed into the sum of individual-specific effects, α_i, that vary over i but stay constant over t; the time-specific effects, λ_t, that vary over t but stay constant over i; and the component, u_{it}, that varies over i and t,

$$v_{it} = \alpha_i + \lambda_t + u_{it}, \qquad \begin{aligned} i &= 1, \ldots, N,\\ t &= 1, \ldots, T, \end{aligned} \tag{1.4.1}$$

because it remains the bulk of panel data literature so far.

Chapter 2 considers a simple linear regressions model where the conditional covariates, $\boldsymbol{x}_{it}$, can be considered strictly exogenous with regard to u_{it}, $E(u_{it}|\boldsymbol{x}_{is}) = 0$, but not

[4] For a formal treatment of this, see Chapter 7.

necessarily between x_{it} and α_i or λ_t, for all s. Fixed-versus random-effects inference (covariance transformation versus generalized least squares) and misspecification tests are discussed.

Chapter 3 considers dynamic panel data where the strict exogeneity assumption of the conditional covariate, $\boldsymbol{x}_{it}$, no longer holds. Inference when either the dimension of N or T is fixed while the other dimension goes to infinity or when both N and T are large in light of the likelihood approach or methods of moment approach or generalized method of moments (GMM) are considered.

Chapter 4 discusses issues when y_{it} and $\boldsymbol{x}_{it}$ are jointly determined (i.e., $\boldsymbol{x}_{it}$ and u_{it} are correlated). Efficient inference for both the reduced form and structural form limited information (single equation) and full information approach is discussed. Identification and estimation of a triangular system are also considered.

Chapter 5 considers a dynamic system. The stationery panel vector autoregressive model and dynamic simultaneous-equations model as well as testing for unit roots and estimation of cointegrated and error-correction models are discussed.

Chapter 6 considers random-effects and fixed-effects inference conditional on minimum sufficient statistics for static and dynamics qualitative choice models. Issues of true and spurious state dependence with or without structural changes (e.g., Lucas 1976) are discussed. Bias reduced estimator and alternative modeling without incidental parameters are also introduced.

Chapter 7 discusses limited dependent variable and sample selection models. The likelihood approach, method of moments approach, and semiparametric approach are considered. The symmetrically trimmed method for implementing moments estimators is introduced.

Chapter 8 considers some widely used nonlinear models, including duration models, count data models and nonparametric models. Complexities of inference for nonlinear panel data are discussed in light of these models.

Chapter 9 considers some miscellaneous topics that do not fit into the specific themes of Chapters 2–8. Panel quantile regression, simulation methods to obtain the method of moments or the likelihood estimators, data with multilevel structures, measurement errors, and issues of missing or rotating or repeated cross-sectional data are briefly discussed. An example of estimating distributed lag models in short panels is provided to illustrate how econometricians identify a model relying on ad hoc assumptions when there is insufficient sample information to identify the model of interest. Discretizing unobserved (continuous) heterogeneity into a finite number of unobserved state variables as a tractable alternative is also discussed.

More complex approaches to control the impact of unobserved heterogeneity to obtain valid inference on structural parameters are considered over the next five chapters.

Chapter 10 considers interactive-effects models. The multiplicative approach (or factor approach) to characterize the impact of omitted variables is more general than the additive approach. It is also a parsimonious way to allow for a complex correlation structure of the error term. However, a simple linear transformation to get rid of the factor structure does not exist. The chapter discusses quasi-maximum likelihood, quasi-difference, common correlated effects, and transformed estimators for both static and dynamic models, as well as factor dimension determination.

Chapter 11 looks at spatial models and tests for cross-sectional dependence. It introduces the notion of weak and strong cross-sectional dependence. Widely used spatial models in regional science and international economics are discussed in light of the panel data approach to model cross-sectional dependence. Various tests of cross-sectional dependence are also considered.

Chapter 12 considers program evaluation using panel data. Issues of measuring treatment effects are discussed, as are the pros and cons of panel data parametric, semiparametric, and nonparametric approach versus cross-sectional data. Issues of measuring the aggregate (or average) and disaggregate treatment effects as well as simulating or ranking different policy options and aggregation are considered.

Chapter 13 focuses on variable coefficient models. Generalizing unobserved heterogeneity over i and t to the regression coefficients are covered. The fixed-effects and random-effects over i and t is considered. To pool or not to pool the data in light of parameter heterogeneity is also discussed.

Chapter 14 considers big data analytics. Challenges of big data to panel data analysis in light of the traditional role of econometrics to verify economic theory and makes predictions are discussed. Basic artificial intelligence algorithms to process high-dimensional, high-volume and high-variety data are introduced. Inference with high-dimensional data and inference with low-dimensional parameters in the presence of high-dimensional data are discussed. Pros and cons on data-based prediction and causal-based prediction and issues of structural breaks are considered. Issues of aggregating micro data and combining different data sources are also raised.

Finally, I should remind readers that although this manuscript is focused on obtaining valid causal inference, all inferential procedures are based on some maintained hypotheses. Maintained hypotheses are typically not testable. Statistical analysis is not a proof of reality but tries to provide information in the data. We should explore every possible consequence when the maintained hypothesis does not hold and be humble in reporting our findings. Otherwise, it will be like what Mark Twain said:

There are lies.
There are damned lies.
There is statistics!

- ? F > < ? 2 5 6 5 C C @ B + ! ! F F F 4 2 = 3 A 6 5 > B G < E 2 > 4 2 A 0 C 2 % 6 1 > G E $ ' A $ ' : C 2 G D # 3) ; + 6 % 4 ' C + ' C & C 9 6 , 2 = 3 A : 5 8 6 ,
9 C C @ B + ! ! F F F 4 2 = 3 A : 5 8 6 C @ B 8 ! 4 5 A 6 ! C 6 A 8 = B " # " # (! * ( ) # " " * " ' ((& ' " " $