CHAPTER 1

Introduction

1.1 INTRODUCTION

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

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

https://doi.org/10.1017/CBO9781139839327.004 Published online by Cambridge University Press

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

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

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

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

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

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

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

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

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

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

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

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

1.2 ADVANTAGES OF PANEL DATA

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

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

https://doi.org/10.1017/CBO9781139839327.004 Published online by Cambridge University Press

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

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

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

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

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

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

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

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

where $\mathbf{x}_{it}$ and $\mathbf{z}_{it}$ are $k_1 \times 1$ and $k_2 \times 1$ vectors of exogenous variables; 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

https://doi.org/10.1017/CBO9781139839327.004 Published online by Cambridge University Press

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 effects of $\mathbf{z}$ through a linear transformation. For example, if $\mathbf{z}_{it} = \mathbf{z}_i$ for all t (i.e., $\mathbf{z}$ values stay constant through time for a given individual but vary across individuals), we can take the first difference of individual observations over time and obtain

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

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

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

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

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

https://doi.org/10.1017/CBO9781139839327.004 Published online by Cambridge University Press

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

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

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

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

https://doi.org/10.1017/CBO9781139839327.004 Published online by Cambridge University Press

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

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

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

https://doi.org/10.1017/CBO9781139839327.004 Published online by Cambridge University Press

$y_{i,t-1} > 0$ 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 Labeaga 1999).

1.3 ISSUES INVOLVED IN UTILIZING PANEL DATA

1.3.1 Unobserved Heterogeneity across Individuals and over Time

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

https://doi.org/10.1017/CBO9781139839327.004 Published online by Cambridge University Press

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

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

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

where u_{it} is the error term, uncorrelated with x, with mean zero and constant variance 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:

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

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

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

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

https://doi.org/10.1017/CBO9781139839327.004 Published online by Cambridge University Press

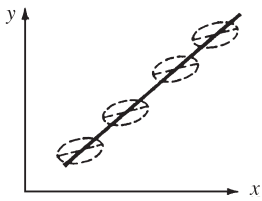

Fig. 1.1

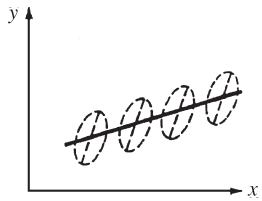

Fig. 1.2

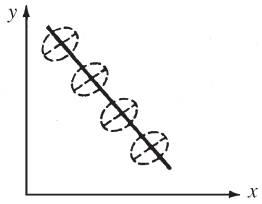

Fig. 1.3

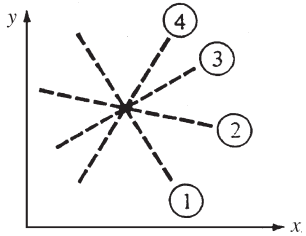

Fig. 1.4

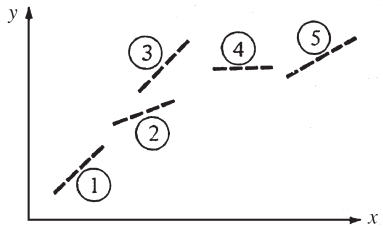

Fig. 1.5

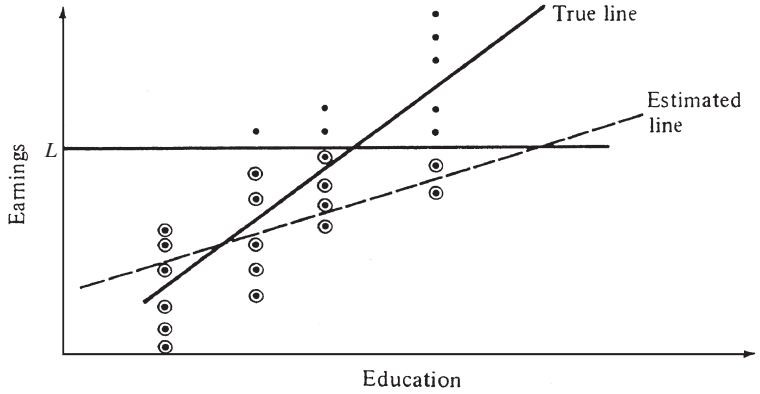

Fig. 1.6

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

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

https://doi.org/10.1017/CBO9781139839327.004 Published online by Cambridge University Press

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

1.3.2 Incidental Parameters and Multidimensional Statistics

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

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

1.3.3 Sample Attrition

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

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

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

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

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

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

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

1.4 OUTLINE OF THE MONOGRAPH

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

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

https://doi.org/10.1017/CBO9781139839327.004 Published online by Cambridge University Press

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

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

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

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

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