

1

Introduction

The increased availability of panel data from household surveys has been one of the most important developments in applied social research in the last thirty years.

Fitzgerald, Gottschalk and Moffitt (1998, p. 252)

1.1 PANEL DATA: SOME EXAMPLES

In this book, the term “panel data” refers to the pooling of observations on a cross-section of households, countries, firms, etc. over several time periods. This can be achieved by surveying a number of households or individuals and following them over time. The data collected from surveying individuals are known as *micro panels* and are collected for a large number N of individuals (usually in the hundreds or thousands) over a short time period T (which may vary from a minimum of two years to a maximum rarely exceeding 10 or 20 years). In contrast, *macro panels* usually involve data collected for a number of countries over time. A macro panel may have a moderately sized N (varying from, say, seven countries of the G7 to a larger set of, say, 20 OECD or European Union countries, or a mix of developed and developing countries, usually between 100 and 200 of them), but the data are usually observed annually over 20 to 60 years. Micro and macro panels require different econometric treatment. For example, the asymptotics for micro panels have to be for large N and fixed T , whereas the asymptotics for macro panels can be for large N and large T . Also, with a long time series for macro panels, one has to deal with issues of nonstationarity in the time series, such as unit roots, structural breaks and cointegration; see Chapter 12. In contrast, for micro panels one does not need to be concerned with nonstationarity issues, since T is short for each individual or household surveyed. Another example of the different treatment required is that in macro panels one has to deal with cross-country dependence, but this is not usually an issue with micro panels where the individuals or households are randomly sampled and hence not likely to be correlated. Chapter 13 studies spatial dependence in panel data as a simple way to model externalities and spillovers across cross-sectional units.

1.1.1 Examples of Micro Panels

Two well-known examples of US micro panel data are the Panel Study of Income Dynamics (PSID) collected by the Institute for Social Research at the University of Michigan (<http://psidonline.isr.umich.edu>) and the National Longitudinal Surveys (NLS), a set of surveys sponsored by the Bureau of Labor Statistics (<http://www.bls.gov/nls/home.htm>).

The PSID began in 1968 with 4800 families and grew to encompass more than 7000 families in 2001. By 2003, the PSID had collected information on more than 65 000 individuals spanning as much as 36 years of their lives. Annual interviews were conducted from 1968 to 1996. In 1997, this survey was redesigned for biennial data collection. In addition, the core sample was reduced and a refresher sample of post-1968 immigrant families and their

2 Econometric Analysis of Panel Data

adult children was introduced. The central focus of the data is economic and demographic. The list of variables includes income, poverty status, public assistance in the form of food or housing, other financial matters (e.g. taxes, interhousehold transfers), family structure and demographic measures, labour market work, housework time, housing, geographic mobility, socioeconomic background, and health. Other supplemental topics include housing and neighbourhood characteristics, achievement motivation, child care, child support and child development, job training and job acquisition, retirement plans, health, kinship, wealth, education, military combat experience, risk tolerance, immigration history, and time use.

The NLS, on the other hand, consists of surveys designed to gather information at multiple points in time on the labour market activities and other significant life events of several groups of men and women:

- (1) The NLSY97 consists of a nationally representative sample of approximately 9000 youths who were 12–17 years old as of 1997.
- (2) The NLSY79 consists of a nationally representative sample of 12 686 young men and women who were 14–22 years old in 1979. These individuals were interviewed annually through 1994 and are currently interviewed on a biennial basis.
- (3) The NLSY79 children and young adults survey includes the biological children born to women in the NLSY79.

The list of variables includes information on schooling and career transitions, marriage and fertility, training investments, child care usage, and drug and alcohol use. A large number of studies have made use of the NLS and PSID data sets. The PSID applications cover a wide range of topics, including: intertemporal models of labour supply; wages and employment over the business cycle; unemployment, job turnover and labour mobility; consumption, income and balance sheet dynamics; extended family behaviour; poverty, welfare and income dynamics; intergenerational transmission of economic status; and antecedents of economic and demographic events.

Panels can also be constructed from the Current Population Survey (CPS), a monthly national survey of about 50 000 households in the USA conducted by the Bureau of Census for the Bureau of Labor Statistics (<http://www.census.gov/cps>). This survey has been ongoing for more than 50 years. The CPS is the primary source of information on the labour force characteristics of the US population. Compared with the NLS and PSID data, the CPS contains fewer variables, spans a shorter time period and does not follow movers. However, it covers a much larger sample and is representative of all demographic groups. The CPS provides estimates of employment, unemployment, earnings, hours of work, and other indicators. These estimates are available by a variety of demographic characteristics, including age, sex, race, marital status, and educational attainment; they are also available by occupation, industry, and class of worker.

Another important source of household survey data for developing countries is the World Bank's Living Standards Measurement Study (LSMS, <http://www.worldbank.org/LSMS>), which was established in 1980. Since 1985, LSMS has conducted surveys in about 20 developing countries from Albania to Vietnam. These tend to involve small samples, in the order of 2000 to 5000 households. In some countries it could be one survey or multiple surveys. In other countries it could be a two- to four-year panel. Three types of questionnaires were used: a household, a community, and a price questionnaire. In some cases, a school or health facility questionnaire was added. The LSMS data have focused mostly on documenting regularities concerning the nature of poverty. Repeated surveys like the LSMS, even if they

might not constitute a genuine panel, can be used to construct a *pseudo panel*, as we will see in Chapter 10.

Although the US panels started in the 1960s, it was not until the 1980s that the European panels began setting up. In 1989, a special section of the *European Economic Review* published papers based on the German Socio-Economic Panel, the Swedish study of household market and nonmarket activities, and the Intomart Dutch panel of households. The first wave of the German Socio-Economic Panel (GSOEP, <http://www.diw.de/soep>) data was collected by the DIW (German Institute for Economic Research, Berlin) in 1984 and included 5921 West German households, involving 12 290 respondents. Standard demographic variables and information on wages, income, benefit payments, level of satisfaction with various aspects of life, hopes and fears, political involvement, etc. were collected. In 1990, after German reunification, 4453 adult respondents in 2179 households from East Germany were also included in the GSOEP. The attrition rate has been relatively low in GSOEP. Wagner, Burkhauser and Behringer (1993) reported that through eight waves of the GSOEP, 54.9% of the original panel respondents have records without missing years. The British Household Panel Survey (BHPS), an annual survey of private households in Britain first collected in 1991 by the Institute for Social and Economic Research at the University of Essex (<https://www.iser.essex.ac.uk/bhps>), comprises a national representative sample of some 5500 households and 10 300 individuals drawn from 250 areas of Great Britain. Additional samples of 1500 households each from Scotland and Wales were added to the main sample in 1999, and in 2001 a sample of 2000 households from Northern Ireland was added as well. The data collected include demographic and household characteristics, as well as information on household organization, labour market activity, health, education, housing, consumption, income, and social and political values. The Swedish Panel Study of Market and Non-market Activities (HUS) was conducted in 1984, 1986, 1988, 1991, 1993, 1996 and 1998 (<http://www.nek.uu.se/faculty/klevmark/hus.htm>), collecting data from 2619 individuals on child care, housing, market work, income and wealth, tax reform (1993), willingness to pay for a good environment (1996), local taxes, public services, and activities in the black economy (1998).

The European Community Household Panel (ECHP) is centrally designed and coordinated by the Statistical Office of the European Communities (EuroStat); see Peracchi (2002). The first wave was conducted in 1994 and included all members of the EU at that time except Austria, Finland and Sweden. Austria joined in 1995, Finland in 1996, and data for Sweden has been obtained from the Swedish Living Conditions Survey. The project was launched to obtain comparable information across member countries on income, work and employment, poverty and social exclusion, housing, health, and many other diverse social indicators of living conditions of private households and persons. The ECHP was linked from the beginning to existing national panels (e.g. those of Belgium and Holland) or ran parallel to existing panels with similar content, such as GSOEP, PSELL (Luxembourg's Panel Socio-Economique Liewen zu Lëtzebuerg) and the BHPS.

Other panel studies include:

- the Canadian Survey of Labor Income Dynamics (SLID), collected by Statistics Canada (www.statcan.gc.ca), which includes a sample of approximately 37 000 households located throughout all ten provinces. Years available are 1993–2000.
- the Japanese Panel Survey on Consumers (JPSC), collected in 1994 by the Institute for Research on Household Economics (www.kakeiken.or.jp). This consists of a nationally

4 Econometric Analysis of Panel Data

representative sample of 1500 women aged between 24 and 34 years in 1993 (cohort A). In 1997, 500 women were added whose ages were between 24 and 27 (cohort B). Years available are 1994–2000. Information gathered includes family composition, labour market behaviour, income, consumption, savings, assets, liabilities, housing, consumer durables, household management, time use, and satisfaction.

- the Russian Longitudinal Monitoring Survey (RLMS), collected in 1992 by the Carolina Population Center at the University of North Carolina (<http://www.cpc.unc.edu/projects/rhms-hse>). This is a nationally representative household survey designed to measure the effects of Russian reforms on economic well-being. Variables include individual health and dietary intake, measurement of expenditures and service utilization, and community-level data such as region-specific prices and information on community infrastructure.
- the Korea Labor and Income Panel Study (KLIPS), available for 1998–2001. This is a survey of 5000 households and their members from seven metropolitan cities and urban areas in eight provinces (<http://www.kli.re.kr/klips/en/about/introduce.jsp>).
- the Household, Income and Labour Dynamics in Australia (HILDA), a household panel survey whose first wave was conducted by the Melbourne Institute of Applied Economic and Social Research in 2001 (<http://www.melbourneinstitute.com/hilda/>). This survey includes 7682 households with 19 914 individuals from 488 different neighbouring regions across Australia.
- the Indonesia Family Life Survey (IFLS), available for 1993/94, 1997/98 and 2000. The sample is representative of about 83% of the Indonesian population and contains over 30 000 individuals living in 13 of the 26 provinces. In 1993, 7224 households were interviewed, and over 7700 households were interviewed in 2000 (<http://www.rand.org/labor/FLS/IFLS.html>).

The above list of panel data sets, though by no means exhaustive, provides a good selection of panel data sets readily accessible for economic research.

1.1.2 Examples of Macro Panels

Several sources of macro panel data commonly utilized by economists are the following:

- (i) The Penn World Table (PWT), available at www.nber.org, provides purchasing power parity and national income accounts converted to international prices for 188 countries for some or all of the years from 1950 to 2004. In addition, the European Union and the OECD provide detailed purchasing power and real product estimates for their countries, and the World Bank gives current price estimates for most PWT countries at the GDP level.
- (ii) The World Bank is a great source of macro panels (<http://data.worldbank.org>), such as the World Development Indicators (WDI). For example, the 2007 WDI includes more than 900 indicators for 152 economies with populations of more than 1 million.
- (iii) The International Monetary Fund (www.imf.org) provides several sources of macro panel data. These include: the World Economic Outlook Databases, which provide time-series data on GDP growth, inflation, unemployment, payments balances, exports, imports, external debt, capital flows, commodity prices, and so on; the International Financial Statistics, which provide approximately 32 000 time series covering more than 200 countries starting in 1948 and containing information about exchange rates, fund

accounts, and the main global and country economic indicators; the Direction of Trade Statistics year book, which provides seven years of trade data for about 186 countries, with quarterly data covering the most recent six quarters and the latest year for about 156 countries. Data are also available on balance of payments and international investment positions, as well as indices of primary commodity prices. In addition, the International Monetary Fund provides member countries' data on international reserves and foreign currency liquidity, as well as Financial Soundness Indicators.

- (iv) The United Nations provides a wealth of macro country panel data (http://unstats.un.org/unsd/economic_main.htm), including information on national accounts, trade, and industry statistics.
- (v) The OECD provides data on its website at www.oecd.org.
- (vi) The European Central Bank (ECB) provides data on European Union member countries at www.ecb.int.
- (vii) The Central Intelligence Agency's World Factbook, available at <https://www.cia.gov/library/publications/the-world-factbook/>, is another rich source of country data.

These are but a few of the agencies that provide macro data on individual countries over time, which can be pooled and used in panel studies. We shall study several special types of panel data encountered in practice, such as unbalanced panels in Chapter 9, nested panels in Section 9.6, unequally spaced panels in Section 5.2.5, rotating panels in Section 10.2, pseudo panels in Section 10.3, spatial panels in Chapter 13, count panels in Section 10.6, and heterogeneous panels in Section 10.5.

1.1.3 Some Basic References

Virtually every graduate text in econometrics contains a chapter or a major section on the econometrics of panel data. Recommended readings on this subject include the Econometric Society monograph *Analysis of Panel Data* by Hsiao (2003), along with two chapters in the *Handbook of Econometrics*: Chapter 22 by Chamberlain (1984) and Chapter 53 by Arellano and Honoré (2001). Maddala (1993) edited two volumes collecting some of the classic articles on the subject. This collection was updated with two more volumes covering the period 1992–2002, edited by Baltagi (2002). Other books on the subject include Arellano (2003), Wooldridge (2010), and Mátyás and Sevestre (2008). Some survey papers on non-stationary panel models are Baltagi and Kao (2000), Choi (2006), and Breitung and Pesaran (2008). Special issues of journals dedicated to panel data include a 1995 special issue of the *Journal of Econometrics* edited by Baltagi, two volumes of the *Annales d'Économie et de Statistique* edited by Sevestre (1999), a special issue of the *Oxford Bulletin of Economics and Statistics* edited by Banerjee (1999), two special issues (year 2000, volume 19, issues 3 and 4) of *Econometric Reviews* edited by Maasoumi and Heshmati, a special issue of *Advances in Econometrics* edited by Baltagi, Fomby and Hill (2000), two special issues of *Empirical Economics* edited by Baltagi (2004) and by Baltagi and Breitung (2011), a special issue of the *Journal of Applied Econometrics* edited by Baltagi and Pesaran (2007) and, more recently, a special issue of *Econometric Reviews* edited by Baltagi and Maasoumi (2013).

The objective of this book is to provide a simple introduction to some of the basic issues of panel data analysis. It is intended for economists and social scientists with the usual background in statistics and econometrics. Panel data methods have been used in political science

(e.g. Beck and Katz, 1995) and in sociology, finance and marketing (e.g. Keane, 1997). While restricting the focus of the book to basic topics may not do justice to the rapidly growing literature, it is nevertheless unavoidable in view of space limitations. Topics not covered in this book include duration models and hazard functions (see Heckman and Singer, 1985), as well as the frontier production function literature using panel data (e.g. Kumbhakar and Lovell, 2000; Koop and Steel, 2001), the literature on time-varying parameters, random coefficients and Bayesian models (e.g. Swamy and Tavlas, 2001; Hsiao, 2003), and the literature on nonparametric and semiparametric panels (e.g. Li and Racine, 2007).

1.2 WHY SHOULD WE USE PANEL DATA? THEIR BENEFITS AND LIMITATIONS

Hsiao (2003) lists several benefits of using panel data. These include the following:

(1) The use of panel data enables us to *control for individual heterogeneity*. Panel data suggest that individuals, firms, states or countries are heterogeneous. Time-series and cross-section studies that do not control for such heterogeneity run the risk of obtaining biased results; see, e.g., Moulton (1986, 1987). Let us demonstrate this with an empirical example. Baltagi and Levin (1986) considered panel data estimation of cigarette demand across 46 American states. Consumption is modelled as a function of lagged consumption, price and income; these variables vary across states and over time. There are, however, many other variables that affect consumption and which may be state-invariant or time-invariant. Let us call the state-invariant variables W_i and the time-invariant variables Z_i . Examples of Z_i are religion and education. For the religion variable, one might not be able to know the percentage of each state's population that is, say, Mormon for every year, but nor does one expect the percentage to change much over time. The same holds true for the percentage of each state's population that has completed high school or holds a college degree. Examples of W_i include advertising on TV and radio; such advertising is typically nationwide and does not vary across states. Some variables may be difficult to measure or hard to obtain, so not all possible Z_i and W_i variables will be available for inclusion in the consumption equation. Omission of such variables will lead to bias in the resulting estimates. By using panel data, one is better able to control for such state- or time-invariant variables, whereas a time-series study or cross-section study cannot. In fact, the data on cigarette demand show that Utah has less than half the average per capita consumption of cigarettes in the USA. This is because the population of Utah is mostly Mormon, and Mormonism prohibits smoking. Controlling for Utah in a cross-section regression can be done with a dummy variable which has the effect of removing that state's observation from the regression. This would not be the case for panel data, as we will shortly discover. In fact, with panel data, one might first difference the data to get rid of all Z_i -type variables and hence effectively control for all state-specific characteristics. This holds whether the Z_i variables are observable or not. Alternatively, the dummy variable for Utah controls for every state-specific effect that is distinctive of Utah without omitting the observations for Utah.

Another example is given by Hajivassiliou (1987), who studied the external debt repayments problem using a panel of 79 developing countries observed over the period 1970–1982. The countries in the study differ in terms of their colonial history, financial institutions, religious affiliations and political regimes. All of these country-specific variables affect the attitudes of the countries with regard to borrowing and defaulting and the way they

are treated by the lenders. Not accounting for this country heterogeneity would cause serious misspecification.

Deaton (1995) gives another example, arising from agricultural economics. This pertains to the question of whether small farms are more productive than large ones. Ordinary least squares (OLS) regressions of yield per hectare on inputs such as land, labour, fertilizer use, farmer's education, etc. usually find that the sign of the estimate of the land coefficient is negative, implying that smaller farms are more productive. Some explanations from economic theory have argued that higher output per head is an optimal response to uncertainty by small farmers, or that hired labour requires more monitoring than family labour. Deaton (1995) offers an alternative explanation: the regression analysis suffers from the omission of unobserved heterogeneity, in this case "land quality", and the omitted variable is systematically correlated with the explanatory variable (farm size). In fact, farms in low-quality marginal areas (semi-desert) are typically large, while farms in high-quality land areas are often small. Deaton argued that while gardens generate more value-added per hectare than a sheep station, this does not imply that sheep stations should be organized as gardens. In this case, differencing may not resolve the "small farms are productive" question, since farm sizes will usually show little or no change over short periods.

(2) Panels give *more informative data, more variability, less collinearity among the variables, more degrees of freedom, and more efficiency*. Time-series studies are plagued with multicollinearity; for instance, in the above example about cigarette demand, there is high collinearity between price and income in the aggregate time series for the USA; this is less likely with a panel across American states, since the cross-section dimension adds a lot of variability, yielding more informative data on price and income. In fact, the variation in the data can be decomposed into variation between states of different sizes and characteristics, and variation within states. The former variation is usually larger. With additional, more informative data, one can obtain more reliable parameter estimates. Of course, the same relationship has to hold for each state, i.e. the data have to be poolable. This is a testable assumption and one that we shall tackle in due course.

(3) With panel data, one is better able to study the *dynamics of adjustment*. Cross-sectional distributions that look relatively stable can hide a multitude of changes. Spells of unemployment, job turnover, or residential and income mobility are better studied with panels. Panel data are also well suited for studying the duration of economic states such as unemployment and poverty, and if the panels are long enough, they can shed light on the speed of adjustments to economic policy changes. For example, in measuring unemployment, cross-sectional data can be used to estimate what proportion of the population is unemployed at a given point in time; repeated cross-sections can then show how this proportion changes over time. However, only panel data can provide estimates of the proportion unemployed in one period who remain unemployed in another period. Important policy questions, such as determining whether families' experiences of poverty, unemployment and welfare dependence are transitory or chronic, necessitate the use of panels. Deaton (1995) argued that, unlike cross-sections, panel surveys yield data on *changes* for individuals or households. Panel data allow us to observe *how* individual living standards change during the development process, and enable us to determine *who* is benefiting from development. Panel surveys also allow us to observe whether poverty and deprivation are transitory or long-lived, i.e. the income-dynamics question. Furthermore, panels are necessary for the estimation of intertemporal relations and the construction of life-cycle and intergenerational models. In fact, panels can relate the individual's experiences and behaviour at one point in time to other experiences and behaviour at another point in time. For

example, in evaluating training programs, a group of participants and non-participants are observed before and after implementation of the training program. Such a panel involving at least two time periods forms the basis for the “difference in differences” estimator; see Chapter 2.

(4) Panel data are more suitable for *identifying and measuring effects that are simply not detectable in pure cross-section or pure time-series data*. Suppose that we have a cross-section of women with a 50% average yearly labour force participation rate. This might be due to (a) each woman having a 50% chance of being in the labour force in any given year, or (b) 50% of the women working all the time and 50% not at all. Case (a) has high turnover, while case (b) has no turnover. Only panel data could discriminate between the two cases. Another example is the determination of whether union membership increases or decreases wages. To answer this question, it is better to observe individual workers moving from union to non-union jobs or vice versa. Holding the individual’s characteristics constant, we would be better equipped to determine whether union membership affects wages and, if so, by how much. This kind of analysis extends to the estimation of other types of wage differentials, holding individuals’ characteristics constant – for example, the estimation of wage premiums paid in dangerous or unpleasant jobs.

Economists studying workers’ level of satisfaction run into the problem of anchoring in a cross-section study; see Chapter 11 of Winkelmann and Winkelmann (1998). Such a survey usually asks the question “How satisfied are you with your life?”, with responses scored on a scale from 0, meaning completely dissatisfied, to 10, meaning completely satisfied. The problem is that each individual anchors their scale at a different level, rendering interpersonal comparisons of responses meaningless. However, in a panel study, where the metric used by each individual is time-invariant over the period of observation, one can avoid this problem by using a difference (or fixed effects) estimator, which will make inference based only on intra- rather than interpersonal comparisons of satisfaction.

(5) Panel data models allow us to *construct and test more complicated behavioural models than do pure cross-section or time-series data*. For example, technical efficiency is better studied and modelled using panels (see Kumbhakar and Lovell, 2000; Koop and Steel, 2001).

(6) Micro panel data gathered on individuals, firms and households can be measured more accurately than similar variables measured at the macro level. Biases resulting from aggregation over firms or individuals may be reduced or eliminated.

(7) Macro panel data, on the other hand, have longer time series and, as we shall see in Chapter 12, panel unit root tests have standard asymptotic distributions and do not suffer from the problem of nonstandard distributions encountered with unit root tests in time-series analysis.

Limitations of panel data include:

(1) *Design and data collection problems*. For an extensive discussion of problems that arise in designing panel surveys, as well as data collection and data management issues, see Kasprzyk *et al.* (1989). These include problems of coverage (incomplete account of the population of interest), nonresponse (due to lack of cooperation of the respondent or interviewer error), recall (respondent not remembering correctly), frequency of interviewing, interview spacing, reference period, the use of bounding, and time-in-sample bias.¹

(2) *Distortions of measurement errors*. Measurement errors may arise because of faulty responses due to unclear questions, memory errors, deliberate distortion of responses (e.g. prestige bias), inappropriate informants, misrecording of responses, and interviewer effects (see Kalton, Kasprzyk and McMillen, 1989). The validation study by Duncan and Hill (1985) on the PSID illustrates the significance of the measurement error problem. They compared

the responses of employees of a large firm with the records of the employer, and found small response biases except in the case of work hours, which are overestimated. The ratio of measurement error variance to the true variance was found to be 15% for annual earnings, 37% for annual work hours, and 184% for average hourly earnings. These figures are for a one-year recall, i.e. 1983 for 1982, and become more than doubled with two years' recall. Brown and Light (1992) investigated the inconsistency in job tenure responses in the PSID and NLS. Cross-section data users have little choice but to believe the reported values of tenure (unless they have external information), while users of panel data can check for inconsistencies in tenure responses with elapsed time between interviews; for example, a respondent may claim to have three years of tenure in one interview and a year later claim six years. This should alert the user of the panel to the presence of measurement error. Brown and Light (1992) showed that failure to use internally consistent tenure sequences can lead to misleading conclusions about the slope of wage–tenure profiles. Section 10.1 deals with measurement error in panel data.

(3) *Selectivity problems.* These include:

- (a) *Self-selectivity.* If people choose not to work because the reservation wage is higher than the offered wage, in this case we would observe the characteristics of the individuals but not their wage. Since only their wage is missing, the sample is censored. However, if we do not observe all data on these people, this would be a truncated sample. An example of truncation is the New Jersey negative income tax experiment: we are only interested in poverty, and people with income higher than 1.5 times the poverty level are dropped from the sample. Inference from this truncated sample introduces bias that is not helped by more data, because of the truncation (Hausman and Wise, 1979). Chapter 11 deals with selectivity problems in panel data.
- (b) *Nonresponse.* This can occur at the initial wave of the panel due to refusal to participate, nobody at home, untraced sample unit, and other reasons. Item (or partial) nonresponse occurs when one or more questions are left unanswered or are found not to provide a useful response. Complete nonresponse occurs when no information is available from the sampled household. Besides the efficiency loss due to missing data, such nonresponse can cause serious identification problems for the population parameters. The seriousness of the problem is directly proportional to the amount of nonresponse. Nonresponse rates in the first wave of the European panels varied from 10% in Greece and Italy, where participation was compulsory, to 52% in Germany and 60% in Luxembourg. The overall nonresponse rate was 28%; see Peracchi (2002). The comparable nonresponse rate for the first wave of the PSID was 24%, for the BHPS 26%, for the GSOEP 38%, and for PSELL 35%.
- (c) *Attrition.* While nonresponse occurs also in cross-section studies, it is more of a serious problem in panels, because subsequent waves of the panel are still subject to nonresponse. Respondents may die, move, or find that the cost of responding is too high; see Chapter 11 for a discussion of the consequences of attrition in panels. The degree of attrition varies depending on the panel studied; see Kalton, Kasprzyk and McMillen (1989) for several examples. In general, the overall rates of attrition increase from one wave to the next, but the rate of increase declines over time. Beckett *et al.* (1988) studied the representativeness of the PSID 14 years after its start. They found that only 40% of those originally in the sample in 1968 remained in the sample in 1981. Nevertheless, they did find that, as far as the dynamics of entry and exit are concerned, the PSID is still representative. The potentially most damaging threat to the value of panel data is

the presence of biasing attrition. Fitzgerald, Gottschalk and Moffit (1998) reported 51% attrition of the original PSID sample by 1989, with the major reasons being family unit nonresponse, death, or a residential move. Attritors were found to have lower earnings, lower education levels, and lower marriage propensities. But despite the large amount of attrition, Fitzgerald, Gottschalk and Moffit (1998) found no strong evidence that it had seriously distorted the representativeness of the PSID through 1989. In the same vein of research, Lillard and Panis (1998) found evidence of significant selectivity in attrition for the PSID; for example, they found that less educated individuals and older people are more likely to drop out, whereas married people are more likely to continue. Propensity to participate in the survey diminishes with increasing duration of the respondent in the sample. Despite this, the effects of ignoring such selective attrition on household income dynamics, marriage formation and dissolution, and adult mortality risk are mild. In Europe, the comparable attrition rates (between the first and second waves) vary from 6% in Italy to 24% in the UK; the average attrition rate was about 10%. For the BHPS, attrition from the first to the second wave was 12%; for PSELL it was 15%. For the GSOEP, attrition was 12.4% for the West German sample and 8.9% for the East German sample; see Peracchi (2002). In order to counter the effects of attrition, rotating panels are sometimes used, where a fixed percentage of respondents are replaced in every wave to replenish the sample. More details on rotating and pseudo panels can be found in Chapter 10. A special issue of the *Journal of Human Resources*, Spring 1998, is dedicated to attrition in longitudinal surveys.

(4) *Short time-series dimension*. Typical micro panels involve annual data covering a short time span for each individual. This means that asymptotic arguments rely crucially on the number of individuals tending to infinity. Increasing the time span of the panel is not without cost either. In fact, this increases the chances of attrition and increases the computational difficulty for limited dependent variable panel data models (see Chapter 11).

(5) *Cross-section dependence*. Macro panels on countries or regions with long time series that do not account for cross-country dependence may lead to misleading inference. In Chapter 12 it is shown that several panel unit root tests suggested in the literature assume cross-section independence. Accounting for cross-section dependence turns out to be important and affects inference. Alternative panel unit root tests have been proposed that account for such dependence. Chapter 13 surveys tests for cross-sectional dependence in panels.

Panel data is not a panacea and will not solve all the problems that a time-series or cross-section study could not handle. Examples are given in Chapter 12 where we cite econometric studies arguing that panel data will yield more powerful unit root tests than individual time-series. This, in turn, should help shed more light on the purchasing power parity (PPP) and growth convergence questions. In fact, this led to a flurry of empirical applications, along with objections from some sceptics who argued that panel data did not really solve the PPP or growth convergence problem; see Maddala (1999), Maddala, Wu and Liu (2000), and Banerjee, Marcellino and Osbat (2004, 2005). Collecting panel data is quite costly, and there is always the question of how often one should interview respondents. Deaton (1995) argues that economic development is far from instantaneous, and so changes from one year to the next are probably too noisy and too short-term to be really useful. He concludes that the pay-off for panel data is over long time periods, such as five years, ten years, or even longer. In contrast, for health and nutrition issues, especially those of children, one could argue the opposite case, i.e. panels with a shorter time span are necessary for monitoring the health and development of children.

This book will make the case that panel data offer several advantages worth their cost. However, as Zvi Griliches argued about economic data in general, the more we have of it, the more we demand of it. The economist using panel data, or any data for that matter, has to know their limitations.

NOTE

1. Bounding is used to prevent the shifting of events from outside the recall period into the recall period. Time-in-sample bias is observed when a significantly different level for a characteristic occurs in the first interview than in later interviews, when one would expect the same level.

