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 latter are known as *micro panels* and are collected for a large number of individuals *N* (usually in the hundreds or thousands) over a short time period *T* (varying from a minimum of 2 years to a maximum rarely exceeding 10 or 20). In contrast, *macro panels* usually involve a number of countries over time. These may have a moderate size *N* (varying from 7 countries say for the G7 countries 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). These are usually observed annually over 20 to 60 years. Micro and macro panels require different econometric care. For example, the asymptotics for micro panels have to be for large *N* and fixed *T*, whereas the asymptotics for macropanels can be for large *N* and *T*. Also, with a long time series for macro panels one has to deal with issues of nonstationarity in the time series, like unit roots, structural breaks and cointegration, see Chapter 12, whereas for micro panels one does not deal with nonstationarity issues, especially since *T* is short for each individual or household. Also, in macro panels one has to deal with cross-country dependence. These are not usually an issue in micro panels where the households are randomly sampled and hence not likely to be correlated.

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) which is 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 has grown to 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 adult children was introduced. The central focus of the data is economic and demographic. The list of variables include 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, labor market work, housework time, housing, geographic mobility, socioeconomic background and health. Other supplemental topics include housing and neighborhood 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, are a set of surveys designed to gather information at multiple points in time on labor 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 currently interviewed on a biennial basis.
(3) The NLSY79 children and young adults. This includes the biological children born to women in the NLSY79.
(4) The NLS of mature women and young women: these include a group of 5083 women who were between the ages of 30 and 44 in 1967. Also, 5159 women who were between the ages of 14 and 24 in 1968. Respondents in these cohorts continue to be interviewed on a biennial basis.
(5) The NLS of older men and young men: these include a group of 5020 men who were between the ages of 45 and 59 in 1966. Also, a group of 5225 men who were between the ages of 14 and 24 in 1966. Interviews for these two cohorts ceased in 1981.

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 used the NLS and PSID data sets. The NLS web site provides a search engine for over 3000 articles, monographs and working papers using the NLS data, while Brown, Duncan and Stafford (1996) estimate that roughly 900 journal and book articles using the PSID have been published. The PSID applications cover a wide range of topics including intertemporal models of labor supply; wages and employment over the business cycle; unemployment, job turnover and labor mobility; consumption, income and balance sheet dynamics; extended family behavior; 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 household survey of about 50 000 households conducted by the Bureau of Census for the Bureau of Labor Statistics (www.census.gov/cps). This survey has been conducted for more than 50 years. The CPS is the primary source of information on the labor force characteristics of the US population. Compared with the NLS and PSID data, the CPS contains fewer variables, spans a shorter 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 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) which was established in 1980 (www.worldbank.org/LSMS). Since 1985, LSMS has conducted surveys in about 20 developing countries from Albania to Vietnam; see Grosh and Glewwe (1998). These tend to be small samples of the order of 2000 to 5000 households. In some countries this could be one survey or multiple surveys. In other countries it could be a two to a four year panel. Three types of questionnaires were conducted: 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 though they may 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 only in the 1980s that the European panels began setting up. In 1989, a special section of the *European Economic Review* published papers using the German Social 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) was collected by the DIW (German Institute for Economic Research, Berlin) in 1984 and included 5921 West German households (www.diw.de/soep). This included 12 290 respondents. Standard demographic variables as well as wages, income, benefit payments, level of satisfaction with various aspects of life, hopes and fears, political involvement, etc. are collected. In 1990, 4453 adult respondents in 2179 households from East Germany were included in the GSOEP due to German unification. The attrition rate has been relatively low in GSOEP. Wagner, Burkhauser and Behringer (1993) report that through eight waves of the GSOEP, 54.9% of the original panel respondents have records without missing years. An inventory of national studies using panel data is given at http://psidonline.isr.umich.edu/Guide/PanelStudies.aspx. These include the Belgian Socioeconomic Panel (www.ufsia.ac.be/CSB/sep_nl.htm) which interviewed a representative sample of 6471 Belgian households in 1985, 3800 in 1988, 3800 in 1992 (including a new sample of 900 households) and 4632 households in 1997 (including a new sample of 2375 households). The British Household Panel Survey (BHPS) which is 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 (www.iser.essex.ac.uk/ulsc/bhps/). This is a national representative sample of some 5500 households and 10 300 individuals drawn from 250 areas of Great Britain. Additional samples of 1500 households in each of Scotland and Wales were added to the main sample in 1999, and in 2001 a sample of 2000 households was added in Northern Ireland. Data collected include demographic and household characteristics, household organization, labor market, health, education, housing, consumption, and income, social and political values. The Swiss Household Panel (SHP) whose first wave in 1999 interviewed 5074 households comprising 7799 individuals (www.swisspanel.ch). The Luxembourg Panel Socio-Economique “Liewen zu Lëtzebuerg” (PSELL I) (1985–94) is based on a representative sample of 2012 households and 6110 individuals. In 1994, the PSELL II expanded to 2978 households and 8232 individuals. The Swedish Panel Study Market and Nonmarket Activities (HUS) were collected in 1984, 1986, 1988, 1991, 1993, 1996 and 1998 (http://www.nek.uu.se/faculty/klevmark/hus.htm). Data for 2619 individuals were collected 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 were conducted in 1994 and included all current members of the EU except Austria, Finland and Sweden. Austria joined in 1995, Finland in 1996 and data for Sweden were 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 indicating living conditions of private households and persons. The EHCP was linked from the beginning to existing national panels (e.g. Belgium and Holland) or ran parallel to existing panels with similar content, namely GSOEP, PSELL and the BHPS. This survey ran from 1994 to 2001 (http://epunet.essex.ac.uk/echp.php).

Other panel studies include: the Canadian Survey of Labor Income Dynamics (SLID) collected by Statistics Canada (www.statcan.ca) which includes a sample of approximately 37 000 households located throughout all 10 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 is a national representative sample of 1500 women aged 24 and 34 years in 1993 (cohort A). In 1997, 500 women were added with ages between 24 and 27 (cohort B). Years available are 1994–2000. Information gathered includes family composition, labor market behavior, 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 (www.cpc.unc.edu/projects/rlms/home.html). The RLMS is a nationally representative household survey designed to measure the effects of Russian reforms on economic well-being. Data include individual health and dietary intake, measurement of expenditures and service utilization and community level data including region specific prices and community infrastructure. The Korea Labor and Income Panel Study (KLIPS) available for 1998–2001 surveys 5000 households and their members from seven metropolitan cities and urban areas in eight provinces (http://www.kli.re.kr/klips). The Household, Income and Labour Dynamics in Australia (HILDA) is 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 includes 7682 households with 19 914 individuals from 488 different neighboring regions across Australia. The Indonesia Family Life Survey (http://www.rand.org/FLS/IFLS) is 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 the country. In 1993, 7224 households were interviewed, and over 7700 households in 2000. This list of panel data sets is by no means exhaustive but provides a good selection of panel data sets readily accessible for economic research.

1.1.2 Examples of Macro Panels

In contrast to Micro Panel surveys, there are several *macro panels* utilized by economists. These include: (i) the Penn World Table (PWT) available at (http://pwt.econ.upenn.edu). The PWT provides purchasing power parity and national income accounts converted to international prices for 188 countries for some or all of the years 1950–2004. In addition, the European Union or the OECD provides detailed purchasing power and real product estimates for their countries and the World Bank makes current price estimates for most PWT countries at the GDP level. (ii) The World Bank is a great source of macro panels including the World Development Indicators (WDI) available at (www.worldbank.org/data). The 2007 WDI includes more than 900 indicators for 152 economies with populations of more than 1 million, as well as for Taiwan and China. (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 for GDP growth, inflation, unemployment, payments balances, exports, imports, external debt, capital flows, commodity prices, etc. Also, the International Financial Statistics which provide approximately 32 000 time series covering more than 200 countries starting in 1948. These include exchange rates, fund accounts and the main global and country economic indicators. The *Direction of Trade Statistics Yearbook* provides seven years of trade data for about 186 countries, and the quarterly data cover the most recent six quarters and the latest year for about 156 countries. There are also data on balance of payments and international investment position, as well as Indices of Primary Commodity Prices. The International Monetary Fund also 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 at http://unstats.un.org/unsd/economic_main.htm. These include national accounts and trade as well as industry statistics. (v) The Organization for Economic Co-operation and Development (OECD) data available at www.oecd.org. (vi) The European Central Bank (ECB) provides data on the European Union member countries at www.ecb.int. (vii) The Central Intelligence Agency's World Factbook available on the web at www.cia.gov/library/publications/the-world-factbook. These are but a few of the agencies providing macro data on individual countries over time, which can be pooled and used in panel studies.

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 Hsiao's (2003) Econometric Society monograph 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 of readings was updated with two more volumes covering the period 1992–2002 and edited by Baltagi (2002). Other books on the subject include Arellano (2003), Wooldridge (2002) and a handbook on the econometrics of panel data which in its second edition contained 33 chapters edited by Mátyás and Sevestre (1996). A book in honor of G.S. Maddala, edited by Hsiao *et al.* (1999); a book in honor of Pietro Balestra, edited by Krishnakumar and Ronchetti (2000); a book with a nice historical perspective on panel data by Nerlove (2002), and an edited book in the contribution to economic analysis series by Baltagi (2006d) with several empirical applications and theoretical contributions not all covered in this book. Survey papers on nonstationary panel models include Baltagi and Kao (2000), Choi (2006) and Breitung and Pesaran (2008). Special issues of journals dedicated to panel data include 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 (Volume 19, Numbers 3 and 4) of *Econometric Reviews* edited by Maasoumi and Heshmati (2000), a special issue of *Advances in Econometrics* edited by Baltagi, Fomby and Hill (2000), a special issue of *Empirical Economics* edited by Baltagi (2004) and a special issue of the *Journal of Applied Econometrics* edited by Baltagi and Pesaran (2007).

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, see Beck and Katz (1995); in sociology, see England *et al.* (1988); in finance, see Boehmer and Megginson (1990); and in marketing, see Erdem (1996) and Keane (1997). While restricting the focus of the book to basic topics may not do justice to this rapidly growing literature, it is nevertheless unavoidable in view of the space limitations of the book. Topics not covered in this book include duration models and hazard functions (see Heckman and Singer, 1985, and Horowitz and Lee, 2004). Also, the frontier production function literature using panel data (see Schmidt and Sickles, 1984; Battese and Coelli, 1988; Kumbhakar and Lovell, 2000; Koop and Steel, 2001), and the literature on time-varying parameters, random coefficients and Bayesian models (see Swamy and Tavlas 2001 and Hsiao, 2003). The program evaluation literature (see Heckman, Ichimura and Todd, 1998 and Abbring and van den Berg, 2004), to mention a few.

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

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

(1) Controlling for *individual heterogeneity*. Panel data suggest that individuals, firms, states or countries are heterogeneous. Time-series and cross-section studies not controlling this heterogeneity run the risk of obtaining biased results, e.g. see Moulton (1986, 1987). Let us demonstrate this with an empirical example. Baltagi and Levin (1992) consider cigarette demand across 46 American states for the years 1963–88. Consumption is modeled as a function of lagged consumption, price and income. These variables vary with states and time. However, there are a lot of other variables that may be state-invariant or time-invariant that may affect consumption. Let us call these Z_i and W_t, respectively. Examples of Z_i are religion and education. For the religion variable, one may not be able to get the percentage of the population that is, say, Mormon in each state for every year, nor does one expect that to change much across time. The same holds true for the percentage of the population completing high school or a college degree. Examples of W_t include advertising on TV and radio. This advertising is nationwide and does not vary across states. In addition, some of these variables are difficult to measure or hard to obtain so that not all the Z_i or W_t variables are available for inclusion in the consumption equation. Omission of these variables leads to bias in the resulting estimates. Panel data are able to control for these state- and time-invariant variables whereas a time-series study or a cross-section study cannot. In fact, from the data one observes that Utah has less than half the average per capita consumption of cigarettes in the USA. This is because it is mostly a Mormon state, a religion that prohibits smoking. Controlling for Utah in a cross-section regression may 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 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 studies the external debt repayments problem using a panel of 79 developing countries observed over the period 1970–82. These countries differ in terms of their colonial history, financial institutions, religious affiliations and political regimes. All of these country-specific variables affect the attitudes that these countries have with regards to borrowing and defaulting and the way they are treated by the lenders. Not accounting for this country heterogeneity causes serious misspecification.

Deaton (1995) gives another example from agricultural economics. This pertains to the question of whether small farms are more productive than large farms. OLS regressions of yield per hectare on inputs such as land, labor, fertilizer, farmer's education, etc., usually find that the sign of the estimate of the land coefficient is negative. These results imply that smaller farms are more productive. Some explanation from economic theory argues that higher output per head is an optimal response to uncertainty by small farmers, or that hired labor requires more monitoring than family labor. Deaton (1995) offers an alternative explanation. This regression suffers from the omission of unobserved heterogeneity, in this case "land quality", and this 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 argues that while gardens add 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 size will usually change little or not at all over short periods.

(2) Panel data 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 example, in the case of demand for cigarettes above, 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, adding 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 bigger. With additional, more informative data one can produce 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 will tackle in due course.

(3) Panel data are better able to study the *dynamics of adjustment*. Cross-sectional distributions that look relatively stable hide a multitude of changes. Spells of unemployment, job turnover, residential and income mobility are better studied with panels. Panel data are also well suited to study the duration of economic states like unemployment and poverty, and if these 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 estimate what proportion of the population is unemployed at a point in time. Repeated cross-sections can show how this proportion changes over time. Only panel data can estimate what proportion of those who are unemployed in one period can remain unemployed in another period. Important policy questions like determining whether families' experiences of poverty, unemployment and welfare dependence are transitory or chronic necessitate the use of panels. Deaton (1995) argues that, unlike cross-sections, panel surveys yield data on *changes* for individuals or households. It allows us to observe *how* the individual living standards change during the development process. It enables us to determine *who* is benefiting from development. It also allows us to observe whether poverty and deprivation are transitory or long-lived, the income-dynamics question. Panels are also necessary for the estimation of intertemporal relations, life-cycle and intergenerational models. In fact, panels can relate the individual's experiences and behavior at one point in time to other experiences and behavior at another point in time. For example, in evaluating training programs, a group of participants and non-participants are observed before and after the implementation of the training program. This is a panel of at least two time periods and the basis for the "difference in differences" estimator; see Chapter 2 and Bertrand, Duflo and Mullainathan (2004).

(4) Panel data are better able to *identify and measure 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 labor force participation rate. This might be due to (a) each woman having a 50% chance of being in the labor 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 these cases. Another example is the determination of whether union membership increases or decreases wages. This can be better answered as we observe a worker moving from union to nonunion jobs or vice versa. Holding the individual's characteristics constant, we will be better equipped to determine whether union membership affects wage and by how much. This 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 Winkelmann and Winkelmann (1998) in Chapter 11. The survey usually asks the question: "How satisfied are you with your life?" with zero meaning completely dissatisfied and 10 meaning completely satisfied. The problem is that each individual anchors their scale at different levels, rendering interpersonal comparisons of responses meaningless. However, in a panel study, where the metric used by individuals is time-invariant over the period of observation, one can avoid this problem since a difference (or fixed effects) estimator will make inference based only on intra- rather than interpersonal comparison of satisfaction.

(5) Panel data models allow us to *construct and test more complicated behavioral models than purely cross-section or time-series data*. For example, technical efficiency is better studied and modeled with panels (see Schmidt and Sickles, 1984; Baltagi and Griffin, 1988b; Kumbhakar and Lovell, 2000; Baltagi, Griffin and Rich, 1995; Koop and Steel, 2001). Also, fewer restrictions can be imposed in panels on a distributed lag model than in a purely time-series study (see Hsiao, 2003).

(6) Micro panel data gathered on individuals, firms and households may be more accurately measured than similar variables measured at the macro level. *Biases resulting from aggregation over firms or individuals may be reduced or eliminated*. For specific advantages and disadvantages of estimating life-cycle models using micro panel data, see Blundell and Meghir (1990).

(7) Macro panel data on the other hand have a longer time series and unlike the problem of nonstandard distributions typical of unit roots tests in time series analysis, Chapter 12 shows that panel unit root tests have standard asymptotic distributions.

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 because of interviewer error), recall (respondent not remembering correctly), frequency of interviewing, interview spacing, reference period, the use of bounding and time-in-sample bias (see Bailar, 1989).[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). Herriot and Spiers (1975), for example, match CPS and Internal Revenue Service data on earnings of the same individuals and show that

there are discrepancies of at least 15% between the two sources of earnings for almost 30% of the matched sample. The validation study by Duncan and Hill (1985) on the PSID also illustrates the significance of the measurement error problem. They compare the responses of the employees of a large firm with the records of the employer. Duncan and Hill (1985) find small response biases except for work hours which are overestimated. The ratio of measurement error variance to the true variance is 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 are more than doubled with two years' recall. Brown and Light (1992) investigate 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 of 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 this panel to the presence of measurement error. Brown and Light (1992) show that failure to use internally consistent tenure sequences can lead to misleading conclusions about the slope of wage-tenure profiles.

(3) Selectivity problems. These include:

(a) *Self-selectivity.* People choose not to work because the reservation wage is higher than the offered wage. In this case we observe the characteristics of these 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 larger 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 (see Hausman and Wise, 1979).

(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, this nonresponse can cause serious identification problems for the population parameters. Horowitz and Manski (1998) show that the seriousness of the problem is directly proportional to the amount of nonresponse. Nonresponse rates in the first wave of the European panels vary across countries from 10% in Greece and Italy where participation is compulsory, to 52% in Germany and 60% in Luxembourg. The overall nonresponse rate is 28%, see Peracchi (2002). The comparable nonresponse rate for the first wave of the PSID is 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 a more serious problem in panels because subsequent waves of the panel are still subject to nonresponse. Respondents may die, or move, or find that the cost of responding is high. See Björklund (1989) and Ridder (1990, 1992) on the consequences of attrition. 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. Becketti *et al.* (1988) study the representativeness of the PSID 14 years after it had started. The authors find that only 40% of those originally in the sample in 1968 remained in the sample in 1981. However, they do find that as far as the dynamics of entry and exit are concerned, the

PSID is still representative. The most potentially damaging threat to the value of panel data is the presence of biasing attrition. Fitzgerald, *et al.* (1998) report that by 1989, 51% of the original sample had attrited. The major reasons were: family unit nonresponse, death, or because of a residential move. Attritors were found to have lower earnings, lower education levels, and lower marriage propensities. Despite the large amount of attrition, Fitzgerald *et al.* (1998) report that there is no strong evidence that this attrition had seriously distorted the representativeness of the PSID through 1989. In the same vein of research, Lillard and Panis (1998) find evidence of significant selectivity in attrition for the PSID. For example, they find that less educated individuals and older people are more likely to drop out. Married people are more likely to continue. This propensity to participate in the survey diminishes the longer the duration of the respondent in the sample. Despite this, the effects of ignoring this 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 wave) vary from 6% in Italy to 24% in the UK. The average attrition rate is about 10%. For the BHPS, attrition from the first to the second wave is 12%. For PSELL it is 15%. For the GSOEP, attrition is 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 the respondents are replaced in every wave to replenish the sample. More on rotating and pseudo-panels 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. Chapter 12 shows that several panel unit root tests suggested in the literature assumed cross-section independence. Accounting for cross-section dependence turns out to be important and affects inference. Alternative panel unit root tests are suggested that account for this dependence.

Panel data is not a panacea and will not solve all the problems that a time series or a 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 the growth convergence questions. In fact, this led to a flurry of empirical applications along with some sceptics who argued that panel data did not save the PPP or the 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, so that changes from one year to the next are probably too noisy and too short term to be really useful. He concludes that the payoff for panel data is over long time periods, five years, 10 years, or even longer. In contrast, for health and nutrition issues, especially those of children, one could argue the opposite case, i.e. those panels with a shorter time span are needed in order to monitor the health and development of these children.

This book will make the case that panel data provide several advantages worth their cost. However, as Griliches (1986) 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 its 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.