

A research design that collects information of the same units repeatedly over time is called a *panel*. Traditionally, panel studies use surveys and focus on individuals. But increasingly, this design is also applied to the analysis of firms, nations, and other social entities using all kinds of source (official statistics, process-produced data, etc.).

The collection of panel data in academic research dates back to the 1940s when Paul F. Lazarsfeld (Lazarsfeld and Fiske, 1938; Lazarsfeld, 1940) started to introduce this methodology from market research into the analysis of public opinion. The first classical panel study (also known as the Erie County study) was an analysis of voting behavior during the 1940 presidential campaign and was conducted by the Bureau of Applied Social Research of Columbia University under the direction of Lazarsfeld himself (Lazarsfeld et al., 1944). Ten years later, the ELMIRA study was published that analyzed some of the open questions of the Erie County study using panel data collected during the 1948 presidential campaign (Berelson et al., 1954).

In the present day, numerous panels are available. They can be found in all social and life sciences. Chapter 6 lists some of the most prominent social science examples (see Table 6.1). The classical examples are the US American National Longitudinal Surveys of Labor Market Experience (NLS) and the University of Michigan's Panel Study of Income Dynamics (PSID) that were started in the 1960s. In many respects, both studies have been prototypes for many other household panels. In Europe, various countries have their own national household panel studies, among them the German Socioeconomic Panel Study (SOEP), the British Household Panel Survey (BHPS), and the Swiss Household Panel (SHP). In response to the increasing demand in the European Union for comparable information across Member States, Eurostat has coordinated in the 1990s a European Community Household Panel (ECHP), which later has been replaced by the European Union Statistics of Income and Living Conditions (EU-SILC). Many countries outside the US and Europe have initiated similar panel studies (e.g., Korea Labor Income Panel Study; Household, Income and Labor Dynamics in Australia Survey). A research project at the Department of Policy Analysis and Management at Cornell University has integrated some of these data in a large comparative panel data set, the Cross National Equivalent

File (CNEF), which includes data from Australia, Canada, Germany, Great Britain, Korea, Switzerland, and the US.

All of the previous panel studies focus on individuals (in households), but Table 6.1 mentions also some other examples. For instance, the Organization for Economic Development (OECD) provides a Social Expenditure Database (SOCX) that includes yearly social policy indicators for 34 OECD countries since 1980. In this case the unit of analysis is the country. Another example is the IAB Establishment Panel (IAB-EP) of the Institute for Employment Research (IAB) of the German Federal Employment Agency. It is a yearly repeated survey of German establishments, which began 1993 in West Germany and 1996 in East Germany. Here the unit of analysis is the single establishment.

As in the aforementioned household panels, the IAB-EP is a survey, while OECD's Social Expenditure Database uses official government statistics. However, the establishments in IAB-EP can be matched with data on employees generated in labor administration and social security data processing. Obviously, the method of data collection varies between different panel studies. Therefore, by using the term "panel" we refer to a specific research design (repeated measurements of identical units) and not to a particular method of data collection.

1.1 Benefits and Challenges of the Panel Design

As the increasing number of panel studies in the recent years shows, the panel design has become increasingly attractive in social research. It can answer more research questions in a much more convincing manner than other research designs. However, a panel is a complex research design and presents many new challenges for social science methodology. We start by summarizing some of its benefits, before we briefly mention the most important challenges.

1.1.1 Benefits

1.1.1.1 Measuring Change at the Individual Level

The main motivation for collecting panel data is an interest in the analysis of change; more specifically, an interest in the analysis of change at the (individual) level of units. What is meant by this can be illustrated with a classical example from poverty research.

How to measure poverty and whether it is a social problem public policy should take care of, is a constant controversy in public discourse. The conventional poverty indicator measures the number of individuals having less economic resources than 40, 50 or 60 % of the median income in their home country. For instance, the European Union defines individuals falling below 60 % of the median income at risk of poverty (Atkinson et al., 2002). Of course, the details of this indicator are much more involved (Which incomes to look at? How to compare single persons and individuals living in families?) but for our present purpose it is enough to say that such a measure exists.

According to the data in Duncan et al. (1993, 231) the average at-risk-of-poverty rate in the US between 1980 and 1985 amounted to 27.9 %. In the following six-year-period from 1981 to 1986, the rate was slightly higher (on average 28.6 %). If more than one-quarter of US citizens according to this definition are poor, it looks like the US had a dramatic poverty problem in the early 1980s. Yet, some scholars argue that the 60 % threshold is much too generous; it measures individuals *at the risk* of poverty, they argue, but not those *in* poverty. For that purpose, the 40 % threshold should be used and according to that measure fewer US citizens were estimated as being poor in the early 1980s (on average 13.6 %) (see Duncan et al., 1993, 231). Whether this percentage is a less dramatic number, is a difficult question because a benchmark is missing. However, if it would be significantly larger than the corresponding poverty rate in other countries or if it would increase over time in the US, it would certainly be a matter of concern.

All of these questions can be answered by using cross-sectional (income) surveys in the corresponding countries and years. Likewise, the aforementioned poverty rates could have been computed from cross-sectional surveys in the years 1980–1986. In other words: For estimates of the *level* and *trend* of poverty rates we do not need panel data. However, if someone asks how many of these poor people are also poor in the following year, cross-sectional data would not provide the answer, because more information is needed than the (aggregate) poverty rate in the following year. One must know the poverty status for each individual in the following year, which presupposes a second (repeated) measurement of the same individual's income. This kind of information measures change (and stability) at the individual level and is only available from panel data. Clearly, a situation in which a significant proportion of this year's poor individuals escapes poverty would be less of a concern than a situation in which the poor remain in poverty for a longer time. Furthermore, transient poverty may have other causes and needs other policy measures than permanent poverty.

Note that similar questions about stability and change at the individual level are asked in other fields of social inquiry, among them voting and consumer behavior where, as we have seen, panel designs were used for the first time. For example, party preferences at the aggregate level may be quite stable, but at the individual level only some voters may have stable preferences, while the majority of voters is not committed to a certain party and may change their party vote quite quickly. Obviously, political parties have an interest in strengthening the bonds to their stable electorate and to convince as many of the undecided voters, and it may be necessary to design different campaigns for both groups of voters. Similarly, producers of consumer goods are confronted with the problem of brand loyalty. On the one hand, they are interested in knowing who the loyal clients are and how to strengthen their preferences for the product. On the other hand, they want to increase their market share and for that reason they need to know how to gain new consumers.

But let us turn back to the poverty example and see what can be done with panel data. Table 1.1 shows the results for the US during the early 1980s. According to these data, 71.3 % (=9.7/13.6) of the severely poor (those below the 40 % threshold) remain in poverty in the following year. Note that this is an average of all yearly

Table 1.1 Family size-adjusted income transition tables for US American families with children (using 40, 50 and 60 percent of median income)

	<40	40–50	50–60	≥60	All
	Percent				
<40	9.7	1.8	0.8	1.3	13.6
40–50	2.1	2.0	1.1	1.5	6.7
50–60	1.1	1.4	2.1	3.0	7.5
≥60	1.6	1.5	3.4	65.6	72.2
All	14.5	6.6	7.4	71.4	100.0

Source: Duncan et al. (1993, 231) using PSID data ($n = 17,427$)

transitions between 1980 and 1986, but it would be easy to compute the corresponding percentage for a specific year, say 1984. A slightly larger (average) stability rate of 79.2 % is observed for the group of US citizens at the risk of poverty (those below the 60 % threshold). All the statistics in Table 1.1 have been estimated using data from the US American Panel Study of Income Dynamics (PSID). However, as the margins of the table demonstrate, panel data can also be used to estimate (cross-sectional) poverty rates for specific years. The right column (labeled “all”) shows that in the period from $t = 1980$ to $t = 1985$ on average 13.6 % have been severely poor and that poverty increased slightly to 14.5 % in the following years $t + 1$ (see the last row labeled “all”).

Hence, besides answering questions on *individual change*, panel data can also be used to answer typical cross-sectional questions about *level* and *trend*. In other words: panel data allow us to address all the research questions that we are used to analyze with cross-sectional data and some additional questions that cross-sectional data cannot deal with; among them the question of individual change.¹ Nevertheless, some purists argue that panel data should be used for the analysis of change only, especially so because panel data have their problems too when it comes to the analysis of long-term trends (see the problem of panel attrition below). We agree, however, with the majority of researchers who think that this would be a waste of resources. If this rich data are available, they should also be used for the analysis of levels and trends, especially so if no other longitudinal information is available. Repeated cross-section surveys are not abundant and often do not include the variables of interest.

The distinction between level and change is one of the guiding principles that structures the material presented in this textbook. Furthermore, we differentiate with respect to the type of the dependent variable that is of interest. Poverty status, party preference, and consumption pattern are called *categorical variables*, while income, political interest, and consumption expenditures are *continuous variables*. This text-

¹Of course, it is true that a cross-sectional survey can also ask retrospective questions and in doing so measure what has changed since some former point in time. However, the amount of retrospective information is usually quite limited and always prone to recall bias.

book will show how to analyze the level and change of continuous and categorical panel data.

1.1.1.2 Separating Age and Cohort Effects

When analyzing change, researchers often want to separate generational from maturation effects. While the former relate to the time when the units of interest started to exist (e.g., year of birth in case of individuals or year of foundation in case of business companies), the latter relate to the time that has passed since the starting date (i.e., the age of individuals or business companies). With a cross-sectional design it is impossible to disentangle both effects, because if one knows the age of an individual (company) at the time of measurement (t), one can easily compute its year of birth (foundation). By definition, with only one measurement both variables are perfectly related to each other: $birth = t - age$. With a panel design, on the other hand, each unit is observed repeatedly over time and hence, units belonging to the same generation are measured at different ages. Now it becomes possible to analyze how maturation (age) affects the characteristics of different generations (sometimes also called cohorts).

In principle, this analysis can also be done by combining *several* cross-sections over time (the *pooled cross-sectional design*), given we are not interested in change at the individual level. For example, a cross-sectional survey conducted in the year 2000 will include individuals from different birth cohorts, among them individuals born in 1950. Another cross-sectional survey sampling the same population in 2005 will again include individuals from the 1950 generation, however at a later age (55 instead of 50). Combining (pooling) both surveys provides us with two measures of age for the 1950 and all other birth cohorts, which also allow us to separate maturation (age) from generation (cohort) effects. However, compared to the panel design, individuals from the 1950 generation sampled in 2005 will not be the same individuals that have been sampled in 2000 (except some rare cases that incidentally have been sampled in both years). Therefore, the pooled cross-sectional design provides us only with so-called *synthetic cohorts*. Analyzing differences with respect to age with synthetic cohorts always has to control for possible chance fluctuations in these differences that are due to sampling repeatedly from the corresponding birth cohorts as is done when using several cross-sections. In case of a panel design, on the other hand, we measure the same members of a birth cohort repeatedly over time and hence, with these “true” cohort data we can make a much stronger case for maturation effects.

1.1.1.3 Controlling for Omitted Variable Bias

Another problem that ails all empirical research is the fact that we often do not know all the determinants of our dependent variables and even if we know them theoretically, we often do not have measures of them. Therefore, we always have to be aware that our models may be incomplete and our estimates possibly biased, because we have omitted important explanatory variables from our models. With cross-sectional data, there is not much we can do about omitted variable bias except make simplifying assumptions about the effects of these omitted variables. The situation is less hopeless with panel data.

As we will show in the following chapters, panel data allow us to control for at least part of this unobserved heterogeneity. The fact that we have access to repeated measurements of the same units allows us to control at least for their unknown characteristics that are constant over time. Units are used as their own controls, a technique known from experimental research as the *pre-test post-test design*. The underlying idea is the following: if a variable X influences the variable of interest Y , then a change of X at some time point t should result in a different value of Y at $t + 1$ than the value of Y at $t - 1$. Since this design compares identical units measured at $t - 1$ and $t + 1$, it also controls for all their characteristics that do not change in between.

1.1.1.4 Assessing Causality

Talking about influences and effects instantly leads to the question of causality. This introductory chapter is not a good place to discuss criteria of causality and causality assessment. Nevertheless, to understand the potential of panel data compared to other research designs, an informal definition of causality is sufficient. According to this definition, (i) two variables X and Y should correlate with each other, when they are causally related. (ii) This correlation should not be spurious in the sense that the correlation between X and Y is due to the correlation of both variables with some other (third) variables. (iii) Finally, whether X has a causal effect on Y (and not Y a causal effect on X) should be demonstrated by manipulating X and analyzing the changes of Y . At least, changes of X should precede changes of Y .

These criteria are most easily assessed with an experiment. One can manipulate X under controlled conditions and analyze whether that results in changes of Y . Other determinants of Y are controlled for by the experimental setup and by selecting units randomly into the treatment and control group (randomization).

A cross-section is the most inappropriate design to assess causality. First, it does not allow one to disentangle the time order of X and Y because all variables are measured at the same point in time. Second, in a real-life situation X is possibly correlated with other variables that cannot be controlled for because they are unknown or have not been measured.

As the discussion in the previous section showed, with panel data it is at least possible to control for those unknown or unmeasured determinants of Y that are constant over time. Moreover, since panel data include repeated measures of X and Y it is much easier to assess whether changes of X precede changes of Y or vice versa. This does not mean that all problems of causality assessment are solved with panel data, but the panel design has much more power than many other designs.

1.1.1.5 Obtaining Larger Sample Sizes

In most cases, small sample sizes are not a problem for survey researchers. Given enough financial resources, it is just a matter of time to collect data on a sample of several thousand individuals. However, social scientists interested in the quantitative analysis of macro phenomena (political systems, national economies, and so on) often have to deal with small sample sizes.

For example, scholars interested in social expenditures in modern capitalist welfare states often decide to analyze OECD countries, simply because the OECD provides so many statistics about them. At present, this population includes only 34 units (countries) and given this low number, it does not make sense to draw a sample. Such small data sets are typical for many analyses at the country level, as you find them in political science, macroeconomics and macro sociology. The limited sample size severely limits possible statistical analyses. In this case, many scholars recommend to extend the data in the time dimension and measure each (macro) unit at several points in time (a panel design). However, it is important to keep in mind that a sample of 30 units observed more than 20 times (see, e.g., the SOCX data base in Table 6.1) is not equivalent to a sample of 600 units, because repeated measurements of identical units do not provide totally independent information. Nevertheless, a panel of this size certainly provides more information than a cross-section of only 30 units.

1.1.1.6 Measurement Error

As we all know, social science data are prone to measurement error, which contaminates the statistical associations that we observe in our data to a greater or lesser extent. Therefore, we would like to have measures of the reliability of our data in order to correct our estimates of the statistical associations. One method to assess the reliability of a variable is to compare several measurements of this variable over time (*test-retest reliability*). This is easily done with panel data, while reliability analyses with cross-section data require that we have *parallel* measurements of the same underlying construct, which may be hard to defend in some cases.

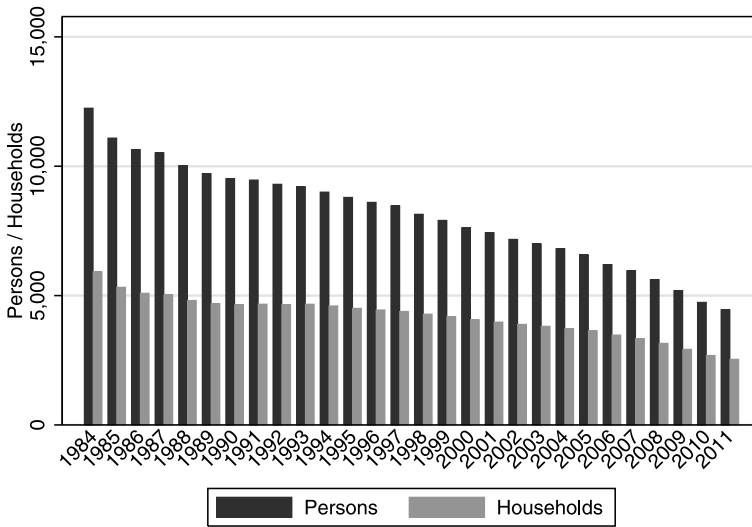
Hence, panel data are a perfect tool to examine measurement error. On the other hand, measurement error is also a challenge for panel data. If we want to analyze change, we have to deal with the problem that part of the observed change is due to measurement error. In order to achieve both a measure of reliability and a measure of “true” error-free change, we need more than just two measurements over time. Therefore, extending statistical models to cope with measurement problems is easily done with panel data, but may raise additional questions of identification.

1.1.2 Challenges

As the discussion in the previous sections has shown, a panel allows answering many more research questions than other kinds of research designs. However, it is no panacea! Naturally, a panel is also a much more complex design that leads to many new challenges when putting it into practice.

1.1.2.1 How to Represent the Population over Time?

The most prominent challenge is the issue of sampling and representing the population over time. Of course, if one studies a census of the population (like the SOCX that includes *all* OECD member states), sampling and representation are not an issue. However, most of the aforementioned panel studies use a sample of a well-defined population.



Source: Kroh (2012, 5)

Fig. 1.1 Successful interviews with persons and households (SOEP: Samples A and B)

For example, when the SOEP was started in 1984, the survey institute selected a stratified random sample of all German private households and interviewed all household members aged 16 and older. The initial sample included 5,921 households and 12,245 individuals. According to the panel design, these 12,245 individuals should have been re-interviewed every year starting from 1985. This is not an easy task. Some of them may have moved away, others may not be available in a particular year, some may refuse to continue participating and finally, a few may have died. All of these events result in missing information for some of the original sample members: either temporarily (no interview in a specific year) or permanently (dropout out of the panel). Temporarily missing information is less of a problem, because it can be imputed from the available information in the other years. The main problem arises when sample members permanently drop out of the panel. This process is called *panel attrition* or *panel mortality* and as Fig. 1.1 shows (Kroh, 2012), the number of dropouts is quite significant, especially when re-interviewing respondents for the first time (in the second wave).

What is so problematic about panel attrition? Since the SOEP is supposed to represent the 16+ population living 1984 in (West) Germany, all dropout events that cannot also happen to a member of the population are potentially harmful to the representativeness of the sample. For example, if a sample member dies and for that reason drops out of the panel, this is a personal tragedy, but from a statistical point of view it is unproblematic because it represents an event that also happens in the population. The same applies to a birth of a child, as long as it is included in the sample (as it is in the population). At age 16, the child will be interviewed for the first time. However, if a sample member cannot be contacted or refuses to participate, this is potentially harmful because in principle every member of the living population can

be contacted and no one can refuse to be member of the population. If these and other kinds of non-response are selective, then the available (non-missing) information provides a biased picture of the population.

For example, the former Table 1.1 is based on all PSID members that provided information for both years t and $t + 1$. Hence, it does not include individuals that refused to participate or could not be contacted in $t + 1$ (or in t). If these are the individuals in permanent poverty, the former stability measures will underestimate the true percentage of permanently poor. Furthermore, the level of poverty in both years t and $t + 1$ will be underestimated too. Since an unbiased estimate of these poverty rates (not the stability measures) can be obtained from cross-sectional surveys, panel attrition is a clear disadvantage of the panel design and a strong argument in support of the cross-sectional design. Although the cross-sectional design cannot answer all research questions (e.g., with respect to individual change), it is not negatively affected by selective non-response that is due to repeated measurements.

Quite generally, representing a population over time is a much more complex issue than representing a population at a given point in time. This is due to the fact that the population itself changes over time. Demographers distinguish between natural changes of the population (births, marriages, divorces, and deaths) on the one side and processes of immigration and emigration on the other.

From the very beginning, SOEP tried to represent natural changes of the population by including all “new” SOEP household members into the survey and by following all “existing” SOEP members founding a (new) household of their own. The first group includes SOEP children that are “born into” the interview age (16) and individuals moving into existing SOEP households (e.g., by marriage). The second group includes, e.g., SOEP youngsters leaving their parent households or SOEP adults that have to found a new household due to divorce. However, these inclusion rules fail if significant changes of the population happen outside existing SOEP households, as is the case if there is heavy migration into or out of the country.

Hence, besides the problem of panel attrition the panel design also suffers from significant *changes of the population* itself. To put it differently: Even if SOEP would not suffer from panel attrition and therefore, represent correctly the German population from 1984, the 1984 population is no more representative of Germany today, which now includes also the population of the former Democratic Republic of Germany and which has experienced a massive immigration of native Germans and other nationals after the fall of the Iron Curtain. A recent cross-sectional survey would not have these problems, because it would sample the present population.

In sum, a panel has the ability to answer research questions that the cross-sectional design cannot address. However, it has selectivity problems due to panel attrition and population change. Hence, to exploit the unique features of the panel design, much effort must be invested to minimize these problems of representation.

Counter measures include intensified field work (a tracking concept) to contact as many of the previously selected households as possible and to motivate as many of the former respondents to continue participating in the panel. The remaining non-response has to be either imputed (in case of temporarily missing information) or

compensated for by re-weighting the remaining units (in case of panel attrition). However, at a certain point the loss due to panel attrition will be so large that weighting the few remaining units does not make sense anymore. At that point, a *refreshment sample* is necessary. Changing populations, on the other hand, can be dealt with by drawing new samples either at regular points in time (called *rotating panels*) or when necessary (e.g., after a period of massive immigration). The EU-SILC is an example of the first sort, the SOEP immigration sample begun in 1995 is an example of the second sort (Schupp and Wagner, 1995).

1.1.2.2 How to Obtain Valid and Reliable Measurements over Time?

If repeated measurements are the main purpose of the panel design, then every effort has to be undertaken to ensure they are valid and reliable. Some scholars argue that the repetition itself may be harmful to the validity of the measures. However, a closer look at the scholarly literature on the so-called *panel effect* (*panel conditioning*) provides positive and negative views.

On the one side, it is correct that posing identical survey questions over and over again elicits stereotypical and streamlined answers. Respondents and interviewers also “learn” how to avoid difficult and time-consuming questions, e.g., by answering filter questions strategically (Van der Zouwen and Van Tilburg, 2001). On the other side, answering repeatedly the same questions over time induces also positive learning effects and attentiveness. Respondents may become more “knowledgeable”, when asked repeatedly over time the same knowledge questions (Das et al., 2011). Complicated questions referring, for example, to the various income sources of the household may be difficult in the first panel wave, but become easier after having answered them several times (Frick et al., 2006; with respect to attitudes see Sturgis et al., 2009). Hence, the panel effect may bias the measures, but also decrease non-response and increase validity. Whether and how it works has to be found out by comparing data from a panel with measurements from independent cross-sections.

Another challenge is to keep the survey instruments equivalent across time. For example, survey questions may need to be changed because their repeated application shows that they have low quality, because they become obsolete during the course of time, because they have to be adapted to the actual historical context, or because survey methods change over time (e.g., changing from face-to-face to telephone interviews). Equally, new questions have to be developed if new aspects attract the attention of researchers. Finally, even if questions are identical across time, their meaning may change over time. Overall, the practice of many panel surveys shows that longitudinal analyses are often hampered by non-equivalent survey instruments over time. All the more reason it is necessary to restrict instrument changes to the absolute minimum and to assess their equivalence at regular intervals.

Traditionally, panel measurements provide information for each unit of analysis at $t = 1, \dots, T$ *discrete* points in time. For example, Lazarsfeld’s Erie county study measured political attitudes at $T = 7$ monthly measurements (May–November) during the 1940 presidential campaign. There was no attempt to measure political attitudes *between* the seven survey dates, assuming that attitude change can be approx-

imated by monthly measurements. Similarly, the large household panels mentioned in Table 6.1 are conducted mostly on a yearly basis. However, researchers have become increasingly interested in what happens in between, also because a yearly interval between the measurements is quite large. For example, measuring income at only one point in time during the course of the year is not very useful, when incomes change rapidly due to changes in employment. In that case, panel researchers try to collect *continuous* employment and income histories either by retrospective questioning or by merging the panel information with process-produced data from other sources. Of course, retrospective questioning is not without risks due to recall bias and seam effects.

For example, the SOEP uses a monthly job calendar, in which respondents can report their employment status for each month in the year before the interview date. If they frequently change their status, they may have problems recalling all transitions with their exact dates. Furthermore, an analysis of the job calendars shows that many job changes happen at the end of the year. This is, however, often a methodological artifact at the seams of the yearly job calendars. Some respondents have forgotten what job they specified for December in the previous job calendar and report a seemingly “new” job for January in the present job calendar.

In sum, although this textbook focuses on managing and analyzing panel data, it should be stressed that the collection of panel data is a methodology of its own. Obviously, the collection of panel data includes many pitfalls that may hamper later statistical analyses. However, it should also be stressed that at the same time repeated measurements are a perfect tool to assess all kinds of measurement errors (see Sect. 1.1.1.6).

1.1.2.3 How Much Does It Cost?

Finally, the question comes up: How much does it cost in terms of money, time, and manpower? And does it not cost too much to make the effort worthwhile? Certainly, a panel is much more costly than a single cross-section. But is it really more costly than a pooled cross-section design that could also answer some of the longitudinal research questions and at various places performed better than the panel design? Both designs need resources for (i) sampling, (ii) data collection, (iii) data management, (iv) weighting, and (v) documentation. Most of these cost factors are more-or-less identical for both designs. In both cases, data have to be put into a data analysis system, weighted and documented. Perhaps data management, weighting, and documenting are a little bit more complicated for panel data, but the differences will not be significant in terms of resources needed.

What is different between both designs is sampling and data collection. While a panel, in the ideal case, only needs a fresh sample at the beginning, the pooled cross-section design needs a new sample for each additional cross-section. Furthermore, resources are needed for collecting data for each panel wave and each cross-section. This is certainly more expensive for the panel design, because a specialized tracking concept is needed to minimize panel attrition. Nevertheless, these additional field work expenses are less costly than selecting new samples for each cross-section. Hence, considering the main cost factors, the panel design does not perform as badly as one might think from the beginning.

1.2 Outline of the Book

As already mentioned, this textbook focuses on methods of managing and analyzing panel data. Hence, most of the chapters will be devoted to statistical methods of panel analysis. The only exception is the following Chap. 2 that shows how to prepare panel data for statistical analysis. In the previous sections, we have shown that panel data are used both for the analysis of level and change. Moreover, statistical models are often differentiated with respect to characteristics of the variables they focus on. Like many other statistical textbooks, we distinguish between continuous and categorical dependent variables and discuss the corresponding panel models in different chapters (Chaps. 4 and 5). Within each chapter we start with models focusing on the level of the dependent variable and then continue with models focusing on change of the dependent variable. All in all, the presented material is quite comprehensive, covering two different types of dependent variable and two modes of analyzing them. Chapter 3 shows how to describe panel data and how to decide between the different models presented in Chaps. 4 and 5. Finally, Chap. 6 concludes with some suggestions on how to do your own panel analysis. It shows you what panel data are available for secondary analysis, but gives you also some references on how to design and collect your own panel data. Moreover, it discusses typical applications in different social science disciplines and mentions other sources that you can read to know more about the specific methods that we only alluded to without discussing them in detail.

1.3 Audience and Prerequisites

There are several excellent textbooks available on panel data analysis (among others Baltagi, 2008; Cameron and Trivedi, 2005; Hsiao, 2003; Wooldridge, 2010), but all of them require a fairly good understanding of matrix algebra and advanced econometric methods (e.g., instrumental variable estimation). At an introductory level, several software and econometric textbooks also treat methods for panel data analysis (e.g., Cameron and Trivedi, 2008; Rabe-Hesketh and Skrondal, 2008; Wooldridge, 2009). However, when things get complicated these sources usually refer to the more advanced literature. Moreover, methods for categorical data are hardly treated in these introductory texts (the textbooks by Cameron and Trivedi (2008) and Hsiao (2003) are exceptions).

This textbook provides an introduction into panel data analysis that does not use matrix algebra and instrumental variables estimation. It does not only focus on linear models and least squares estimation; it also provides an introduction into maximum likelihood estimation, which is a necessary tool when modeling categorical data with non-linear models. The focus is on applications of panel models and less so on the underlying statistical theory. We illustrate all methods with real research examples from scholarly journals from different social science disciplines (sociology, political science, economics).

Readers should be familiar with linear regression and have a good understanding of ordinary least squares estimation. It is also helpful to have some experiences

with logistic regression and perhaps maximum likelihood estimation, but as already mentioned these techniques will be introduced in greater detail in this book (see Chap. 5). Naturally, restricting ourselves to such a limited set of mathematical and statistical tools implies that we cannot go into more advanced methods of panel data analysis. We hope, however, that the basic panel regression models are introduced in a way that few questions remain open and readers can go on to the more advanced literature.

The text is written without any specific software in mind to estimate these models, but certainly statistical software is needed to do this job. Stata is a perfect choice, when it comes to regression models for panel data, but other statistical software like SPSS or SAS is equally well suited, at least for the models discussed in this textbook. On the web site of this textbook you find all the data sets used in our examples accompanied by Stata syntax files that replicate our results (see Sect. 7.3).

Leaving matrix algebra and instrumental variables aside does not mean that we can refrain from mathematics. Indeed, rather than simple introductions, we want to make sure that readers understand the mathematics behind the basic panel regression models. Nevertheless, we tried to keep a simple and unified mathematical notation across all chapters. Most of it will be explained in the methodological overview (see Chap. 3) and if necessary in the method-specific chapters (Chaps. 4 and 5).

At this point you only need to know that we distinguish between variables and the values (realizations) that these variables obtain for each unit of analysis. Variables are symbolized with capital letters (Y, X, Z, T, U, E), their realizations (values) with small letters (y, x, z, t, u, e). We distinguish between dependent (Y) and independent (explanatory) variables (X, Z), process time (T), and independent (explanatory) variables that are unobserved (U, E). Realizations of these variables refer to measurements for a specific unit at a given point in time. To denote them as precisely as possible we use the indices i and t (for example, y_{it}). Estimation results are presented in tables and interpreted in the text. Estimates in the text are usually rounded and hence, slight differences between text and tables may happen because of rounding errors.

Acknowledgements This text has a long history and parts of it have served as background literature in lectures, seminars, and workshops both in Cologne and in other places. We thank the participants of these courses for their questions and comments, which helped us to formulate our ideas more precisely and hopefully in a more coherent form. Our special thanks go to our colleagues Josef Brüderl, Kenneth Bollen, Romana Careja, Marco Gießelmann, Achim Goerres, Heiner Meulemann, Henning Lohmann, Luis Maldonado, Ulrich Pötter, Götz Rohwer, and Hawal Shamon who discussed numerous versions of this text with us and contributed valuable improvements. Thorsten Meiser, Ingo Rohlfing, Elmar Schlüter, and Dirk Temme provided helpful literature references from their field of methodological expertise. A special word of thanks goes to all the people that supported our intention to write an applied textbook introducing panel analysis with real research examples from scholarly journals. Helmut Dietl, Bernd Fitzenberger, Geoffrey Garrett, Karsten Hank, Guido Heineck, David Johnson, Markus Klein, Richard Lucas, Pasi Moisio, Stephanie Moller, and John Stephens provided us with their data. Some of them will be used in this textbook, the rest will be provided on the web site for secondary analysis. We especially like to thank Jan Goebel at the Research Data Center of the SOEP and Heather Laurie at the Institute for Social and Economic Research for the permission to use anonymized versions of SOEP and BHPS data in our textbook. Evelyn Funk, Claudia Ubben, and Ravena Penning helped with typesetting

the manuscript and producing tables, figures, and the index. Donatas Akmanavičius at VTeX Book Production did the final editing. Finally, Joscha Dick rewrote all our Stata syntax files to publish them on the book's web site. Martin Spitzenpfeil prepared the data files that are available for secondary analysis on the web site. He also programmed Excel spreadsheets to illustrate examples from Sect. 7.2.

Applied Panel Data Analysis for Economic and Social
Surveys

Andreß, H.-J.; Golsch, K.; Schmidt, A.W.

2013, XV, 327 p., Hardcover

ISBN: 978-3-642-32913-5