Gayle, Vernon, and Paul Lambert. "Introduction to quantitative longitudinal data." *What is Quantitative Longitudinal Data Analysis?*. London: Bloomsbury Academic, 2018. 1–12. *Bloomsbury Collections*. Web. 26 Apr. 2023. <<http://dx.doi.org/10.5040/9781472545268.0008>>.

Downloaded from Bloomsbury Collections, www.bloomsburycollections.com, 26 April 2023, 12:13 UTC.

Copyright © Vernon Gayle and Paul Lambert 2018. You may share this work for non-commercial purposes only, provided you give attribution to the copyright holder and the publisher, and provide a link to the Creative Commons licence.

1 Introduction to quantitative longitudinal data

Tempus et maris aestus neminem expectant (Time and tide wait for no man)

The earliest known record is possibly from St. Marher, 1225

Introduction

The subject of this book is quantitative longitudinal data analysis, and it focuses on data that are collected in large-scale social surveys. The most universal definition of longitudinal social science data is any data that have a temporal (i.e. time) dimension. A rudimentary distinction is routinely made between cross-sectional designs, where data are collected at only one point in time, and longitudinal designs, where information is collected from the same units on multiple occasions. Temporal data can be collected in cross-sectional designs, and it is mildly misleading to assume that longitudinal analyses can never be undertaken with cross-sectional data.

Longitudinal social science data are critical to understanding social change over time. They are also critical to understanding social stability, but this is often overlooked. The analysis of quantitative longitudinal data has made a number of important intellectual contributions. Here are three examples. First, the analysis of longitudinal data on British children revealed a link between smoking during pregnancy and subsequent child development (see Butler and Goldstein, 1973). Second, the Whitehall Study revealed that inequalities in health were not limited to the health consequences of poverty and demonstrated that occupational hierarchy was intimately related to health and life chances (Marmot and Brunner, 2005; Marmot, 2004). Third, in the early 1990s much of the analyses of income in Britain focused on poverty and social inequality. Inequality and poverty rates then flattened off, and it appeared that there was little or no change in the income distribution from one year to the next. The analysis

of longitudinal household data revealed that underneath the apparent stability there was a hidden flux. Household incomes fluctuated between one year and the next, and there was substantial turnover in the membership of the low-income population. This was well known in countries like the United States and Germany, which already had longitudinal household surveys. These new findings made a contribution to the understanding of poverty and to the development of new economic theory. They also influenced the Labour government's welfare reforms. The concept that household poverty is dynamic now influences the way living standards are measured and monitored in Britain (see Jenkins, 2000).

Many social science research questions can be adequately answered using cross-sectional data. Most social science research projects can be improved by incorporating suitable longitudinal data. Some social science research questions can only be sensibly answered using longitudinal data. The following chapter introduces concepts associated with analysing large-scale quantitative social science data and highlights the rich variety of data resources that are available to researchers. In this book we focus upon observational social survey designs, but most of the issues can be directly extended to other forms of data, for example administrative data or experimental designs.

Cross-sectional social surveys

Cross-sectional social surveys are often overlooked as a source of longitudinal information. The majority of the data in a cross-sectional social survey will relate to only one point in time. Cross-sectional social surveys sometimes collect a small amount of information that has a temporal dimension. For example, a cross-sectional survey might collect information on a respondent's current occupation and also on their father's occupation when they themselves were aged 14. While the information is collected at only one point in time, the measures still contain a temporal dimension and can inform analyses of social change (i.e. intergenerational occupational mobility). Similarly a respondent in a cross-sectional survey might be asked about their current job and also about their first job. These data could effectively be used to analyse intragenerational occupational change. The point here is that while the survey is cross-sectional, the measures collected contain some temporal information.

Many cross-sectional surveys are carried out on multiple occasions. These surveys are not based on repeated contacts with the same individuals (or households), but they offer the possibility of comparing similar data for different points in time. Combining data from repeated cross-sectional surveys is a particularly effective way to study trends over time. There are numerous large-scale repeated cross-sectional surveys. A notable example is the Labour Force Survey, which has been carried out in the UK since 1979 (and for which a similar survey is undertaken in many other countries). Further examples are the General Household Survey, which commenced in the UK in 1971, and the Family Expenditure Survey, which dates back to the 1950s. The US Current Population Survey is another example of a repeated cross-sectional survey.

For many social science analyses, cross-sectional surveys provide highly appropriate temporal data on social patterns. Repeated cross-sectional surveys offer opportunities to analyse macro-level trends, sometimes extending over lengthy periods of time. They do not, however, offer any insights into micro (i.e. individual)-level changes over time. This is because cross-sectional surveys generally do not include information that links up individuals in the surveys at different time points. We will return to the analysis of multiple cross-sectional surveys for quantitative longitudinal social science data analysis in Chapter 3.

Longitudinal social surveys with repeated contacts

Central to the collection of longitudinal social survey data is the concept of a panel. A panel is a sample of respondents who are contacted and surveyed on multiple occasions. The idea of researching a panel of respondents was pioneered by sociologist Paul Lazarsfeld for opinion research in the 1930s (see Lazarsfeld and Fiske, 1938). The sample unit will usually consist of individuals, but could also consist of households, firms, farms, schools or hospitals, or any other unit of social science research interest.

A cohort study is a special type of panel study. It is principally concerned with charting the development, or progress, of a particular group from a certain point in time. A notable example is a birth cohort. This form of study tracks or charts the development of a group of babies born in a particular year and follows them through childhood and into adulthood.

In Britain we are richly furnished with birth cohort datasets and we will outline the main studies later.

Not all cohort datasets track individuals from birth. For example, the Youth Cohort Study of England and Wales (YCS) monitored the behaviour and decisions of representative samples of young people aged 16 and upwards as they made the transition from compulsory education to further or higher education, or into the labour market (Gayle, Murray and Connelly, 2016). The British Medical Association Cohort Study of Medical Graduates (2006) followed a group of new doctors, who were all at the same career stage, and tracked them from qualification through the first ten years of their professional careers.

The term 'balanced panel' is used to describe a dataset where every respondent has the same number of repeated contacts. Balanced panels are more common in experimental situations where the study is specifically designed to collect the same number of data points for each participant. 'Unbalanced panels' are ubiquitous in longitudinal social surveys. This is because routinely some individuals leave the study, new participants enter the study and some participants will not have provided data at every wave. Unbalanced panel data are far more usual in the social sciences, therefore in Chapter 5 we will concentrate on methods that can be applied to unbalanced repeated contacts data.

Modes of survey data collection

It is useful to distinguish between two forms of temporal data collection. Prospective data are collected as the study moves forward. Measures are collected at each subsequent contact, for example in annual interviews. By contrast, retrospective data contain historic or 'back-dated' information. For example, when a new study begins, participants might be asked to supply retrospective information on their employment history from the time that they entered the labour market up to their present job. Most of the large-scale longitudinal social surveys have prospective designs, but in practice also collect some retrospective data.

Historically, interviews and traditional questionnaires were common in large-scale longitudinal studies but multiple modes of data collection are now usual. Data collection has been enhanced by computer-aided interviewing techniques, and electronic communication strategies

are increasingly being used to collect data in large-scale longitudinal studies.

The Research value of repeated contacts quantitative longitudinal data

At the beginning of his insightful chapter on longitudinal social science data analysis, Davies (1994) states that researchers are more likely to be encouraged to analyse longitudinal data if they understand its research value. In this section we will introduce the main research benefits of quantitative longitudinal data.

Micro-level dynamics

We will begin with a vignette.

Jason Jones is aged 10. One day he comes home from school and he notices that his Dad's car is not outside of his house. Jason goes into his house and finds his Mum standing in the kitchen crying. Jason's Mum informs him that his Dad has left them. Jason's Dad rapidly loses contact with the family and does not provide any financial support. They are plunged into poverty and family life becomes very difficult.

One day, a few years later, Jason returns from school and he notices that an unfamiliar car is parked outside of his house. He enters his house and his Mum introduces him to her new 'friend', a man called Mike. Mike is very friendly and starts spending time at Jason's house. After a while he buys Jason an Xbox. Later that year Jason and his Mum go on holiday with Mike. Shortly after, Mike comes to live with Jason and his Mum, and their financial position improves dramatically. Life for Jason and his Mum gets much better.

One day, about a year later, Jason returns from school and notices that Mike's car is not outside. He enters his house and finds his Mum is in the kitchen crying. Jason glances into the sitting room and notices that his Xbox is gone.

It is easy to guess what has happened, and we can speculate on the likely changes to Jason and his Mum's life and their financial situation.

This snapshot into the life of young Jason Jones is disheartening, but we have chosen it because it provides a vivid illustration of our first methodological point. Imagine that data were collected on the Jones family via a cross-sectional survey at any time during the period of this vignette. The cross-sectional survey would accurately collect information on the family's current circumstances, and a little retrospective information might also be collected. It is unlikely that with cross-sectional information a researcher could genuinely understand the journey that Jason and his mother have been on. Understanding the micro-level social processes that are associated with the flows into and out of childhood poverty is an example of a research area that cannot be properly studied without comprehensive temporal information.

Many other social phenomena are inherently temporal in their nature. For example, family migration is intrinsically linked to the movement from one place at a certain time to another place at a later point in time. Nowok et al. (2013) is one example of a study that uses repeated contacts data to study family migration. This study found that on average UK migration was preceded by a period when individuals experienced a measureable decline in happiness for a variety of reasons and migration itself caused a boost in happiness, and brought people back to their initial levels.

Another situation where temporal information is required is the study of individual development. Bergman and Magnusson (1990) describe longitudinal data as the lifeblood of the study of individual development (e.g. in childhood), and it has been pointed out many times that the most important questions concerning individual development can only be answered by adopting a longitudinal design whereby the same individuals are followed through time. Turner, Alborz and Gayle (2008) provide an example of the research value of repeated contacts data for studying individual development. Their study analysed data on children with Down's syndrome and their families. Their findings suggest that notwithstanding the dominant effect of the severity of intellectual impairment, a number of factors within and outside the family may also contribute to higher attainment in reading, writing and numeracy. In particular, mainstream schooling for those with less severe disabilities appears to have benefitted the children in this study.

Temporal ordering of events

The ‘arrow of time’ flies in only one direction. Experiences at nursery school could affect a young person’s chances of going to university. Teenage smoking could influence health in old age. We cannot conceive of circumstances where university entry could influence experiences at nursery school or how health in old age could influence smoking behaviour decades before. The reversal of the arrow of time would render the external social world nonsensical. Repeated contacts longitudinal data help facilitate the accurate temporal ordering of information, which can inform narratives about the direction of influence. Research studies have routinely taken advantage of the ‘arrow of time’ and have used longitudinal data to investigate the direction of social influences.

The association between ill health and unemployment serves as a useful example (see Bartley, 1994). This association is consistent with both the theoretical claim that unemployment causes ill health and the alternative theoretical claim that ill health causes unemployment.

Table 1 shows a hypothetical panel of data on ill health and unemployment. Repeated contacts data were collected each month (for 12 months) from Person A. This respondent had relatively good health at the start of the study. Their level of health was 17 on a scale where 20 is very good health. Person A was employed for the first three months of the study and

Table 1 Hypothetical data on ill health and unemployment (Person A)

Month	Level of health (20 = very good health)	Employment status
1	17	Employed
2	17	Employed
3	17	Employed
4	17	Unemployed
5	17	Unemployed
6	10	Unemployed
7	9	Unemployed
8	5	Unemployed
9	4	Unemployed
10	3	Unemployed
11	2	Unemployed
12	1	Unemployed

became unemployed in month 4. We observe that their health started to decline in month 6 and deteriorated to level 1 by the end of the study. This pattern is consistent with the theoretical claim that becoming unemployed negatively affected their health.

Table 2 shows another hypothetical series of data on ill health and unemployment. Repeated contacts data were collected each month (for 12 months) from Person B. This respondent also had relatively good health at the start of the study. Their level of health was 17 on a scale where 20 is very good health. Person B was employed for the first three months of the study and became unemployed in month 4. We observe that their health started to decline in month 2 and deteriorated to level 1 by the end of the study. This pattern is consistent with the theoretical claim that their ill health affected their employment status.

Despite the very different temporal ordering of events, if both of these people were included in a cross-sectional survey in month 12, Person A would have been unemployed for 9 months and have a health score of 1 and Person B would also have been unemployed for 9 months and have a health score of 1. This is an obvious example of how panel (i.e. repeated contacts) data can make an essential contribution to untangling micro-social processes.

Table 2 Hypothetical data on ill health and unemployment (Person B)

Month	Level of health (20 = Very good health)	Employment status
1	17	Employed
2	1	Employed
3	1	Employed
4	1	Unemployed
5	1	Unemployed
6	1	Unemployed
7	1	Unemployed
8	1	Unemployed
9	1	Unemployed
10	1	Unemployed
11	1	Unemployed
12	1	Unemployed

Residual heterogeneity

There is another important benefit of repeated contacts data, which is more subtle and less immediately obvious. Data collection instruments often fail to capture the detailed nature of social life, and there is, almost inevitably, considerable variation (or heterogeneity) in response variables even among respondents that share the same characteristics across all of the explanatory variables that are included in an analysis. This is sometimes referred to as 'residual heterogeneity'. An alternative term is 'omitted variable bias', because hypothetically the heterogeneity could be captured by other important measures that should have been included in the model. A technical account of this issue is provided by Hedeker and Gibbons (2006).

Within social science studies some variables will inevitably be unmeasured so they cannot be included in an analysis. Some variables may even be unmeasurable in the context of a survey and cannot ever be collected (and therefore cannot be included in an analysis). As long as we make the assumption that (at least some of) these residual effects are enduring, there are techniques for providing increased control for the effects of omitted explanatory variables if we have repeated contacts data. We will elaborate on these techniques later.

State dependence

Economist J. J. Heckman stated in his Nobel lecture that a frequently noted empirical regularity in the analysis of unemployment data is that those who were unemployed in the past are more likely to be unemployed in the future and those who worked in the past are more likely to be working in the future (Heckman, 2001). Much of human behaviour is influenced by previous behaviour and outcomes (i.e. there is positive 'feedback'). This is known as 'state dependence'.

State dependencies are relatively easy to conceptualize. For example, if you are employed in May you are more likely to be employed in June. Similarly if you are currently married you are very likely to be married next year, and if you travel to work by car this week you are more likely to travel to work by car next week. McGinnis (1968) outlined the 'axiom of cumulative inertia'. The axiom states that the probability of remaining in any state increases as a strict monotonic function of the duration of prior residence in that state. Panel data open up the possibility of explicitly including past

behaviour (and duration in a state) within a statistical modelling framework in order to better understand substantive outcomes.

Age, period and cohort effects

It is a stylized fact that older people are more conservative in their political attitudes (see Glenn, 1974). Do people simply become more conservative as they age? In the contemporary era, is it just that a conservative attitude fits more naturally with the interests of older people? Or have older cohorts been raised in less liberal times and maintained their more conservative political attitudes? A further benefit of longitudinal data is that they facilitate the exploration of the effects of three distinct but related temporal processes; these are ageing, time period and cohort membership. In this context the term 'cohort' describes a common group that is being studied. The term 'age' describes the amount of time since the cohort was constituted. The term 'period' describes the moment of observation. As Yang and Land (2008) assert, a common goal of age-period-cohort analyses is to assess the effects of one of the three factors on some outcome of interest, net of the influences of the other two time-related dimensions.

Age effects represent the variation associated with different age groups that is brought about by changes in circumstances (e.g. the accumulation of social experience, or human capital). This contrasts with period effects, which affect all age groups simultaneously (e.g. a societal change or a change in a policy). Cohort effects are associated with changes across groups of individuals who experience an initial event at the same time (e.g. they are all born in the same year or qualify as doctors at the same time). Cohort effects may be present because individuals had the same formative experiences (e.g. they grew up in Britain immediately after the Second World War, or they attended medical school at about the same time and received similar training).

The tangled effects of ageing, period and cohort membership can be present in both longitudinal and cross-sectional datasets. In general it is possible to make some inroads into distinguishing between these effects with data that have information on people from the same cohort at different ages. Statistical models that attempt to separate the independent effects of age, period and cohort encounter inherent identification problems; this is because $age = period - cohort$, and therefore, knowing the value of any two of these time-related factors can automatically determine the value of the third. This creates a serious problem in the

attempt to decompose the effects of the three time-related influences in a single analytical framework (e.g. a standard statistical model) (see Harper, 2015; Glenn, 1976). More recently a number of advanced techniques have been proposed to help to deal with the problem of identification in age-period-cohort models. Yang and Land (2008) assert that advanced methodological guidance is needed to address the fundamental question of how to determine whether the phenomenon of interest is cohort based or whether some other factors, such as age or calendar year, are more relevant.

Conclusions

The research value of quantitative longitudinal data lies in the variety of important analytical contributions that they can make to addressing social research questions. Temporal data are clearly required to support statements about social trends and macro-level change. The capacity of repeated contacts data to better elaborate micro-level social processes is an obvious attraction. The ability to place events in a correct sequence and therefore to enable a better understanding of their temporal ordering is another key attraction of panel data analysis. The capacity to support analyses that have improved control for residual heterogeneity and state dependence is also a desirable feature of panel data. In principle it is possible to make some inroads into distinguishing between age, period and cohort effects with longitudinal data, but in our experience this is not always straightforward in practice.