Gayle, Vernon, and Paul Lambert. "Adopting the long view: A review of analytical methods." *What is Quantitative Longitudinal Data Analysis?*. London: Bloomsbury Academic, 2018. 111–118. *Bloomsbury Collections*. Web. 26 Apr. 2023. <<http://dx.doi.org/10.5040/9781472545268.0013>>.

Downloaded from Bloomsbury Collections, www.bloomsburycollections.com, 26 April 2023, 09:34 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.

6 Adopting the long view: A review of analytical methods

The main message

In the absence of data with a temporal dimension there is little that can reliably be concluded regarding social change or social stability. For many research questions cross-sectional data will be sufficient, but the answers to most research questions can be valuably enhanced with the addition of some longitudinal data analyses. Some questions can only be sensibly and comprehensively answered using longitudinal data.

The generality and flexibility of large-scale longitudinal data resources facilitates their use in a wide variety of circumstances. Repeated cross-sectional surveys often provide useful sources of information, particularly when the main research focus is describing longer term social trends. Cohort studies tend to be well suited to the study of growth or development in specific contexts. Cross-cohort comparisons are especially effective when the research focus is investigating generational differences. Household panel studies usually collect a broad spectrum of data, and they tend to be suited to studying micro-level social change especially over shorter time periods. Sometimes it is also possible to construct pseudo-cohorts from household panel studies to study growth or development.

There are a number of methodological benefits to using repeated contacts data in social science research. Repeated contacts data provide detailed information on transitions and changes experienced by individuals, and they ensure that researchers accurately disentangle the temporal ordering of events by charting exactly how events unfold in the individual's lifecourse. Repeated contacts data also support the estimation of statistical models that provide improved control for residual heterogeneity (or omitted explanatory variables). Residual heterogeneity is almost inevitable because within analyses some important variables will not have been included, either because the measure was not available or because it was not possible to measure it in a survey context. There is a frequently noted

empirical regularity that current behaviour is influenced by previous behaviour. Repeated contacts data allow models to be estimated that explicitly provide improved control for previous behaviour (state dependence).

Using existing datasets

Longitudinal data are time consuming and expensive to collect and we would always advise researchers to begin by investigating existing longitudinal data resources before embarking on collecting new longitudinal data. There is now a broad portfolio of survey data available to social science researchers. The existing large-scale longitudinal datasets are often 'omnibus' studies that collect hundreds of measures on a broad spectrum of topics.

A large-scale longitudinal data resource might not have exactly the variable that a researcher is looking for to undertake a particular analysis. In these circumstances it is often good advice to compromise a little and we would advise the researcher to think hard about what they wish to achieve. For example, imagine a situation where one option is to use a variable from the BHPS that is less than fully optimal. Another option is to begin a new data collection exercise attempting to collect a more appropriate or optimal variable. The BHPS provides nearly two decades of systematically collected data on a very large sample of individuals and households. Measures within the BHPS are robust and their reliability and validity have been investigated and tested. In this example there seems little contest. This is not to argue that large-scale social surveys are a 'data panacea' for all research questions, but they should be preferred whenever they provide relevant data.

Statistical models for quantitative longitudinal analyses

The known universe of statistical techniques is constantly expanding, and the array of techniques may seem bewildering even for more experienced researchers. It is valuable to have a broad understanding of the most important data analytical options, and in our review we have discussed some of the most prominent approaches to longitudinal data analysis in the social sciences. It is good advice not to stray too far from established conventions. Our observation is that many of the most successful applied

social researchers are relatively pragmatic in their use of statistical methods and focus on those approaches with a tradition of relevance to their own field of social science research. There are always new and emergent ideas in social statistics, but some of the more esoteric developments may not provide a suitable intellectual return on the outlay of erudition required to gain sufficient expertise to apply them.

The positive message is that most researchers undertaking longitudinal data analysis with large-scale social science datasets will use a core of techniques that shelter under the umbrella term 'generalized linear mixed models' (glmm). The positive message is that these models are the natural extensions of models in the generalized linear model (glm) family (which include standard regression models and logistic regression models). Models within the glmm family are flexible enough to be well suited to an enormous range of empirical research questions using longitudinal social science data.

The technical work outlined in Clark and Linzer (2015) and Bell and Jones (2015) provides persuasive methodological evidence of the flexibility and generality of random effects panel models from the glmm family. For many social science research questions where large-scale repeated contacts data are used, a random effects panel modelling approach will probably be adequate because of its generality. There are also many scenarios where the fixed effects panel model is more compelling. A strong account of the statistical foundations of panel models and their application in social science research is provided by Halaby (2004). Researchers requiring further practical advice on estimating panel models using Stata should consult Longhi and Nandi (2014).

Recent advances in estimating generalized linear mixed models mean that increasingly they can also incorporate information on the design of complex social surveys (e.g. clustering, stratification and sampling weights). Many models can be conveniently estimated using Stata's *meqglm* suite, which estimates multilevel mixed-effects generalized linear models and can be combined with the *svy* suite.¹ This brings together two previously disparate advanced statistical areas that are potentially pertinent to the analysis of the BHPS, the UKHLS and the MCS, all of which have

¹ A brief overview and a succinct example are provided on the Stata website <http://www.stata.com/new-in-stata/multilevel-models-survey-data/> (accessed 24 June 2016).

complex designs (see Taylor et al., 2010; Lynn, 2009b; Lynn and Kaminska, 2010; Ketende and Jones, 2011a).

Further methodological issues

This book is part of the *What Is* series that focuses on introducing readers to new methodological areas. Due to its orientation and length there are a number of more advanced topics that have been strategically omitted. These topics include the estimation and interpretation of interaction effects in statistical models (see Ai and Norton, 2003; Norton, Wang and Ai, 2004; Mitchell and Chen, 2005), post-estimation measures and model evaluation (see Long and Freese, 2014) and missing data (see Carpenter and Kenward, 2012). A range of advanced methodological issues relating specifically to quantitative longitudinal social surveys (e.g. sample attrition, panel conditioning, interviewer effects and data collection modes) are also not covered (see Lynn, 2009a).

Speculation on the future of longitudinal social science datasets and other emerging resources

An emerging area that we are confident to speculate upon is administrative social science data. Administrative data are records and information that are gathered in order to organize, manage or deliver a service (Elias, 2014). Although they are not primarily collected for research, some administrative data resources contain information that has great potential for longitudinal social research. Historically, social scientists have had very limited access to administrative data, but there is increasing pressure to improve its availability. In the UK, initiatives such as the Economic and Social Research Council (ESRC) funded Administrative Data Research Network² seek to improve access to data that are routinely collected by government departments and other agencies. Administrative social science datasets often have a temporal aspect and repeated observations are made on the same units (e.g. individuals, households, schools etc.). This makes them suitable for longitudinal analyses. Administrative datasets are not the product of systematic social science data collection activities, rather they

² See <http://www.adrn.ac.uk/> (accessed 6 June 2016).

are the by-product of some other process (see Harford, 2014; Connelly et al., 2016c). Therefore, administrative social science data tend to be far messier than other social science datasets that are specifically collected and then curated to facilitate research.

In comparison with large-scale social surveys, especially omnibus surveys such as the UKHLS, which are specifically designed to support multi-disciplinary data analyses, the number of available explanatory variables in most administrative datasets is extremely limited. For example, variables measuring a person's ethnicity or level of education, which are key variables in many analyses, might be completely absent in a dataset because they are irrelevant for the administrative process that the data are collected for.

Administrative data offers great potential for social science research, but our view is that it is likely to prove most effective when linked to well-designed longitudinal studies. Progress is already being made and a range of administrative sources including welfare benefits, health and education have been linked to some large-scale surveys (Sala, Burton and Knies, 2012). The UKHLS has already linked administrative data and there are plans to include further administrative data on education, health, economic circumstances and transport, and data on other topics are likely to be included in the future. In the longer term the UKHLS is also likely to be enriched by the collection of a range of data that go beyond the constraints of the standard survey interview. These activities may include using the survey as a sampling frame for qualitative longitudinal research, using diary methods to collect information (e.g. on time use) and developing experiments (Buck and McFall, 2012). Ermisch et al. (2009) provide an interesting example of an experimental design on trust that was developed using a sampling frame of households that were formerly members of the BHPS.

Some methodologists have suggested that the increasing use of administrative data and 'big data' more generally might lead to a transcendence in methods of statistical analysis (Playford et al., 2016). There is absolutely nothing that convinces us that a longitudinal analysis drawing upon administrative data should ignore the important lessons that have emerged from many decades of research into longitudinal analytical methods in social statistics and statistically orientated areas of social science such as econometrics.

A new phase in quantitative longitudinal social surveys is the collection of biomeasures (which are also referred to as biomarkers). The main

motivation for the collection of these measures in surveys such as the UKHLS is to augment survey data with objective health measures and genetic information to better support investigations of biological and environmental influences on human behaviour and health (McFall et al., 2012). Further information on the collection of biomeasures in the UKHLS is provided in McFall, Conolly and Burton (2014), and Davillas et al. (2016) is one empirical research example using these data. Biomeasures have also been included in major longitudinal studies such as the ELSA (see Read and Grundy, 2012) and the NCDS (1958 birth cohort) (see Calvin et al., 2011).

The UK has an impressive set of large-scale birth cohort studies with national coverage. The birth cohort studies support a wide range of analyses associated with development and change (or stability) over the lifecourse. The Life Study was the name given to the next UK birth cohort study which aimed to recruit babies born between 2014 and 2018. In October 2015 the UK ESRC and the UK MRC agreed that funding for the Life Study was discontinued due to the serious challenges encountered in recruiting participants.³ The future of the existing UK birth cohort studies appears to be secure at the present time, but the establishment of further new cohort studies is currently unclear.

Conclusions

Quantitative longitudinal data analysis is essential for researchers who are serious about providing comprehensive answers to research questions that require, or would benefit from, a longitudinal perspective. The analysis of appropriate quantitative longitudinal data will provide intellectual returns for the extra analytical efforts that inevitably are required. On many occasions we have found quantitative longitudinal data analysis to be an all-absorbing activity, and sometimes it has even been enjoyable. Nevertheless, the endeavour itself is not without its challenges, and Angrist and Pischke (2008) playfully remarked that if applied research was easy then theorists would do it. They also reassure readers that applied research is not as hard as the dense pages of *Econometrica* might lead us to believe.

Some of the challenges of analysing longitudinal data lie with the complex array of data analysis techniques through which we have tried to

³ See <http://www.lifestudy.ac.uk/about/faqs> (accessed 24 June 2016).

navigate in the preceding pages. Many of the challenges relate to issues of organizing and understanding data and preparing it for subsequent analysis (i.e. 'enabling data'). Quantitative longitudinal social science data are messy and confusing, and that is because the contemporary social world is messy and confusing. The deeper you dig into quantitative longitudinal data, the more tricky and bewildering it becomes. This need not be a permanent or insoluble state and given a little time and modest effort it is possible for researchers to increase their understanding of a dataset.

The skills required for enabling data, managing the workflow and undertaking longitudinal data analyses are initially perplexing. In the closing chapter we hope to assist readers who want to start analysing longitudinal data, primarily by focusing on the practical issues of enabling and preparing complex datasets for analyses.