Edinburgh Research Explorer

Modelling key variables in social science research

Modelling key variables in social science research: Introduction to the special section

Background

There is general agreement between methodologists, research societies and research funders that high-quality empirical social science needs high-quality data sources and an appropriate supporting infrastructure (UK Data Forum, 2013). In the United Kingdom, and in many other nations, there is an increasing volume of large-scale social surveys designed to support secondary data analyses, which are made available to researchers through national data archives[1]. These surveys can either be cross-sectional or longitudinal and are usually conducted at either the household or individual level. In the United Kingdom, in particular, the substantial scale and extensive supporting infrastructure of many large-scale multi-purpose social surveys makes them especially appealing for sociological research.

High-quality research information can only be derived from survey data resources if their measures and variables are used with appropriate consideration. A major characteristic of the omnibus surveys is that they regularly collect an array of the same 'key variables' which measure concepts that are central to a wide range of social science inquiries. Burgess (1986) defines 'key variables' as measures that are regularly collected in different social surveys, and that are almost always of relevance as explanatory measures. We draw on this conceptualisation throughout this special section, which concentrates on three key variables, occupations, education and ethnicity.

As far back as the 1950s, social scientists have issued warnings and guidance on using key variables in secondary analyses (see Blumer, 1956; Bulmer et al., 2010; Burgess, 1986; Stacey, 1969). More recently, the volume of social science survey data available to researchers has dramatically increased. The Internet has provided an unparalleled global facility for delivering survey data to secondary analysts, and for allowing researchers to share results. Desktop computers have become both quicker and more powerful and generally have large storage capacities. At the same time, techniques for analysing data in a multivariate framework have galvanised. Standard data analysis software packages have become much more advanced and incorporate the functionality necessary to organise and manage large-scale datasets, and to undertake analyses using advanced statistical techniques. This changing landscape has inspired us to return to the issue of analysing 'key variables' in secondary survey data analyses.

Overview of papers

This special section comprises four papers. The first three focus on specific key variables (occupations, education and ethnicity). The final paper corals a series of pragmatic and technical methodological issues and provides some recommendations for researchers using key variables in secondary social survey data analyses.

The first paper in this section focuses on occupation-based measures of socioeconomic positions. Occupations are a key element of contemporary social life and occupation-based indicators are central to sociological research. It begins with a review of alternative strategies for measuring occupations. We then introduce a series of issues associated with using occupational measures in sociological research. A central recommendation is that researchers should use existing occupation-based measures appropriately and avoid deploying them in an *ad hoc* manner. We also advise that researchers should not develop new measures without a strong justification, and in these circumstances, the new measures should be transparently documented.

The second paper in the section focuses on education. Measures of education are routinely incorporated into analyses of a wide variety of social outcomes and in analyses of social and population change. Education is a powerful explanatory factor influencing a number of economic phenomena, most notably both participation and success in the labour market. Education is also important in far less obvious fields such as health. Measuring education appropriately is more difficult than researchers might initially assume, because there is no simple, universal or agreed upon measure of education. Most societies have complex educational systems that have often changed over time and the seemingly prosaic activity of measuring an individual's education within a social survey is far from straightforward.

The third paper focuses on the key variable of ethnicity. Ethnicity is frequently taken to represent a self-claimed or subjective identity linked to a perception of shared ancestry as a result of some combination of nationality, history, cultural origins and possibly religion. There is an extensive literature which discusses the meaning and use of the term ethnicity and how this concept differs and overlaps with the neighbouring concepts of race and national identity. A central aim of this paper is to provide information relevant to using these measures for survey data analysts who are not experts in the field of ethnicity.

While there are many texts orientated towards technical statistical analysis, there are relatively few which focus on the more practical activities that are routinely associated with the secondary analysis of social survey datasets. There is usually little or no discussion of the issues surrounding selecting key social science variables, assessing their scope and limitations and including them in statistical models. Instead, many textbooks use simplified examples of social science variables to aid clear communication. In genuine secondary analyses of large-scale social survey datasets the researcher is likely to encounter a number of challenges when incorporating key variables into their sociological analyses. The aim of the final paper is to highlight several issues which are of generic importance for good quality statistical modelling in social survey analyses. We deliberately focus on interpreting the effects of key variables within the framework of non-linear regression models, since these models are common in sociological research. We also illustrate some alternative strategies for reporting and communicating the results of statistical models that include key variables.

A theme that runs through each of the papers is the value of undertaking 'sensitivity analyses'. We adopt the term sensitivity analysis to describe the process of systematically evaluating alternative social science measures, for example, different operationalisations of key variables, and exploring the influences that minor perturbations in the statistical modelling process have on substantive results. Although sensitivity analyses are recommended by some methodologists (see Dale, 2006; Treiman, 2009), this aspect of the data analysis process is often overlooked. An overall message is that sensitivity analysis must become a more prominent part of the workflow (or research process) within sociological analyses of large-scale social surveys. We also argue that sensitivity analyses should routinely be made public, for example, in data supplements or on websites. The publication of sensitivity analyses is one aspect of our wider recommendation that secondary data analysts should engage in providing clear and accessible documentation that supports their research. This recommendation chimes squarely with wider calls for increased openness in research. We argue that social science is incremental, and therefore clear and consistent documentation provides suitable building blocks that are essential for replication.

We hope that the reviews of occupations, education and ethnicity provided in this special section convince the reader that the optimal operationalisation of key variables in social survey research does not happen automatically, and should be treated as a serious part of the analytical process. An aim of the later paper on '*Statistical Modelling of Key Variables*' is to offer some useful practical prescriptions on modelling key variables in sociological research. The material presented in this special section updates earlier work on key variables in light of recent developments in survey datasets, statistical methods, and computing and infrastructural resources. We don't, however, expect this to be the last word on the subject.

Note

1. The UK Data Archive is curator of a large collection of digital data in the social sciences with several thousand contemporary and historical datasets (see http://www.data-archive.ac.uk/).

References

Blumer H (1956) Sociological analysis and the "variable". *American Sociological Review* 21(6): 683–690.

Bulmer M, Gibbs J and Hyman L (2010) Introduction. In: Bulmer M, Gibbs J and Hyman L (eds) *Social Measurement Through Social Surveys: An Applied Approach*. Farnham: Ashgate, pp. 1–9.

Burgess R (ed.) (1986) *Key Variables in Social Investigation*. London: Routledge & Kegan Paul.

Dale A (2006) Quality issues with survey research. *International Journal of Social Research Methodology* 9(2): 143–158.

Stacey M (ed.) (1969) *Comparability in Social Research*. London: Heinemann.

Treiman DJ (2009) *Quantitative Data Analysis: Doing Social Research to Test Ideas*. San Francisco, CA: John Wiley & Sons.

UK Data Forum (2013) *UK Strategy for Data Resources for Social and Economic Research 2013–2018*. London: UK Data Forum.

**Roxanne Connelly[1], Vernon Gayle[2]
and Paul S. Lambert[3]**
[1]The University of Warwick, Coventry, UK
[2]The University of Edinburgh, Edinburgh, UK
[3]University of Stirling, Stirling, UK