CITY UNIVERSITY LONDON
EST 1894

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Ipsos MORI
Social Research Institute

Tackling survey nonresponse:

The role of geocoded auxiliary data

Thursday 26th May 2016   17:00 – 19:15

London Art House

Low response rates - and the potential this has to lead to bias - are one of the major challenges facing survey research today. One commonly suggested approach to address non-response is to append auxiliary data available for both respondents and non-respondents to the dataset. However, despite the growing array of data that can be appended to sample frames using geocodes such as postcode, doing so can often be a time-consuming and frustrating task and research to date has struggled to identify auxiliary variables which are sufficiently correlated with both response propensity and the survey variables of interest to be useful.

This workshop provided an opportunity to hear findings from the ADDResponse Project - which investigates the scope that auxiliary data provide to understand and overcome nonresponse bias in the European Social Survey (ESS) in the UK - and discuss their implications for survey research and practice. Incorporating auxiliary data from a variety of sources and at multiple levels of aggregation - including small-area government data, household level commercial data and local geographic information - the project provides a uniquely in-depth insight into the potential afforded by geocoded auxiliary data to address the problem of nonresponse in social surveys.

Here we summarise the main findings emerging from the **Presentations**

and subsequent **Panel Discussion Presentations**

Sarah Butt and Kaisa Lahtinen
ADDResponse: Getting to grips with different types of auxiliary data: Was it worth it?

Kathrin Thomas and Rainer Schnell
ADDResponse: Predicting nonresponse with small area auxiliary data

Some of the main findings presented include:

- It is possible to append a wide range of auxiliary data from different sources to survey data using geocodes such as postcode.
- Using auxiliary data presents a number of challenges including: missing data, differences in geocodes (and data sources) over time and across countries within the UK, inconsistencies in the timing of data collections and incomplete documentation of data.
- Auxiliary data appear to do a poor job of predicting survey non-response. This may be down to the limitations of the data available.   It may also provide reassurance that there is no systematic bias in survey nonresponse across areas.

- There is a lot of small-area data available for analysis, with many variables providing highly correlated measures of similar dimensions. Isolating the best predictors of survey nonresponse – and interpreting the causal mechanism behind any correlations - is not straightforward.
- The ONS "Hard to Count" measure developed for the 2011 Census did emerge as a significant predictor of nonresponse on the ESS.
- Household level data purchased from commercial companies may provide some useful insights with, for example, missingness from commercial databases a strong predictor of ineligibility and non-contact. However, overall the addition of commercial geodemographic segmentation variables or measures of specific household characteristics does not add significantly to our ability to predict survey nonresponse. There are also concerns about the fact that many commercial variables are produced as the result of "black box" modelling rather than being directly based on real data.
- Ordnance Survey Points of Interest (POI) data provides a potentially rich source of contextual data for both substantive and methodological analysis. However, it has proved difficult to identify POI measures which are correlated with response propensity and ESS survey variables.

Panel Discussion

Presentations on key project findings were followed by a discussion featuring contributions from leading survey methodologists and practitioners.

- Patrick Sturgis, University of Southampton and NCRM
- Tom W. Smith, NORC at University of Chicago
- Patten Smith, Ipsos MORI
- Michael James, ONS

Panellists were asked to consider the following questions:

- What are the key findings emerging on using auxiliary data to study nonresponse bias, from ADDResponse and/or other research projects you are aware of?
- What are the main challenges associated with using auxiliary data to explore survey nonresponse?
- What are the most promising areas for further research combining auxiliary and survey data?
- How will/should the growing availability of auxiliary data influence survey practice in the future?

Patten Smith started the discussion by noting that one of the main findings from ADDResponse was that purchasing household or individual level data from commercial companies such as Experian appears not to contribute much to our ability to predict or understand survey nonresponse. Whilst on the one hand this is disappointing it is good to have access to systematic analysis confirming this so that resources are not wasted investing in these data unnecessarily. He also pointed out that whilst the auxiliary sources used in ADDResponse may not have proved useful for nonresponse analysis, auxiliary data can be useful for substantive analysis, for example to identify food deserts or travel patterns. He argued that it remains important to be vigilant and keep our eyes open for new sources of auxiliary data to be used alongside survey data.

Patrick Sturgis said that it was useful for analysis of nonresponse bias to consider external sources of auxiliary data rather than relying solely on interviewer observations which, for a long time, has been one of the main sources of information on survey nonrespondents. There are concerns about the meaning and accuracy of interviewer observations with for example, evidence that interviewer perceptions of how safe a neighbourhood is bears little resemblance to residents' perceptions or to crime rates. He said he was quite reassured by the findings from the project and that, despite the wide range of auxiliary data employed it had not been possible to identify systematic sources of nonresponse bias. This may point to the fact that most nonresponse is situational rather than systematic. However, it is also worth noting that many of the auxiliary variables used are noisy and reliant on aggregate area level measures whereas we know that most of the variation in response behaviour occurs at the individual level. He also expressed concern about the "black box" nature of data from commercial companies. One further use for the types of data collected through the ADDResponse project may be to compare how sample composition varies over the course of fieldwork and whether, for example, extra contact attempts lead to the achieved sample becoming more representative of the general population.

Michael James talked about work underway at ONS to prepare for the 2021 Census and the possibilities of using auxiliary data to make data collection more efficient. In 2011 £6M was spent on chasing responses from empty properties. If these could be identified in advance fieldwork costs could be reduced significantly. Similarly, information on broadband coverage could be useful in informing online data collection strategy. The use of auxiliary data to streamline fieldwork will be trialled in the 2017 census test conducted in 7 -8 local authorities. He mentioned ONS' "Beyond 2011" programme of work which explored the options for using other sources of data as a possible alternative to the census in providing small area population statistics. Obtaining and linking data can be time consuming and problematic even within ONS or government departments. A project is underway to append a Unique Property Reference Number for each UK address as a standard identifier on all government data sources to facilitate easier linkage. Government Digital Services are working on an address matching tool. He confirmed that a similar indicator to the 2011 "Hard to Count" measure would be constructed for 2011.

Tom Smith argued for the need to continue to explore auxiliary data sources in more detail, to build up a picture of which data sources and variables within data sources are of good/bad quality (income variables in commercial data are particularly likely to be inaccurate whereas other variables may be more reliable for example). He also argued for using auxiliary data in creative ways – for example, constructing indicators of missingness or agreement across multiple data sources. Households or individuals for whom data is missing in multiple data sources – or where there is disagreement between the different sources – may be among the transient or hard to reach population and so less likely to respond to surveys. He pointed out that one of the strongest predictors of survey nonresponse, especially in the USA, remains whether you live in an urban or rural area and that there are systematic area-based differences in response propensity. The reasons for this should be further explored i.e. whether there is something inherent in living in an urban area (what?) that depresses response or whether there is something further to explain.