City Research Online

City, University of London Institutional Repository

Citation: Butt, S. and Lahtinen, K. Using auxiliary data to model nonresponse bias The challenge of knowing too much about nonrespondents rather than too little?. Paper presented at the International Workshop on Household Nonresponse 2015, 02 Sep 2015 - 04 Sep 2015, Leuven, Belgium.

This is the accepted version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/14510/

Link to published version:

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

Using auxiliary data to model nonresponse bias
The challenge of knowing too much about nonrespondents rather than too little?

Kaisa Lahtinen (kaisa.lahtinen@city.ac.uk)
Sarah Butt (sarah.butt.1@city.ac.uk)
City University London

Research Questions to be addressed:

- Are multilevel multisource auxiliary data useful for understanding and predicting patterns of nonresponse in UK social surveys?
- What is the best method for selecting auxiliary variables for nonresponse adjustment?

Introduction

There is growing potential to harness information from multiple sources of auxiliary data and employ what has been termed a multilevel integrated database (MIDA) approach to study nonresponse (Massey and Tourangeau, 2013; Smith and Kim, 2013). This approach, which involves attaching as much auxiliary information as possible to the entire sample, may offer a valuable opportunity to overcome one of the main challenges of nonresponse adjustment – that of identifying suitable variables which are correlated both with response propensity and survey outcome variables. However, it also introduces new challenges. First, there are various methodological, practical and legal issues associated with sourcing and matching a wide range of data sources (Smith and Kim, 2013). Second, once the data are matched, researchers must identify which of the large number of possible variables available may be of most use in nonresponse adjustment.

The first step in nonresponse adjustment using sample-matched auxiliary data is usually to model response propensity using some variant of logistic regression. Assuming that the researcher is lucky enough to have access to more than a few basic auxiliary variables, they must make decisions about which variables to include in the model, usually a process of "multiple steps" based on "heuristics and experience" (Bethlehem et al., 2011). The problem of variable selection is magnified under the MIDA approach, especially once we factor in possible interactions between variables and, often, limited sample sizes. One commonly used approach to variable selection is what might be termed a researcher-led or theory-drive approach where auxiliary variables are pre-selected based on existing literature. Whilst a useful starting point, this should not necessarily provide the only basis on which variables are selected; part of the value of the MIDA approach is the potential it offers to explore possible new correlates of nonresponse in addition to the 'usual suspects'. An alternative approach, motivated by the goal of prediction, is to adopt a machine-learning approach and rely on statistical criteria for variable selection, perhaps using some variant of step-wise regression, or increasingly, regression trees (Toth and Phipps, 2014). The risk of this approach, however, is that it results in models using combinations of variables which are substantively meaningless and/or not easily replicable in future studies. With a large pool of variables and a (relatively) small sample size it may be difficult to arrive at a stable solution and some preliminary variable selection is still likely to be required.

The ADDResponse project (www.addresponse.org) explores the potential for using auxiliary data from multiple sources to understand and correct for nonresponse bias in general social surveys in the UK. Data from the census and other administrative sources together with consumer profiling data and

geographic information about local neighbourhoods have been matched to data from Round 6 of the European Social Survey in the UK.[1] Preliminary bivariate analysis suggests that a large number of these variables may be associated with response propensity and worthy of further investigation. Here we discuss some of the preliminary steps we have taken to try and identify the most likely candidates for nonresponse adjustment and compare the results from propensity models employing theory-driven vs. automated variable selection. We would welcome further suggestions or comments on how to approach the task of modelling response propensity when faced with the (enviable) problem of having too many auxiliary variables from which to choose rather than too few.

Data sources

The choice of auxiliary data sources for ADDResponse has been informed by previous research into nonresponse and theoretical considerations but is also intended to be exploratory and make use of additional sources of auxiliary data not previously used in the analysis of nonresponse. We are also interested to explore different auxiliary measures of the same phenomena and to consider measures at different levels of aggregation to test whether some measures perform better than others and/or could be used as alternatives in future applications.

In common with many other studies of nonresponse, an important source of auxiliary data for our purposes is the national census. We match variables from the 2011 UK Census intended both as proxies for household/individual characteristics (e.g. age/household composition) and as measures of neighbourhood characteristics (e.g. ethnic fractionalisation). We also include some variables which measure change in neighbourhood composition (for example in terms of class composition) over the period 2001-2011 to test whether change may be a driver of response behaviour.

In addition to census variables and measures of area deprivation (relatively common sources of auxiliary data for UK nonresponse analysis), we also appended small-area data from a range of other less well-explored sources of administrative data. This includes recorded crime figures, data on electricity consumption and fuel poverty from the Department of Energy and Climate Change, data on benefit claimants, local area estimates of personal wellbeing, and local election results. Such data are increasingly available in the public domain at relatively low levels of aggregation (lower super output areas for example cover around 1000 households).

We also consider local geographic information i.e. information on the location of sampled addresses in relation to environmental features such as roads, green space, shops or leisure amenities. As far as we are aware this type of information has not previously been used in nonresponse analysis. It is of interest to the extent that their local environment may provide useful indicators of the types of people included in the sample and their day to day experiences. “Points of Interest” (POI) data were obtained from the Ordnance Survey, Britain’s official mapping service, and OpenStreetMap, and mapped onto the sample file using the British National Grid coordinates. A variety of POIs are considered including: the presence of pawn brokers, discount stores and gambling establishments (possible indicators of socio-economic deprivation); the presence of an “evening economy” (which may be associated with

[1] Household level consumer profiling data are not considered as part of the analysis presented here which focuses on the sources of aggregate data.

crime levels and social disorder); and access to transport links, green space and cultural amenities (possible lifestyle indicators). Different ways of representing POIs are considered including counts, distances and relative propensities of one type of POI over others.

The ESS in the UK uses an address-based sample with respondent selection on the doorstep. Auxiliary variables were matched to the list of 4520 addresses sampled to take part in ESS Round 6 (see www.europeansocialsurvey.org for further information) using postcode. 2286 (50.6%) addresses generated an interview, 265 were ineligible, 306 classified as non-contacts, 1268 as refusal and 244 as other nonresponse. In addition to the auxiliary data from external sources, ESS paradata are also available including a small number of interviewer observations for each sampled address.

These data sources generated 284 separate auxiliary variables which could potentially be used to examine patterns of nonresponse in the ESS in the UK. This represents a substantial repository of data to explore. For nonresponse modelling some preliminary variable selection is required.

Preliminary variable selection

The first step was to identify where auxiliary variables were essentially duplicates or alternatives for one another and, where this was the case, to decide which version to use. This was done via a combination of correlation analysis and visualisation to explore the distribution of auxiliary variables. Where there was a choice of alternative measures which were found to be highly correlated with one another, we opted for the variable with greatest coverage, which seemed most theoretically robust, and was more strongly correlated with response in bivariate analysis. We prioritised measures of violent crime over burglary for example because the latter are not available for Scotland and benefit focused measures of unemployment over those in the Census as the former are updated annually rather than every 10 years.

Two noteworthy findings: First, the level of aggregation at which variables are measured appears to make little difference with the same measures at different levels of aggregation (OA vs. LSOA vs. MSOA level) highly correlated with one another. We opted to prioritise OA level measures of household characteristics and LSOA level measures of neighbourhood characteristics. Second, most POI measures (counts of POIs or distance to nearest) were almost perfectly correlated with urban-rural and population density measures and so are unlikely to provide additional information. However, ratio measures indicating the prevalence of one type of POI over another (e.g. fast-food outlets as a proportion of all restaurants) were less correlated with other variables and may perhaps provide some additional insights into the type of area.

Whilst this initial selection process significantly reduces the number of variables in contention (to 129) further data reduction is probably still required.

At this stage we were keen not to restrict the choice of variables based on those which were or were not significantly associated with response propensity in bivariate analysis. This is to account for the possibility that even variables not found to be significant in global bivariate analysis may prove to be important when combined with other variables. We also, at this stage, do not make any variable selections based on the correlations between auxiliary variables and survey outcomes, in part because, the ESS is a general social survey covering a wide range of topics, and so there may be value in considering a more universal approach to response adjustment.

Two data reduction techniques were explored, principal component analysis (PCA) and clustering using partitioning around medoids (PAM) (Kaufman and Rousseeuw, 1990) to see whether certain sub-sets of variables could be identified. PCA did not prove informative. Various versions of the PCA were run including different combinations and numbers of variables but failed to identify meaningful factors with sizeable loadings.

Clustering was more insightful. Sixteen clusters of variables were identified (based on maximising the lowest level of correlation between a pair of variables within a cluster whilst minimising the number of clusters) as shown in Table 1 in the Appendix. Some of these clusters are quite tightly focused, for example cluster 4 which captures old age, In these cases, it may be sufficient simply to include the medoid of the cluster i.e. the variable most correlated with other variables in the cluster and most representative of that cluster (shown in bold in Table 1). The interpretation of other clusters is less clear (for example cluster 12) and we may not necessarily wish to restrict attention to a single variable from within each of these clusters. Additionally it might be useful to consider whether we should include more than one variable from some of the clusters if specific variables are considered to be theoretically important for nonresponse analysis.

Moving to a multivariate model of response propensity

Having made a preliminary selection of variables, we move on to consider how to employ the remaining variables in a multivariate model of response propensity. We present here findings from two different models of response propensity – one in which variables were selected by the researchers on theoretical grounds and a second more statistics-driven approach employing automatic variable selection. In both cases response is treated as a binary outcome (responded vs not) and the base is all eligible addresses in Great Britain (N=4290). At this stage, only the results of a global model with main effects and no interactions is considered.

The theory-based analysis consisted of a logistic regression model containing 16 variables. Variables were selected to be representative of the 'usual suspects' found to be important in previous studies of nonresponse and informed by the clustering analysis. Automatic variable selection was done using LASSO (Least Absolute Shrinkage and Selection Operator) regression in preference to the more common step-wise regression. Step-wise regression has some statistical issues (often producing biased significance tests and inflated coefficients as well as omitting predictors with insignificant individual contributions that may hide a significant joint contribution) which Lasso is able to overcome. Lasso regression is a shrinkage and selection method that involves penalising the absolute size of the regression coefficients and including or excluding variables from the model on this basis. Lasso was fitted with the subset of 129 variables that resulted from preliminary variable selection.

A summary of the variables included in each model, and the direction and magnitude of their relationship with response, is provided in Table 2 in the appendix. Key findings include:

- Both models identify some auxiliary variables that are correlated with response propensity. However, neither model does a particularly good job of predicting the outcome. What, if anything, can be done to improve model fit?
- It is interesting to note that the LASSO model identifies 10 output area or household level proxy measures from the census as significant compared with two LSOA or neighbourhood level measures. This may provide a useful clue as to the level of aggregation to prioritise in further variable selection?
- Lasso identified more and, importantly, different auxiliary variables compared with those included in the theory-based model. It included variables from nine different auxiliary data

sources, including some of the more 'novel' measures such as measures of subjective wellbeing and the evening economy POI measure. It is possible to identify reasons why these variables might be important in predicting response behaviour and hence justify their inclusion. This suggests that there may be value in exploring multiple sources of auxiliary data. However, there are also practical costs involved in using data from multiple sources. Might we want to prioritise more common or readily available measures in our model?

- In some cases variables selected by LASSO are alternative measures of concepts included in the theory-based model. For example, might LASSO or a similar approach be useful in selecting between alternative measures of those higher level constructs which theory or previous experience suggests may be important to consider?

Summary

It is now possible to match a wide variety of auxiliary variables to address-based sample files and use these data to investigate nonresponse bias in social surveys. Preliminary analysis of their relationship with response propensity in the context of the UK ESS suggests that some of these variables may be useful in nonresponse adjustment and, crucially, that there may be value in moving beyond the 'usual suspects' to consider alternative data sources. However, despite the wide range of (nearly 300) variables available and the use of different modelling techniques, a robust, strongly predictive model of response propensity remains elusive. Neither theory-based nor machine learning approaches on their own appear sufficient. There may perhaps be some value in adopting an iterative approach to variable selection informed by both approaches e.g. using machine learning to refine theory driven choices and then checking selections for theoretical relevance.

Questions for discussion:

- Neither of the models presented here do a particularly good job of predicting response propensity. What, if anything, should we consider to improve model fit?
- What should we prioritise when it comes to variable selection and model building? For example, the best fitting LASSO model suggests data from nine different data sources be used to predict nonresponse. Is this practical in most instances?
- How can we make best use of machine learning/automatic variable selection techniques in nonresponse analysis?
- Are there other approaches to variable selection, besides those discussed here, that we should consider?

References

Bethlehem, J., Cobben, F., Schouten, B., 2011. Handbook of Nonresponse in Household Surveys. John Wiley & Sons.

Kaufman, L., Rousseeuw, P.J., 1990. Partitioning Around Medoids (Program PAM), in: Finding Groups in Data. John Wiley & Sons, Inc., pp. 68–125.

Massey, D.S., Tourangeau, R., 2013. Where Do We Go from Here? Nonresponse and Social Measurement. Ann. Am. Acad. Pol. Soc. Sci. 645, 222–236.

Smith, T.W., Kim, J., 2013. An Assessment of the Multi-level Integrated Database Approach. Ann. Am. Acad. Pol. Soc. Sci. 645, 185–221.

Toth, D., Phipps, P., 2014. Regression Tree Models for Analyzing Survey Response. Presented at the Joint Statistical Meetings (JSM), Boston, Massachusetts.

The project "**A**uxiliary **D**ata **D**riven Non**Response** Bias Analysis" (ADDResponse) is a collaboration between the Centre for Comparative Social Surveys, City University London; the giCentre, City University London; and the Department of Statistics, LSE. It is funded by the Economic and Social Research Council grant number: ES/L013118/

Appendix

Table 1: Partitioning Around Medoids results

Auxiliary variable	Number of the cluster	Cluster name
% owner occupation, % social renting**, %** detached housing**,** % divorced**, %** under crowding, % 2 or more cars, 1 or more cars**,** % owner occupation lsoa	1	OA level affluence
% single, % private renting, % living alone 35 year old and under, % full-time student hholds, % sharing multi-adult hholds, % married, % 16 to 24 year olds, % 45 to 64 year olds, % proving unpaid care, % working part-time, % full-time students, % commuting by foot or by bike, change % NSSEC routine lsoa (01-11), % voted Green party	2	Young (urban) areas
% flats_lsoa, % flats, % families with non-dependent children , % commuting by car, % overcrowding, % private renting, population density lsoa, change in population density lsoa (01-11), extraversion, openness, % no access to gas network, internal population flow, prevalence of evening economy, access to culture amenities, access sport amenities, distance to transport links	3	Urban areas
% 65 years old and older, % living alone 65 years old and over, % 25 to 44 year olds, % 85 years old and up, % with limiting disability, % retired	4	Old people
% hholds with dependent children, % living alone, % lone parents with dependent children, % 0 to 4 year olds, % 5 to 15 year olds	5	Families
% NS SEC managerial, % with bad health, % with no qualifications, % with level 4 qualifications (or higher), % full-time work, % self-employed, % working 49h+ a week, % NS SEC routine, NS SEC managerial lsoa, % NS SEC routine lsoa, change % NS SEC managerial lsoa (01-11), % voting turnout	6	Employment/ social class
IMD, % unemployed, % long term unemployed, % social renting lsoa, % detached houses lsoa, % unemployed lsoa, long term unemployed lsoa, IMD income, IMD employment, IMD health, IMD education, IMD housing, IMD crime, electricity msoa	7	Deprivation
% UK born lsoa, % white, % mixed, % Asian- Indian, % black, % UK born, % Christian, Fractionalisation index lsoa, % EU born lsoa, % Poland born lsoa, % Christian lsoa, IMD access, change in Fractionalisation index lsoa (01-11), change in % UK born lsoa, agreeableness, international population flow, prevalence of fast-food outlets, prevalence of private schools, distance to police stations	8	Urban multicultural areas
% Muslims, % Asian- Pakistani, % Asian- Bangladeshi, % Asian all % Muslims lsoa	9	Asian ethnicity
% no religion lsoa, % no religion	10	No religion
Crime: violent LA, conscientiousness, crime: sexual LA	11	Sexual and violent crime
% out of work benefit lad, neuroticism**, %** job seekers allowance lad**, %** educational absence lad**, %** voting conservative**, %** voting Labour**,** crime: damage LA**,** golf**,** water	12	
% voting for non- major party, % voting Liberal Democrats**, %** voting national parties**, %** voting independent candidates	13	Voting for non-major party
Life satisfaction, happiness**,** feeling worthwhile**,** anxiety	14	Subjective well-being
Prevalence of pawnbrokers Prevalence of gambling Prevalence of special food stores Prevalence of discount stores	15	Not nice PoIs
Prevalence of frozen food stores	16	Frozen foods

Variables in bold are the cluster medoids

Table 2: Regression results

Logistic regression (1=respondent 0=nonrespondent)	Lasso
Type of housing (reference category: detached)	% social renting (+)
Flat (-)	% detached housing (+)
Other (-)	% flats (-)
Semi-detached house (-)	% living alone 35 year old and younger (-)
Terraced house (-)	% sharing multi-adult hhold (+)
Access: not locked (+)	% 5 to 15 year olds (+)
% 65 year olds and older (+)	% providing unpaid care (-)
% owner occupation (-)	% full-time work (-)
% hholds with dependent children (+)	% part-time work (+)
% flats (-)	% Asian Indian (+)
% 16 to 24 year olds (+)	% social renting lsoa (+)
% level 4 qualifications (or higher) (+)	% EU Born lsoa (-)
% NSSEC routine (+)	IMD access (-)
Fractionalisation index lsoa (-)	Openness (+)
Population density lsoa (+)	jobseekers allowance lad (+)
Crime: violence LA (-)	% voting Liberal Democrats (+)
Out of work benefit lad (-)	% voting national parties (+)
IMD (-)	% voting non- major parties (+)
Prevalence of pawnbrokers B (-)	Happy (-)
Life satisfaction (+)	Anxiety (-)
	crime: sexual LA (-)
	prevalence of evening economy (-)
	prevalence of private schools (-)
	distance to golf courses (+)
	distance to water (+)

Sign in brackets indicates the direction of the relationship. Faded variables included in logistic regression but not statistically significant at the 5% level.