**Taking into account unknown eligibility in nonresponse correction**

Paper to be presented at the 23rd International Workshop on Household Survey Nonresponse, Ottawa, September 2012

Olena Kaminska and Peter Lynn, ISER, University of Essex, UK (contact olena@essex.ac.uk)

**Abstract**

This short note describes an innovative method of calculating response rates and estimating nonresponse correction weights for the cases of unknown eligibility. While the methods used so far are deterministic in nature and aim to assign a definite status of ‘eligible’ or ‘not eligible’ to the cases of unknown eligibility, our method is probabilistic and works by assigning probabilities to be eligible to all the cases of unknown eligibility. The eligibility probabilities are then used as weights for estimating response rates, and for estimating nonresponse propensity models. This method appears to have advantages over the methods used so far, including the possibility to account for uncertainty in the eligibility status among the unknown cases.

**Introduction**

Unknown eligibility is a recognised issue in the nonresponse field (AAPOR standard definitions, 2011; Smith, 2009). It occurs in the situation where not all units on the sampling frame are eligible for a study, and the eligibility status cannot be known before fieldwork. This means that one task of the fieldwork is to define eligibility, something that may not be known for all issued cases at the end of the fieldwork.

Unknown eligibility types vary depending on the mode of the survey, sampling frame and the population of interest. These may include ‘ring, no answer’, or ‘always busy’ in telephone survey, ‘not returned’ in mail or email survey, and ‘not contacted’ in face-to-face survey. Another special category is when the population of interest is much smaller than the population represented in the sampling frame and a screening step is required before eligibility is established. For example, the National Immunisation Study (NIS) in the US looks for households with children between 19 and 35 months using an RDD sample. In this study, all phone numbers not reached and those contacted but where no information on the presence and age of children is provided are of unknown eligibility.

A number of methods for calculating response rate in the presence of unknown eligibility have been suggested (Smith 2009). Minimum and maximum allocation methods treat all cases of unknown eligibility as noneligible or eligible respectively. This corresponds to AAPOR response rates RR1 and RR2 for a maximum allocation; and RR5 and RR6 for minimum allocation (see AAPOR standard definitions, 2011). A slightly more realistic approach is to assume that the proportion of eligible cases (e) is the same among cases of unknown eligibility as among cases of known eligibility. This corresponds to AAPOR’s RR3 and RR4.

Other methods use additional information to determine whether the case is eligible or not (see Smith, 2009, for a review). This is done by using final outcome codes (either by assigning eligibility status directly to each code or by estimating it using other additional information), using survival analysis to predict eligibility based on number of attempts, or using record or geographical linkage to help determine the eligibility status. Other methods learn more about the eligibility of the case, either through extensive follow-up procedures, or through obtaining information about the phone number or address from external company.

All these methods aim to assign the eligibility status (0 or 1) to each case of unknown eligibility. Even the methods that estimate eligibility probability use this probability only to assign the predicted status to each case.

In this paper we suggest and demonstrate empirically how the predicted probability of being eligible can be used directly in estimating response rates and in developing nonresponse correction weights.

**Data and Method**

We demonstrate our method using wave 1 data from the UK Household Longitudinal Study (UKHLS). One aim of this study was to collect detailed information on ethnic minorities. For this, a separate sample was selected in areas with high proportions of ethnic minorities and only households with at least one ethnic minority member were eligible for an interview (for details on sampling see Berthoud, Fugamalli, Lynn and Platt, 2009). The households in the selected areas were screened with questions on the ethnicity of household members. Thus, only households who were contacted and provided information on ethnic minority composition of the household are of known eligibility. All households which were not contacted or contacted but refused information on ethnicity are of unknown eligibility. Note, selection probabilities differed across ethnic minority groups and months of issue (see Berthoud, Fugamalli, Lynn and Platt, 2009, for details).

Ideally we would like to know the eligibility status for every sample household. In this situation only eligible cases would be used in response rate calculation and modelling nonresponse. While eligibility information is not available directly, we have other auxiliary information for all issued cases. The most relevant information to eligibility include month of issue (screening probabilities changed at month 12 and month 21 for some ethnic subgroups out of 24 months fieldwork period ); and predicted ethnic minority population density of Indian, Pakistani, Bangladeshi, Caribbean and African in the postal sector (from sampling frame, used to select clusters with different probabilities). Given that screening probabilities (and therefore eligibility) varied by ethnic group and month of issue by design, these variables are crucial for our model. Other information used includes region (GOR) and several neighbourhood-level variables from the Census and other administrative sources linked to our dataset.

The process involved the following steps. We first excluded categories of ineligible which should always be identified (and would not have a chance to be of unknown eligibility): vacant house, not yet built, non-residential, institution, etc. We then divide the remaining sample into three subgroups: those *known to be eligible* (including interviewed households and those who refused an interview but were successfully screened in); those *known to be ineligible* (these are households which were contacted and went through successful screening, and were not selected into the sample either because they didn’t have any member of ethnic minority group of interest or were screened out by design according to selection probability – e.g. because the areas had high number of Indians, the households with them were selected with the probability of 0.5 in the first 12 months of the fieldwork and 0.62 thereon); and finally those of *unknown eligibility* (these include household noncontacts, as well as those who refused information on ethnic minority composition of the household).

Our aim is to estimate probabilities of eligibility for each household of unknown eligibility. For this, using only households with known eligibility, we run a backward stepwise logistic regression predicting eligibility. The auxiliary variables described in the previous section include month of issue, ethnic minority population density, region, and a number of neighbourhood characteristics from administrative data sources. Only significant predictors at the 0.01 p-value level remain in the final model. Using the final model (note, predictors are available for all issued households) we predict eligibility for all households of unknown eligibility.

The eligibility likelihood is assigned to each household in the following way: households known to be eligible receive the value of 1; households known to be ineligible receive a value of 0; and households of unknown eligibility receive model- predicted probability of eligibility. This eligibility probability is now used as a weight component, and we suggest this can be released as part of a design weight. Note, this component does not change the design weight for responding households (as it is 1 for all of them), but provides a value for nonresponding households of unknown eligibility. This is useful for future analysis on nonresponse.

Note, unlike the usual weight calculation which is equal to the inverse of the probability (of selection or response), the weight component of eligibility is directly equal to the probability of being eligible. This is because we need households with higher chance of being eligible to contribute more to the nonresponse model – so the relationship is direct.

**Response rate calculation**

To estimate response rate, we simply run weighted frequencies of an outcome code, where the weight is the probability of being eligible (eligibility weight). Note, the weight should be the product of design weight and eligibility probability, but considering the complexity of the design weight in UKHLS and its lack of relevance to this paper, we use only eligibility probability here.

Table 1 in the Appendix shows unweighted and weighted outcome frequencies in the ethnic minority boost part of wave 1 UKHLS. Out of 40,748 issued households (the pre-screening ineligible households are excluded), we had 4,066 households who completed an interview, 30,590 households screening out (ineligible), 2,472 nonresponding households who are known to be eligible, and 3,620 nonresponding households of unknown eligibility. The weighted column represents estimated number of eligible households in each outcome category and the last column represents an average eligibility probability in that column. Interestingly, among households of unknown eligibility, households unknown whether residential are estimated to have rather low eligibility probability (0.22 or 0.24), non-contacted households also have low eligibility probability on average (between 0.22 and 0.26) compared to contacted ones (generally between 0.29 and 0.62), and households with language difficulties have relatively higher eligibility probability (0.38 or 0.56).

Following AAPOR standard definitions (2011), the RR2 under the assumption that all households of unknown eligibility are eligible would be equal to 4066/10158=40.0%; the RR6 assuming that all households of unknown eligibility are ineligible would be equal to 4066/6538=62.2%; RR4 with the proportional allocation would be equal to 4066/7175.46=56.7%; and our response rate taking into account predicted probability of being eligible is 4066/7843.97=51.8%. Interestingly, the proportional allocation response rate of 56.7% is higher than our response rate of 51.8%. This is because the eligibility ratio among those known to be eligible 6538/37128=0.176 is lower on average than 0.36 (std=0.30) as estimated from eligibility model. This is likely a result of ethnic minorities having a lower probability of being contacted or responding to the screening questionnaire, and therefore there being more eligible households among noncontacted and not screened households than amongst those who were successfully screened.

**Incorporating eligibility probability into nonresponse correction**

In running a nonresponse model ideally we would like to exclude all ineligible households, and predict response only among eligible ones. In the presence of unknown eligibility we suggest that nonrespondents of unknown eligibility should contribute to the model proportionally to their likelihood of being eligible. We therefore use a weight equal to the predicted probability of being eligible in a weighted backward stepwise logistic regression which predicts response to the survey.

An alternative method which would account for unknown eligibility would be a two stage modelling where the first model predicts successful screening among all issued households, and the second model predicts response among households known to be eligible. The nonresponse correction is then inversely proportional to the product of two probabilities for these models. Note, such an approach assumes that the ratio of eligible households among those of unknown eligibibility is the same as among those of known eligibility (successfully screened). We expect that in our example this method will result in a different nonresponse adjustment considering that our method predicts a very different eligible proportion amongst cases of unknown eligibility (0.36) than the proportion observed amongst cases where eligibility was established (0.176).

In table 3 we compare two methods of adjusting for unknown eligibility. The first column presents model results predicting the chance of successful screening; the second is for the model predicting response among those known to be eligible, and the third is the product of these two odds ratios, the inverse of which can be used as a nonresponse correction weight. The fourth column presents odds ratios for our weighted model. For comparison we used selected range of predictors and we included all of them regardless of p-value. Note, the final odds ratios in models differ and therefore two methods would result in different nonresponse correction weights. The variation in weights appears to be greater with the two-stage method.

**Points for discussion**

We suggest a way of improving response rate calculation and nonresponse correction in the situation when a part of issued units are of unknown eligibility. This is achieved through predicting the probability of eligibility and then using this prediction directly instead of trying to impute a binary indicator of eligibility for each case.

While this method seems to have advantages and potential, this method may be controversial with respect to response rates considering that the value would depend on the eligibility model.

Advantages of our method over alternative methods for nonresponse correction are still unclear and we are looking for suggestions of alternatives as well as criteria of comparison among them.

We are also eager to learn of any possible refinements to our method as well as other situations where it may be useful.

Table 1

Final outcome codes of household response in UKHLS ethnic boost, wave 1, unweighted and weighted by eligibility probability

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **outcome code** | **eligibility status** | **Un-weighted** | **weighted by eligibility probability** | **average eligibility probability** |
| completed household interview | eligible | 4066 | 4066 | 1 |
| refusal before interview by selected respondent | eligible | 1,997 | 1997 | 1 |
| broken appointment no recontact | eligible | 384 | 384 | 1 |
| proxy refusal | eligible | 55 | 55 | 1 |
| refusal during interview (unproductive partial) | eligible | 23 | 23 | 1 |
| lost on laptop | eligible | 10 | 10 | 1 |
| interview achieved but resp requested data deletion | eligible | 3 | 3 | 1 |
| office approval only: issued but not attempted | unknown eligibility | 1 | 1 | 1 |
| office approval only: other unproductive | unknown eligibility | 492 | 325.6 | 0.66 |
| contact made: information refused on number of households | unknown eligibility | 39 | 24.19 | 0.62 |
| language difficulties with hhold as a whole | unknown eligibility | 230 | 128.4 | 0.56 |
| contact made; information refused about household | unknown eligibility | 349 | 174.1 | 0.50 |
| contact made; information refused about address | unknown eligibility | 128 | 54.68 | 0.43 |
| other unknown eligibility | unknown eligibility | 27 | 11.14 | 0.41 |
| contact made at dwelling but not with selected household | unknown eligibility | 5 | 2.01 | 0.40 |
| uncertain eligibility: language barrier | unknown eligibility | 18 | 6.839 | 0.38 |
| screener not completed: lack of knowledge | unknown eligibility | 24 | 8.043 | 0.34 |
| contact made at hhold but not with resp adult | unknown eligibility | 53 | 17.13 | 0.32 |
| information refused whether residential | unknown eligibility | 17 | 5.191 | 0.31 |
| contact made at address, not with anyone at selected dwelling | unknown eligibility | 11 | 3.357 | 0.31 |
| contact made at address, not with household | unknown eligibility | 84 | 24.63 | 0.29 |
| refusal to answer screening questions | unknown eligibility | 216 | 62.09 | 0.29 |
| office use only: not issued to interviewer | unknown eligibility | 12 | 1.775 | 0.27 |
| other unknown eligibilty | unknown eligibility | 41 | 10.91 | 0.27 |
| office approval only: inaccessible | unknown eligibility | 50 | 13.24 | 0.26 |
| no contact after 6+ calls | unknown eligibility | 836 | 214.5 | 0.26 |
| uncertain eligibility: lack of knowledge | unknown eligibility | 10 | 2.448 | 0.24 |
| residential unknown eligible: no contact 6+ calls | unknown eligibility | 169 | 41.25 | 0.24 |
| information refused whether resident eligible | unknown eligibility | 115 | 26.37 | 0.23 |
| unknown whether residential: no contact after 6+ calls | unknown eligibility | 121 | 26.59 | 0.22 |
| no contact with anyone at selected household | unknown eligibility | 440 | 95.76 | 0.22 |
| office refusal | unknown eligibility | 120 | 21.5 | 0.18 |
| contact not confirm presence of resid hhd | unknown eligibility | 12 | 3.228 | 0.15 |
| ineligible (ethnic boost) | ineligible | 30,590 | 0 | 0 |

Table 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Alternative 2-stage model | | | our model |
|  | stage 1 (screening prob) | stage 2 (response prob) | Combined odds ratios |  |
|  | Odds Ratio | Odds Ratio |  | Odds Ratio |
| Indian density (mid) | 1.21 | 0.90 | 1.09 | 1.00 |
| Indian density (high) | 0.88 | 0.94 | 0.83 | 1.01 |
| Pakistani (mid) | 0.98 | 0.97 | 0.95 | 0.98 |
| Pakistani (high) | 0.68 | 0.96 | 0.65 | 0.93 |
| Bangladeshi (mid) | 0.59 | 0.90 | 0.53 | 0.86 |
| Bangladeshi (high) | 0.37 | 0.93 | 0.35 | 0.78 |
| Carribean (mid) | 0.97 | 0.87 | 0.85 | 0.96 |
| Carribean (high) | 0.75 | 0.72 | 0.54 | 0.81 |
| African (mid) | 0.69 | 1.29 | 0.89 | 1.11 |
| African (high) | 0.37 | 1.11 | 0.41 | 0.80 |
| GOR North West | 1.49 | 1.46 | 2.18 | 1.20 |
| GOR Yorkshire | 1.24 | 1.69 | 2.10 | 1.28 |
| GOR East Midlands | 0.92 | 1.24 | 1.13 | 0.94 |
| GOR West Midlands | 0.96 | 1.08 | 1.04 | 0.82 |
| GOR East of England | 2.43 | 1.09 | 2.65 | 1.14 |
| GOR Inner London | 1.14 | 1.14 | 1.31 | 0.96 |
| GOR Outer London | 1.14 | 1.10 | 1.26 | 0.92 |
| GOR South East | 1.98 | 1.07 | 2.12 | 1.08 |
| GOR South West | 1.31 | 2.03 | 2.66 | 1.78 |
| Intercept | 20.11 | 1.61 | 32.31 | 1.34 |