Failing to locate panel sample members: minimising the risk

Peter Lynn, ISER, University of Essex

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

Abstract

At each wave of a panel survey each sample member must first be located before attempts can be made to make contact and gain co-operation. Failure to locate sample members can contribute a sizeable proportion of nonresponse. Methods to maximise location rates have, however, been largely neglected in the survey methods literature, with a focus instead on reducing non-contacts and refusals. We propose the use of between-wave interventions targeted at sample members at highest risk of not being located. To implement such interventions, two methodological challenges must be met: a) to identify which sample members are at highest risk, and b) to identify effective interventions. In this paper we present an empirical attempt to meet the first challenge and some theoretical discussion regarding the second challenge.

We analyse data from the first two waves of *Understanding Society: the UK Household Longitudinal Study*. We fit logistic regression models to predict the propensity to not be located at wave 2, based on a range of variables collected at wave 1. We identify the variables that are most useful for prediction of risk of not being located. We then develop a theoretical discussion of the likely effectiveness of prediction-based interventions.

Introduction

At each wave of a panel survey, the process of obtaining the response of a sample member involves three stages: location, contact and co-operation (Lepkowski & Couper 2002). Much of the literature on panel attrition is devoted to description and discussion of the contact and co-operation stages (e.g. Hill & Willis 2001, Uhrig 2008, Watson & Wooden 2009). Little attention has been paid to the location stage (Couper & Ofstedal 2009). However, failure to locate sample members may account for an increasingly large proportion of non-response over waves of a panel survey, as both non-contact rates and refusal rates tend to decrease, conditional on response to the previous wave. Failure-to-locate rates will tend to be higher the longer the interval between survey waves and the greater the levels of mobility in the study population. Furthermore, non-response due to a failure to locate sample members is clearly not random. Non-respondents in this category consist solely of people who have moved home, a subgroup with distinct characteristics. For these reasons, identifying ways to reduce the proportion of panel sample members who are lost to a panel survey due to a failure to locate them is important, particularly for panel surveys carried out amongst relatively mobile populations.

Identifying "at-risk" sample members

The objective is to identify sample members who would have a relatively high probability of becoming non-respondents due to a failure to locate them at a subsequent wave, in order that special interventions designed to reduce that probability could then be implemented for those individuals. We do this as follows. We first develop a general model of the probability of being not located at a particular wave, based on data available from the previous wave. We then use this model to generate predicted values for respondents to the latest wave. These values determine which respondents will receive the special interventions. To identify relative risks, the model should be developed using data in which all sample members were treated equally with regard to any procedures that might affect the propensity to locate the sample member at the next wave.

We use data from the first two waves of *Understanding Society: the UK Household Longitudinal Study*, a large general population survey involving face-to-face interviews with a 12-month interval between waves.

Identical procedures were administered to all sample members, both at wave 1 and between waves 1 and 2. Relevant procedures included:

- asking for extensive contact details at wave 1, including the details of a friend or relative who would be likely to the know the whereabouts of the sample member in the event of a move;

- leaving an "address update card" with each respondent, to be returned by freepost in the event of a move;

- sending a between-wave mailing to all respondents. This mailing included a brief report on findings and another address update card;

- initiating tracing procedures *prior to* wave 2 for any sample member whose between-wave mailing was returned by the Post Office as "undeliverable" or "not known at address". These procedures utilised the contact details provided at wave 1, where available;

- initiating tracing procedures *during* wave 2 for any sample member who was found during wave 2 field work to have moved and for whom the interviewer was not able to obtain a new address.

We fit a logistic regression model to predict the propensity to not be located at wave 2 (2010-11), based on a range of socio-demographic and behavioural variables collected at wave 1 (2009-10). The base consists of all sample members for whom data were successfully collected at wave 1 of the survey (either personal or proxy individual interview), aside from those known to have died or moved abroad prior to wave 2. Analysis is restricted to the "year 1 sample", which consists of the first 12 of 24 monthly sample replicates (the *Understanding Society* sample is divided randomly into 24 monthly samples for fieldwork management purposes, so each wave of data collection takes two years). This was done because final outcome data for the year 2 sample was not yet available at the time of analysis. The year 1 sample contained 24,525 wave 1 respondents, of whom 337 were known to have become ineligible by the time of the wave 2 field work, leaving an analysis sample of 24,188. Of these, 4.2% were not located at wave 2.

A set of 21 potential predictor variables were selected, based on non-response theory and previous studies. These are listed in table 1. The model was developed using forward stepwise procedures. For most variables, alternative forms were tested and the one with the lowest p-value at that step was selected for inclusion.

Results

In all, 17 of the 21 variables made a significant contribution to the model. The two strongest predictors were:

- A 7-category indicator of housing tenure ("tenure");
- A binary indicator of whether the sample member reported at wave 1 that they expected to move home in the next 12 months ("xpmove").

A model with tenure as the sole predictor has a pseudo-R^2 of 0.160, while a model with xpmove as the sole predictor has a pseudo-R^2 of 0.075. A model with these two variables has a pseudo-R^2 of 0.186. The other 15 significant predictors made smaller contributions to the final model, which has a pseudo-R^2 of 0.24. This is a rather powerful model. Such predictive power is rare is non-response research, indicating that a failure to locate someone may be rather more predictable than a non-contact or a refusal. This is promising as it suggests that targeting interventions at the "at-risk" could be quite efficient.

Model results are presented in table 2. The effects of the predictor variables can be summarized as follows. An increased propensity to not be traced is associated with:

Renting accommodation from a private landlord, particularly if the accommodation is furnished;

Reporting an expectation to move within the next 12 months;

Age 20-24 (and, to slightly lesser extent, 25-29);

Male;

Single (never married, divorced, separated, widowed) rather than in a couple;

Not currently in employment;

Born outside the UK;

Did not complete the self-completion questionnaire;

Proxy rather than personal interview;

Not living at the same address for at least 20 years or since birth;

In England and Wales;

No-one under 15 in the household;

Living in a dwelling attached to business premises, a converted flat, or sheltered accommodation

General health rated as excellent;

Being aged 17-18 (16-17 at previous wave) *and* in the August sample;

Not having internet access in the home;

Having no qualifications, or a degree or higher qualification.

The variables that were tested but rejected (no significant contribution to the model) were: number of adults in the household, number of floors in the building, main floor level of accommodation, presence of an entryphone.

Efficiency of interventions

The strong predictive power of the model suggests that intervention could be cost-effective. In designing an intervention, there are two key decisions:

a) To whom should the intervention be administered?

b) What should the intervention consist of?

In this section we address question a). Question b) is addressed in the next section. If we administer an intervention to all sample members with a predicted probability, p, greater than a critical value p^*, then question a) becomes one of identifying the optimal value of p^*. The variable cost of administering the intervention will be approximately proportional to the number of sample members, n^*, with $p > p^*$. If we assume that the effectiveness of the intervention (i.e. the probability, r, that the intervention enables location of a mover who would otherwise not have been located) is independent of p, then the number, m, of additional sample members who will be retained in the sample as a result of the intervention will be proportional to the number within the treatment group who would otherwise have been not located at the next wave. Figure 1 shows the predicted rate of not located at the subsequent wave under this assumption, for $r = 0.4$, for $0.025 < p^* < 0.25$. We do not consider values of $p^* > 0.25$ as this would result in small n^* (in our data, $n^*/n = 0.038$ for $p^* = 0.25$) and hence small impact on the rate of failure to locate. Values of $p^* < 0.025$ can of course be considered, but this no longer constitutes targeting the high-risk group (as p has a mean of 0.044 and median of 0.014). An efficient targeted intervention can be thought of as one that optimizes the trade-off between the likely reduction in the rate of failure to locate (the reduction increases as p^* decreases) and the cost of implementation (cost increases as p^* decreases). The relationship between cost and p^* is displayed in figure 2. It can be seen that costs reduce sharply as p^* increases to around the mean of 0.044. The rate of cost reduction then slows. The increase in the predicted rate of failure to locate is, however, close to linear in the range $0.03 < p^* < 0.09$, so we suggest that an efficient choice of p^* may be around 0.05, provided sufficient budget is available.

Identifying effective interventions

An effective intervention is one that maximizes r at an affordable cost. A key consideration is the need to make contact with a sample member at a time when, a) they already know their new contact details, but b) they are still contactable via their old contact details. Contact attempts, however they are made, are unlikely to be successful in obtaining the new contact details if they are made outside of this window of opportunity.

Figure 1: Predicted proportion not located, by *p

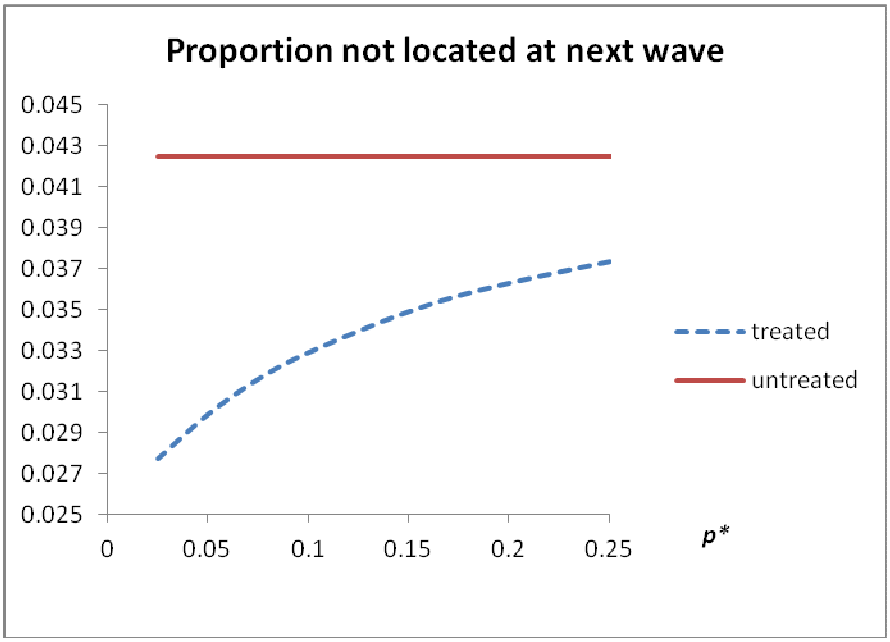
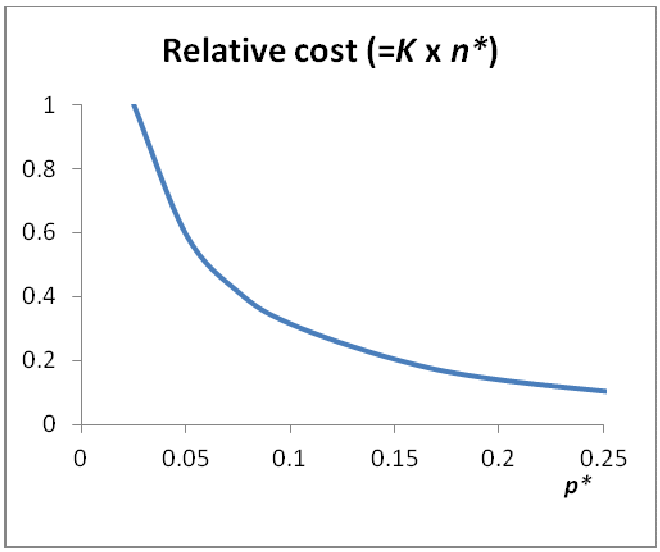


Figure 2: Relative cost of intervention, by *p



We therefore consider that the most important aspect of an intervention is likely to be its timing. For sample members who move between waves, the process of administering an intervention can be decomposed into three parts, for each of which timing could be important:

a) Successfully making contact with the sample member;

b) Conditional on contact, the sample member knowing his/her new contact details;

c) Conditional on knowing the contact details, reporting them to the survey organization (i.e. responding to the intervention).

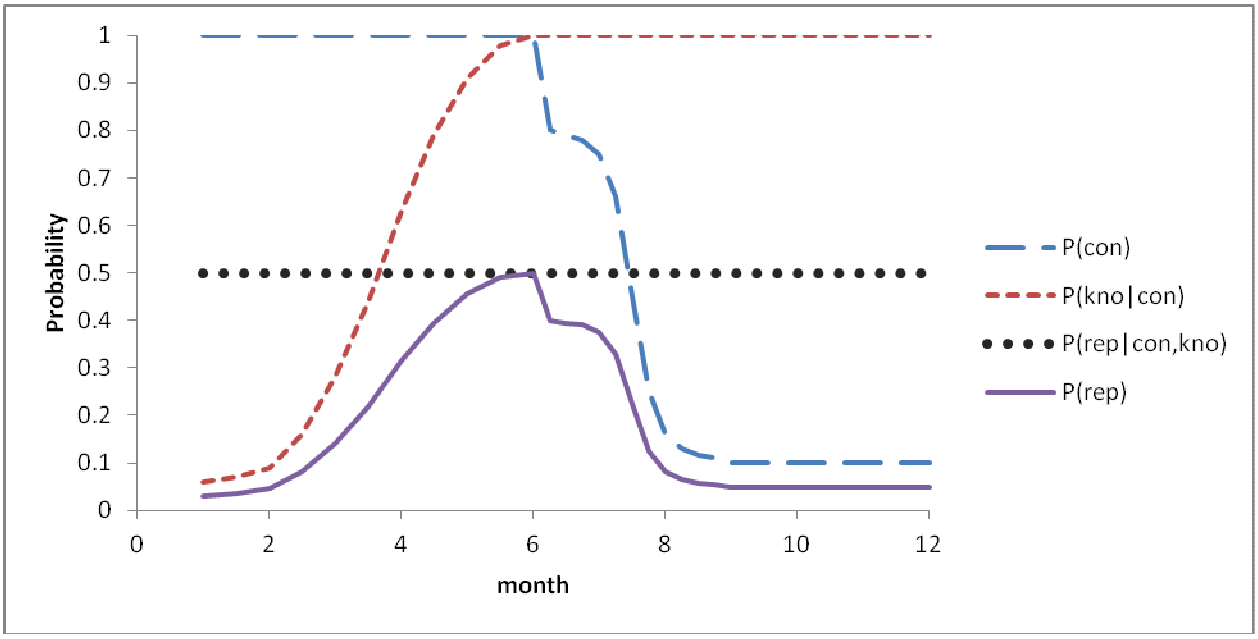
The probability of successfully obtaining updated contact details is the product of the three conditional component probabilities. It may be instructive to consider how each of the three component probabilities may vary over time, relative to the date of an actual move. The probability of successfully making contact

should be close to 1.0 prior to the move, and independent of the month of the attempt. It is then likely to fall sharply immediately after the move. However, some movers will have their mail redirected for a limited period (typically 1, 2 or 3 months in the UK) or will leave their new contact details with the new resident of their previous address. For these people, the probability of making contact will fall off over time, notably as the redirect periods end. This is shown in figure 3 as P(con). The second component, the probability of the sample member knowing his or her new contact details, will rise rapidly in the months prior to the move, reaching 1.0 just before the move. This is shown in figure 3 as P(kno|con). The third component, the conditional probability of the sample member reporting the new address, may be fairly constant over time. This is shown as P(rep|con,kno). The product of the three components is plotted as P(rep). This suggests that there may be a period of around three months – two months before the move and one after – when there is a fairly good chance of an intervention producing the new contact details. This is the window of opportunity referred to at the start of this section. Any intervention that takes place outside of that period has a greatly reduced chance of success.

The challenge, then, is to minimize the proportion of moves by sample members that take place outside of any window of opportunity to report the move. This can be done by timing interventions so that they cover the majority of the periods between waves for high risk individuals.

Having decided the timing of the intervention, remaining decisions concern the nature and content of the intervention. Options concern the mode of approach, the wording of the request, the design of materials if a mail approach is used (on this point, see McGonagle et al 2011 and Lynn et al 2012), and incentivisation.

Figure 3: Probability of obtaining new address, by timing of intervention, for a move in month 6



Points for discussion

How best to predict the propensity to not be located at next wave;

How best to use predicted propensities to identify the target group for the intervention;

How to determine the content/nature of the intervention (mail, telephone, in-person, incentive, …);

Should the content/nature of the intervention be different for different sample subgroups?;

How to determine the timing of the intervention (including whether there should be more than one intervention between waves).

References

Couper M P and Ofstedal M B (2009) Keeping in contact with mobile sample members, 183-203 in P Lynn (ed.), *Methodology of Longitudinal Surveys*, Chichester: Wiley.

Fumagalli L, Laurie H and Lynn P (2012) Experiments with methods to reduce attrition in longitudinal surveys. *Journal of the Royal Statistical Society Series A (Statistics in Society)*, published online 27 July 2012.

Hill D H and Willis R J (2001) Reducing panel attrition: a search for effective policy instruments, *Journal of Human Resources* **36**, 416-438.

Lepkowski J M and Couper M P (2002) Nonresponse in the second wave of longitudinal household surveys, 259-272 in R M Groves et al (ed.s) *Survey Nonresponse*, New York: Wiley.

McGonagle K A, Couper M P and Schoeni R F (2011) Keeping track of panel members: an experimental test of a between-wave contact strategy, *Journal of Official Statistics* **27**, 319–338.

Uhrig S C N (2008) The nature and causes of attrition in the British Household Panel Study, *ISER Working Paper* 2008-05.

Watson N and Wooden M (2009) Identifying factors affecting longitudinal survey response, 157-181 in P Lynn (ed.), *Methodology of Longitudinal Surveys*, Chichester: Wiley.

Table 1: Potential predictor variables

Variable name	Description	Format / categories
xpmove	Whether respondent expects to move in next 12 months	0 = no; 1 = yes
age	Age at wave 1	1 = 60+; 2 = 16-17; 3 = 18-19; 4 = 20-24; 5 = 25-29; 6 = 30-34; 7 = 35-39; 8 = 40-44; 9 = 45-49; 10 = 50-54; 11 = 55-59
sex	Sex	1 = male; 2 = female
mastat	Marital status	1 = single, divorced, separated or widowed; 2 = married; 3 = living as couple
employed	Current employment status	1 = currently employed; 2 = not currently employed
tenure	Housing tenure	1 = owned outright; 2 = mortgage; 3 = rented from local authority; 4 = rented from housing association; 5 = rented from employer; 6 = privately rented, unfurnished; 7 = privately rented, furnished
nonukborn	Whether respondent was born in the UK	0 = born in UK; 1 = not born in UK
selfcomp	Whether respondent completed the wave 1 self-completion questionnaire	0 = no; 1 = yes
proxy	Whether wave 1 participation was by personal interview or by proxy	1 = personal interview; 2 = proxy interview
nomv20yr	Whether the respondent has moved home in the past 20 years	0 = moved home at least once since 1989; 1 = has not moved home since 1989 (or since birth if born after 1989)
country	Country of residence	1 = England & Wales; 2 = Scotland; 3 = Northern Ireland
nunder15	Whether there is a child aged under 15 in the respondent's household	0 = no person aged under 15 in household; 1 = at least one person aged under 15 in household
dwelltype	Type of dwelling	1 = detached house; 2 = semi-detached house; 3 = end-terrace house; 4 = mid-terrace house; 5 = purpose-built flat in block with fewer than 10 floors; 6 = purpose-built flat in block with 10 or more floors; 7 = converted flat in block with fewer than 10 floors; 8 = converted flat in block with 10 or more floors; 9 = dwelling attached to a business; 10 = bedsit; 11 = sheltered accommodation; 12 = other
genhealth	General health	1 = 'excellent'; 2 = 'very good'; 3 = 'good', 'fair' or 'poor'
august	Whether respondent is in August sample and, if so, whether he/she is aged under 18	1 = not August sample; 2 = aged 18 or over, August sample; 3 = aged 16 or 17, August sample
web	Whether there is a computer and internet connection in the respondent's household	1 = PC and internet in household; 2 = no PC or no internet in household
quals	Highest level of qualification obtained by the respondent	1 = degree or higher; 2 = A levels; 3 = GCSE; 4 = other; 5 = none
numadult	Number of persons aged 16 or over in the respondent's household	1 = 1; 2 = 2; 3 = 3 or more
floors	Number of floors in the respondent's dwelling	1 = 1; 2 = 2; 3 = 3; 4 = 4 or more
flrlevel	Floor level of the living accommodation	0 = 1 to 4; 1 = 5 or more ("high rise")
entryphone	Whether there is an entryphone at the respondent's address	0 = no; 1 = yes

Table 2: Final model

Variable	Value	Odds ratio	Standard error	p
xpmove * age	no / 16-17	2.62	0.78	0.001
	no / 18-19	3.62	0.98	0.000
	no / 20-24	5.42	1.13	0.000
	no / 25-29	4.77	0.99	0.000
	no / 30-34	3.53	0.78	0.000
	no / 35-39	2.98	0.68	0.000
	no / 40-44	3.08	0.69	0.000
	no / 45-49	1.77	0.46	0.028
	no / 50-54	1.64	0.45	0.070
	no / 55-59	1.28	0.37	0.388
	yes / 60+	3.69	1.12	0.000
	yes / 16-17	7.77	2.62	0.000
	yes / 18-19	12.16	3.33	0.000
	yes / 20-24	11.92	2.52	0.000
	yes / 25-29	9.03	1.97	0.000
	yes / 30-34	9.31	2.11	0.000
	yes / 35-39	9.30	2.30	0.000
	yes / 40-44	8.55	2.34	0.000
	yes / 45-49	6.13	1.89	0.000
	yes / 50-54	5.51	2.08	0.000
	yes / 55-59	5.03	2.08	0.000
sex	female	0.85	0.06	0.023
mastat	married	0.70	0.07	0.000
	living as a couple	0.82	0.08	0.057
employed	not currently employed	1.41	0.12	0.000
tenure	mortgage	1.17	0.18	0.316
	rent LA	1.74	0.31	0.002
	rent HA	1.85	0.35	0.001
	rent employer	3.44	0.94	0.000
	rent private unfurnished	3.92	0.64	0.000
	rent private furnished	6.65	1.10	0.000
nonukborn	not born in UK	1.58	0.14	0.000
selfcomp	self-completion completed	0.70	0.07	0.000
proxy	proxy interview	1.68	0.26	0.001
nomv20yr	has not moved	0.59	0.08	0.000
country	Scotland	0.75	0.11	0.056
	Northern Ireland	0.49	0.09	0.000
nunder15	under-15 in household	0.87	0.07	0.101
dwelltype	semi-detached house	1.09	0.15	0.523
	end-terrace house	1.79	0.28	0.000
	mid-terrace house	1.46	0.20	0.005

	purpose-built flat <10 floors	1.67	0.27	0.002
	purpose-built flat 10+ floors	1.82	0.30	0.000
	converted flat <10 floors	1.92	0.36	0.000
	converted flat 10+ floors	2.68	0.93	0.005
	attached to business	4.90	2.66	0.003
	bedsit	0.64	0.34	0.394
	sheltered accommodation	2.56	1.60	0.132
	other	1.23	0.26	0.345
genhealth	very good	0.86	0.08	0.111
	good, fair or poor	0.81	0.08	0.028
august	aged 18 or over, August sample	1.35	0.17	0.013
	aged 16-17, August sample	2.59	1.34	0.065
web	no internet or no PC	1.22	0.10	0.019
quals	A levels	0.97	0.10	0.769
	GCSEs	0.85	0.09	0.122
	other	0.62	0.13	0.023
	none	1.07	0.12	0.584
constant		0.01	0.00	0.000

Table 3: Association between model predictions and observed outcomes

Predicted probability	< 0.025	[0.025, 0.05)	[0.05, 0.075)	[0.075, 0.10)	[0.10, 0.15)	[0.15, 0.20)	[0.20, 0.30)	[0.30, 0.40)	[0.40, 0.50)	[0.50, 1.00]
Observed proportion	0.0086	0.0390	0.0696	0.0942	0.1320	0.1582	0.2222	0.3614	0.4096	0.5612
n	15,890	3,363	1,494	828	917	531	558	321	188	98