MODELLING FOR NON-RESPONSE IN A LONGITUDINAL SURVEY

Lecily Hunter, Sylvie Michaud, Virginia Torrance, Statistics Canada
L. Hunter, 5D5 Jean Talon Bldg, Tunney's Pasture, Ont., Canada K1A 0T6

KEY WORDS: Longitudinal surveys, non-response, logistic regression

1. Introduction

Statistics Canada will launch a major panel survey of households in 1994 called the Survey of Labour and Income Dynamics (SLID). The survey will follow individuals and families for six years, collecting information on their labour market experiences, income and family circumstances. SLID has a strong base within Statistics Canada. Its origins are in several surveys, including the Labour Force Survey (LFS), the Survey of Consumer Finances (SCF) and the Labour Market Activity Survey (LMAS). Both LFS and SCF are cross-sectional surveys. As cross-sectional surveys, they offer a series of snapshots and are useful and efficient tools for monitoring trends at aggregate levels. The LMAS served both as a longitudinal and as a cross-sectional survey. Two panels have been conducted to date, a two year panel (1986-1987) and a three year panel (1988-1990). For each longitudinal panel, people that participated in the first wave were interviewed and traced. All people living with them in the following waves were also interviewed (but not traced).

Because SLID wants to interview people for six years, conducting two interviews per year, it is felt that non-response rates and attrition of the sample are key issues to examine in the survey design. Different studies are currently being conducted on non-response to the LMAS in hopes of finding approaches that will minimize the impact of non-response on the SLID data. This paper will discuss one of the studies that is underway, to investigate the possibility of fitting a model or models to (a) adjust for non-response in weighting and (b) predict non-response in the following year.

2. LMAS survey design and non-response.

For the first interview of the panel, LMAS is conducted as a supplement to the January Labour Force Survey (LFS). All eligible respondents from the LFS are included in the LMAS sample. In the subsequent waves, for the longitudinal component of LMAS, all people that were respondents to the first wave are interviewed in January of the following year(s). People are traced if they have moved.

LFS uses a multiple stage sample design. A stratum is defined based on a geographical and a size breakdown of Canada. At least two distinct PSU's (primary sampling units) are selected within each stratum (to allow unbiased variance estimates). LFS initial weights go through a series of adjustment factors at the stratum level (stabilisation weight, cluster weight, non-response adjustment...) to produce a sub-weight. This sub-weight is then adjusted to population estimates (with a province/age-group/sex adjustment, plus an adjustment by economic region and Census metropolitan area) to produce a final weight. More details can be found in [1].

For the LMAS longitudinal sample, non-response adjustment is done at the stratum- component level (component correponds to a PSU or a group of PSU's), as defined for the LFS. A post-stratification is then done to adjust the non-response adjusted weights to population estimates (province/age-group/sex).

When the LMAS file was evaluated, it was found that non-response was quite different among certain groups of people:

-movers had a non-response rate (including people that could not be traced) of 20% while non-response for non-movers was about 2%,
-based on characteristics from wave1, people that were employed in wave 1 were responding more (after three years) than people who were unemployed in wave1,
-similarly, people that were married in wave 1 were more responding in year 3,
-people who lived in non-urban areas in year 1 were also more represented in the sample after three years.

The different characteristics between respondents and non-respondents suggested a couple of possibilities. First, that non-response adjustments should perhaps be done at some different level than stratum-component. Second, that it may be possible to predict the next year's non-respondents on the basis of the current year's observed characteristics.

3. Modelling

Two possible approaches were considered for the non-response adjustments in weighting: ratio adjustments within population subgroups, and regression model-fitting. The model-fitting approach was chosen because it was felt that this work could be used to serve other purposes as well. In particular, there are two possible uses of a non-response model in the context of our longitudinal survey: the prediction of non-response and a non-response weighting adjustment.

While the same model could probably not be used for the two purposes, it was hoped that a base set of variables could be identified which would be common to the models. A small set of additional variables would be unique to the different purposes of the models. For example, the characteristic most correlated with non-response is whether or not the person moved since the time of the last interview (non-response being the result of the inability to trace in this case); clearly this information could be used in the model for weighting, but would not be available at the time of the previous interview (since the event had not yet taken place). At best, we could hope to find a variable or set of variables that is correlated with subsequent moves to use in the prediction model.

The Model

Logistic regression was used to create the model. This type of a model was chosen because non-response is a binary dependent variable. Logistic regression was preferred over discriminant analysis since logistic regression has fewer assumptions and is essentially as efficient as discriminant analysis (Harrell, 1983).

The multiple logistic response function is

$$E\{Y|\mathbf{X}\} = [1 + \exp(-\beta^T\mathbf{X})]^{-1} \qquad (1)$$

where

Y is the dependent variable,
β is the column vector of regression parameters,
X is the n x (p-1) matrix of independent variables.

Equation (1) expands to

$$E\{Y|\mathbf{X}\} = [1 + \exp(-\beta_0-\beta_1X_1- \cdots -\beta_{p-1}X_{p-1})]^{-1}. (2)$$

The dependent variable, Y_i, in this analysis indicated if the ith respondent to the 1986 survey had become a non-respondent to the 1987 survey. Therefore, for the i^{th} individual

$Y_i = 1$ if the i^{th} individual did not respond in 1987,
$Y_i = 0$ if the i^{th} individual did respond in 1987.

The multiple logistic regression model states that Y_i are independent Bernoulli random variables with

$$E\{Y_i|\mathbf{X}_i\} = [1 + \exp(-\beta^T\mathbf{X}_i)]^{-1} \qquad (3)$$

and X_i is the vector of p-1 independent variables associated with the i^{th} individual.

Denoting $P(Y=1|\mathbf{X})$ as $\pi(\mathbf{X})$, the logit transformation is defined as

$$g(X) = \ln\left[\frac{\pi(X)}{1-\pi(X)}\right]$$
$$= \beta_0 + \beta_1X_1 + \beta_2X_2 + \ldots$$

The Data

The 1986/87 panel of LMAS was used to fit and evaluate the non-response models. The dataset consisted of 66,817 individuals, of which 3,385 (5%) were non-respondents to the 1987 interview. Demographic variables that were likely to be related to non-response were chosen from the 1986 LMAS master file as possible independent variables for the model. One additional variable was collected in 1987 for all individuals - whether or not the person changed address since the 1986 interview.

Variables

The variables examined for inclusion in the non-response model were:

Province at 1986 interview
Urban/Rural area at 1986 interview

Household size at 1986 interview
Type of dwelling (house; other) at 1986 interview
Status of dwelling (owned; rented) at 1986 int.

Sex
Age at 1986 interview
Marital status at 1986 interview
School attendance (full/part time/none) in 1986
Highest level of education at 1986 interview

Any employment in 1986
Any unemployment in 1986
Any out-of-labour-force in 1986
Number of jobs in 1986
Any short tenure jobs (< 2 years) held in 1986
Any long tenure jobs (2+ years) held in 1986
Any absences from work in 1986
Industry of job(s) in 1986

Average weekly income (over all jobs) in 1986
Received any unemployment insurance in 1986
Received any welfare in 1986

Moved (changed address between 1986 and 1987 interview)

Categorical variables were analyzed and manipulated so that the final representation of this information was by groups of binomial (0-1) variables. The differences between respondents and non-respondents with respect to the independent variables were analyzed. The correlations between all pairs of these variables were examined to find any potential multicollinearity.

The Sample File

PROC LOGIST in SAS was used to fit the logistic regression model. Because of the size of the dataset, the procedure required a large amount of computer resources. Therefore, it was decided to select a sample of households from the original file to be used for the model-building stage. The sample file consisted of all households that contained a non-respondent plus a random selection of an equal number of households containing respondents only. This was preferable to a simple random sample since the variables associated with non-response could be more easily identified by using all the non-response information that was available. The parameters of the regression model were estimated using the full dataset.

Regression Procedures

First, a stepwise linear regression procedure was used to identify potentially useful variables for the modelling. This reduction in the choice of variables resulted in fewer variables to be entered into the logistic procedures saving considerable computer resources.

The variables given in the STEPWISE procedure were entered into the SAS procedure PROC LOGIST with the BACKWARD and FAST options. These options allowed LOGIST to use an approximate backward elimination method to eliminate nonsignificant variables. Different logistic regression models were fitted to the full dataset using combinations of the most significant variables identified from the sample file. A consideration in choosing the model was the number of variables. It was desired to have a model with a small number of variables so that utilizing the model would be simple.

4. Prediction Model

For the prediction model, it was decided to focus on the largest group of non-respondents: those who changed address since the 1986 interview and we were not able to trace. It was felt that procedures could be implemented in the field most effectively for this group. The SLID interview will be conducted using computer-assisted interviewing. If an appropriate model can be found, it may be possible to identify at the time of the current year's interview those people with a higher probability of becoming movers and non-respondents the following year so that extra efforts can be made to keep them in the sample (ie. collecting extra information for tracing, more frequent contacts between waves, more feedback on the importance of their participation, etc). Because the amount of information collected in the survey is large and memory space may be a concern, it is important that the model we choose for prediction be kept as simple as possible.

The BACKWARD option of PROC LOGIST run on the sample file identified eight variables as good predictors of move/non-response.

Male	(MALE)
Single	(SINGLE)
Rented dwelling	(RENT)
Any unemployment	(ANYUN)
Any out-of-labour-force	(ANYOUT)
Received welfare	(WELFARE)
Household size	(HHS)
Age	(AGE)

Before fitting the models on the full dataset, the two continuous variables (household size and age) were examined for linearity in the logit. Plots of the variables showed that neither appeared linear. Non-response was high for ages 16-24, low for ages 25-54, and rose slightly for ages 55+. Some transformations were attempted, but without success. It was decided instead to create age groups, and replace the continuous age variable with two binomial variables for age (AGE1, AGE2). Because few people in the sample came from very large households, it was decided to group together households of size 8 or more and assign a value of 8 to the recoded variable. A plot of non-response versus the recoded household size showed essentially a V-shaped distribution. The transformation ABS(HHS - 4.5) was used to linearize the variable. The transformed household size variable was called HHSTRANS.

Four models were fitted to the full dataset: (1) using all eight variables; (2) using all except SINGLE; (3) using all except SINGLE and ANYUN; (4) using all except SINGLE and MALE. The statistics for evaluating the fit of the models indicated very little differences between the four models. For operational reasons, we preferred to keep the number of variables to the minimum that would provide a good fit. Therefore, the choice was between model (3) and (4). The Pearson residuals were plotted against the fitted values and the residual plots were examined. Model (3) residuals indicated a slightly better fit with fewer extreme values. Again using the sample file, the data were examined for the presence of two-way interactions between the variables in the model. None were found to be significant.

Classification Tables

The intended use of this regression model is to identify people who are likely to move and not respond to the next wave of the survey. Based on the model, people were given a probability of moving and non-responding based on the 1986 characteristics. In order to identify a person as a potential move/non-respondent, it is necessary to choose a cut-off for the predicted values calculated for each person. Any person having a predicted value higher than the cut-off is classified as a potential non-respondent; people with predicted values equal to or below the cut-off are classified as potential respondents.

For the model to be useful in prediction it must be possible to find a cut-off such that the number of move/non-respondents classified correctly is high, while the number of people mis-classified as move/non-respondents is low. For each person on the 1986 file, the probability of move/non-response was calculated under the fitted model. Using different levels of cut-offs, each person above the cut-off was classified as a potential move/non-respondent. The classification was then compared with the person's actual moved and response status from the 1987 interview. Tables are given below for cut-off values of .02 and .05. The classification tables show the number of actual move/non-respondents who were correctly classified, along with the number of people who were incorrectly classified as move/non-respondents. For example, when the cut-off is changed from 0.02 to 0.05, the number of correct classifications decreases to 1137 and the number of mis-classifications also decreases to 11039. Note that although the number of mis-classifications is high, when it is taken as a proportion of all people who were not move/non-respondents, the proportion is low.

Table 1. Classification table for prediction model. (Cut-off value= 0.02)

	CLASSIFIED AS MOVEDNR	TOTAL
MOVEDNR = no	25645	64662
	39.7%	100.0%
MOVEDNR = yes	1650	2155
	76.6%	100.0%

Table 2. Classification table for prediction model. (Cut-off value= 0.05)

	CLASSIFIED AS MOVEDNR	TOTAL
MOVEDNR = no	11039	64662
	17.1%	100.0%
MOVEDNR = yes	1137	2155
	52.8%	100.0%

Whether or not this model will be practical to implement will likely depend on the cost of keeping people in the sample relative to the amount of sample attrition over the six years. Until it is tested in the field, it is unknown if the additional efforts made for the potential move/non-respondents will actually be successful in making tracing easier.

5. Weighting Model

The second use for a non-response model in a longitudinal survey is to make adjustments to the weights of the respondents in the second year (1987). For this model, the dependent variable was total non-response, and the independent variables were characteristics observed the previous year (1986) plus the current year's information (1987) on whether or not the person moved.

The BACKWARD option of PROC LOGIST was used with the sample file to identify eight variables related to non-response.

Male	(MALE)
Single	(SINGLE)
Rented dwelling	(RENT)
Any employment	(ANYEMP)
Highest educ=secondary	(EDUCSEC)
Moved since 1986 interview	(MOVED)
Household size	(HHS)
Age	(AGE)

Before fitting the models on the full dataset, the two continuous variables (household size and age) were examined for linearity in the logit. As with the prediction model, the age variable was replaced with two binomial variables for age (AGE1, AGE2), and the same transformation was applied to household size (HHSTRANS).

Four models were fitted to the full dataset: (1) using all eight variables; (2) using all except RENT; (3) using all except EDUCSEC; (4) using all except EDUCSEC and AGE. Although all eight variables were significant using the sample file, when the models were fitted to the full data file, certain ones no longer appeared important. However, it was decided to retain them in the models anyway. The statistics for evaluating the fit of the models indicated few differences between the four models. The Pearson residuals were plotted against the fitted values and the residual plots were examined. Model (3) residuals indicated a slightly better fit with fewer extreme values. Again using the sample file, the data were examined for the presence of two-way interactions between the variables in the model. Two sets of interactions were added to the model: the (AGE1 AGE2)*HHSTRANS and (AGE1 AGE2)*SINGLE. A summary of the fitted values for this model is given below. Note that the age and single variables as well as their interactions are not statistically significant. Nevertheless, when a model was fitted with these variables removed, it was found that there were more extreme values in the residuals.

Using the parameter estimates from the final model, predicted probabilites of non-response were calculated for all respondents to the 1987 interview. The non-response weighting adjustment was done by dividing the 1986 starting weight by (1-predicted probability). This gave a non-response adjusted 1987 weight. A post-stratification was then done to adjust the weights to population control totals. This was done by ratio adjusting the weights within categories of province-sex-agegroup, to produce a 1987 final weight.

Evaluation of the Weights

Because the LMAS is a longitudinal survey, the same people are present in the sample in both years. The only difference between the people on the 1986 file and the people on the 1987 file is that some are missing from the 1987 file because of non-response. If the non-response weighting adjustment is adequate, there should be no difference in estimates obtained from the 1986 respondents and estimates obtained from the 1987 respondents when tabulating on 1986 characteristics. A number of demographic and labour-related characteristics were evaluated. Estimates were calculated using the 1986 weights, the 1987 model-adjusted weight, and the 1987 regular weights (doing a ratio-adjustment at low geographic levels for non-response adjustment). For each characteristic a 95% confidence interval was calculated for the estimate based on the 1986 weights. The two 1987 estimates were compared for differences to the 1986 estimates as well as differences to each other.

Of all the characteristics compared, only one 1987 estimate was outside the 1986 confidence interval: weeks employed=49-52 using the regular weighting. One pattern was clear, however. The estimates using the model-based weights were consistently closer to the 1986 estimates than those using the regular method of weighting. The two 1987 estimates were also compared using the sub-weights (before doing the post-stratification adjustment) and at provincial as well as national levels. In general, differences between the 1987 estimates were greater using sub-weights than they were using final weights. Differences were also greater for labour-related characteristics than for demographic characteristics; differences were greater for variables included in the non-response model; differences were greater at provincial level than at national level. Although the size of the differences are small, the indications are that the model-based approach is performing better. It is expected that when the non-response is extended over more years, the gains will be greater.

6. Future work

The usefulness of the model to predict move/non-response is not clear. Although we are able to correctly identify a large proportion of actual move/non-respondents, at the same time the model incorrectly classifies a large number of people who were not move/non-respondents. Even if the model is implemented, there is no guarantee that the extra measures taken to prevent the move/non-response outcome from occurring will succeed. For example, if the preventive action is to collect an extra contact name, how much good will that do; the first contact name provided by the respondent was not successful so how reliable will any additional information be? On the other hand, if the preventive action is successful, is it operationally easier or better to apply it to all respondents instead of just those identified under the model? It may be necessary to answer these questions before any testing or further modelling work can be done.

The weighting model on the other hand does show a great deal of promise. Although the differences realized with the model-based approach to weighting were small when tested over a one-year interval of non-response, it is expected that the gains will be greater over a longer period. For the future then, we would like to test the stability of the model with the 3-year panel of LMAS data. There are some operational questions which must first be answered. For instance, when the third year of data is added, some of the non-respondents will have information from both the first and second year, while others will have information from the first year only. In addition we may have third year respondents for whom we have no second-year information. Exactly how these complexities will be handled in the model has yet to be decided. Once these issues have been solved and tested on the 3-year LMAS panel, we would next like to do a simulation study to look at the longer-term behaviour of the model since the SLID respondents will be followed for six years. If this approach to weighting seems feasible, we would like to investigate the possibility of combining the non-response adjustment and the post-stratification into a single regression model. Finally, we have yet to look at variance estimates under the model-based approach; this will be essential if we are to make proper evaluations of the model-based estimates.

References

[1] M.P. Singh, J.D. Drew, J.G. Gambino, F. Mayda, Methodology of the Canadian Labour Force Survey 1984-1990, Statistics Canada publication, Catalogue 71-256

[2] The Labour Market Activity Survey, 1986-87 Longitudinal File, Microdata User's Guide, Special Surveys Group, Statistics Canada

[3] D.W. Hosmer,Jr., S. Lemeshow, Applied Logistic Regression, John Wiley & Sons, 1989

[4] F.E. Harrell, "The LOGIST Procedure", SUGI Supplemental Library Guide, Version 5 Edition, Cary, NC: SAS Institute Inc., 1986