Model-Based Reweighting for Nonresponse Adjustment

David A. Binder, Sylvie Michaud, and Claude Poirier
Statistics Canada

In virtually every survey, no matter how carefully it is designed, we must accept the fact that some data will be missing. Some measures must be taken to deal with such nonresponse. Over the years, a host of techniques has been developed. Many of the methods for coping with nonresponse make use of models, either explicitly or implicitly. Even the most ardent advocates of the pure design-based school will resort to some model assumptions when it comes to adjusting for nonresponse. This presents a new set of problems associated with the statistical inferences, since the randomization distributions on which the inferences are based are no longer purely design-based, unless the nonresponse mechanism can be considered to be part of that design.

In this paper, we shall focus on the implications of the estimation method to be used and the amount of information about the nonrespondents that is available. It will be assumed that the prime focus of the survey is to obtain estimates of descriptive statistics, such as means, totals, differences, and ratios. We will also restrict on unit nonresponse. In practice, this definition is extended to other cases, where there is insufficient usable data from the respondents. The usual method for dealing with unit nonresponse is to use an "appropriate" weighting procedure to compensate for the nonresponse. (We define weighting procedures here broadly to include weight adjustments implied by regression, ratio or similar estimation techniques using auxiliary data.)

In the next section, we discuss the basic theory underlying many of the adjustment methods. In the two sections which follow, we give examples of two surveys at Statistics Canada, where some of these models have been studied recently. We summarize our findings in the last section of this paper.

■ Some Generalities

Estimation

In general, we are interested in means, totals, ratios, etc. of survey variables. We denote the value of the i-th survey variable for the k-th respondent as y_{ik}. In cases where the occasion, t, is relevant, we can use y_{ikt} instead. A sample is selected according to some well-defined sampling plan.

We use s to refer to the selected sample. The problem that we are addressing here is the case where the y-values are unobtainable (as opposed to values that could be obtained but are in error). We denote by $s' \subseteq s$ the set of units for which we obtain useable y-values. (The subscript t is implied, where appropriate, for longitudinal surveys.)

Formally we assume that, given the sample, s, the set of responding units, s', follow a probability distribution $p(s' \mid s)$. This is completely general, allowing for correlated response patterns. It also allows for the classical case, where it is assumed that the response behaviour is nonrandom and is an inherent attribute of the selected respondents, just like the survey variables. We now consider methods of nonresponse adjustment which we refer to as generalized reweighting methods. Associated with each responding unit, k, we have an adjusted weight given by

$$w_k'(s',s) = g_k'(s',s) w_k(s) \quad ,$$

where $g_k'(s',s)$ is a weight adjustment that makes use of auxiliary frame data, as well as other information that may be available for the nonresponding units. This allows the weight adjustment to depend

on survey values that were observed on previous occasions from a longitudinal survey. We assume that the estimator of a total for a y-variable on the t-th occasion is given by

$$\hat{Y}_{it}^{(GR)} = \sum_{k \in s_t'} w_k'(s_t', s_t) y_{ikt} \quad . \qquad (1)$$

We let $\rho_k(s)$ be Pr($k \in s' | s$). (2)
Conditions required to be asymptotically consistent with respect to the original design and the response probabilities are:

1) the probability distribution of s' given s depends only on the auxiliary data and the survey data from previous occasions, but not directly on the y-values for the current occasion,

2) the limiting expectation of

$$g_k'(s', s) \text{ is } \{E[\rho_k(s)]\}^{-1} , \qquad (3)$$

3) the variance of $\hat{Y}^{(GR)}$ is asymptotically zero. (4)

If (3) is violated, then the expectation of $\hat{Y}^{(GR)}$ is

$$\sum E[g_k'(s', s)] E[\rho_k(s)] y_k \quad . \qquad (5)$$

The form of this bias is important, because if one were to impose model assumptions on the y-variables, it is possible that the model-bias becomes small. However, for those who wish to make the fewest model assumptions, it is clear that one should restrict attention to adjustment methods which yield condition (3) as closely as possible. This implies that the weight adjustment should reflect the propensity to respond as nearly as possible. Of course, the probability mechanism generating these response probabilities is generally unknown, so the weight adjustment must necessarily be model-based.

Another important feature of (3) is that if there are some "hard-core" nonrespondents -- that is, units where $\rho_k=0$ -- there would be no consistent estimates.

A lot of examples of different weighting adjustments may be found in the literature. Some examples of different forms with some basic properties can be found in Binder et al. (1994).

The next two sections deal with the use of the logistic regression to predict nonresponse.

■ Surveys of Labour and Income Dynamics

In 1994, Statistics Canada launched a major panel survey of households called the Survey of Labour and Income Dynamics (SLID). The survey follows individuals and families for six years, collecting information on their labour market experiences, income and family circumstances. Its origins are in several surveys, including the Labour Market Activity Survey (LMAS). (LMAS was a panel survey. Two panels were conducted -- a two-year panel in 1986 and a three-year panel in 1988.) Different studies are currently being conducted on nonresponse to the LMAS in hopes of finding approaches that will minimize the impact of nonresponse on the SLID data.

Similarly to its predecessor (LMAS), the longitudinal sample for SLID is selected from the sample of dwellings that participated in the Labour Force Survey (LFS) in January 1993. The LFS has a response rate of 95%. Out of those respondents, close to 90% agreed to participate in SLID. This subsample of respondents (a sample of 15,000 households) is defined as the longitudinal sample, representative of the Canadian population as of January 1993.

Attritional nonresponse will be compensated with a weighting adjustment. Imputation will be used to compensate for some nonresponse (for example, nonresponse that is non-attritional). The weighting will include the following steps:

- calculation of the initial weight (based on the sample design),
- nonresponse adjustment, and

- post-stratification (province, age groups, and sex) to the 1993 population estimates.

The longitudinal panel of LMAS has been used as the research vehicle for the nonresponse modelling and weighting adjustments.

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

When the LMAS file was evaluated, it was found that nonresponse was quite different among certain groups:

- Movers (including people that could not be traced) had a nonresponse rate of close to 20%, while nonresponse for non-movers was about 2%. This was by far the characteristic that presented the most differences.

- Based on characteristics from Wave 1, persons who were employed in Wave 1 had higher response rates after three years than those who were unemployed in Wave 1.

- Similarly, persons who were married in Wave 1 had higher response rates in Year 3, compared to those who were single in Year 1.

- Persons who lived in non-urban areas in Year 1 had higher response rates after three years.

The different characteristics between respondents and nonrespondents suggested that nonresponse adjustments should be done at some level different than stratum-component. Logistic regression was used to model the nonresponse behaviour. The multiple logistic response function is

$$\text{logit}(p) = \log[p/(1-p)] = \beta' x \quad ,$$

where p is the probability of response to the 1987 survey for a 1986 survey respondent, β is the column vector of regression parameters, and χ is the vector of independent variables.

The dataset for the 1986/87 panel of LMAS consisted of 66,817 individuals, of which 3,385 (5%) were nonrespondents to the 1987 interview. Demographic variables that were likely to be related to nonresponse were chosen from the 1986 LMAS master file as possible independent variables for the model. More than 20 variables were examined for inclusion in the nonresponse model. The model was fitted on a subsample of records.

First, a stepwise linear regression procedure was used to identify potentially useful variables for the modelling. The model is used to make adjustments to the weights of the respondents in the second year (1987). For this model, the dependent variable was total nonresponse, 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. Eight variables were identified as begin related to nonresponse:

- male
- single
- rented dwelling
- any employment
- highest education=secondary
- moved since 1986 interview
- household size, to a maximum of 8, and
- age.

Before fitting the models on the full dataset, the two continuous variables (household size and age) were examined for linearity on the logit scale. As with the prediction model, the age variable was replaced with two binomial variables for age (AGE1 for persons aged 25-54, AGE2 for persons aged 55-69; the survey was conducted for person aged 16-69), and a transformation was applied to household size (HHSTRANS=|HHS-4.5|). Two sets of interactions were added to the model: the (AGE1 AGE2)*HHSTRANS and (AGE1 AGE2)*SINGLE. 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 probabilities of nonresponse were calculated for all respondents to the 1987 interview and a nonresponse adjustment was made. Finally, a post-stratification adjustment to population control totals at the province/sex/age-group level yielded the 1987 final weight.

If the nonresponse 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 nonresponse adjustment). The two 1987 estimates were compared for differences to the 1986 estimates, as well as differences to each other.

The estimates using the model-based weights were consistently closer to the 1986 estimates than those using the regular method of weighting. A number of the results are presented in Binder et al. (1994). Differences to the benchmark were done before and after the post-stratification adjustment. Differences were much larger before the post-stratification than after post-stratification adjustment. Differences were greater for labour-related characteristics than for demographic characteristics; differences were greater for variables included in the nonresponse model; and differences were greater at provincial levels than at the 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 nonresponse is extended over more years, the gains will be greater.

The variables obtained through the stepwise regression were compared to the ones obtained with an Automatic Interaction Detection system. Moving was the first variable found. Then, separately for the movers and for the non-movers, it determined what is the best discriminating variable. The discriminant variables were quite different between the movers and the non-movers. Among movers, there was higher nonresponse found if the person had received welfare in the previous year, while for non-movers, nonresponse was higher if the persons were renting their dwelling in the previous year.

Regressions have been redone, including the interaction terms that were the most significant. As expected, those terms are now coming out as significant in the model. However, the impacts of adding these interaction terms on the estimates and the variances, instead of the previous interactions, were not significantly different.

The current results seem to indicate that a modelling approach could compensate for some of the nonresponse bias that occurs in the attrition of the longitudinal sample. However, there may be a bias in the first year of selection (those who refused to participate) that will not be taken into account. More evaluation of these nonrespondents will be done by comparing results from an administrative file match.

Since all our analyses were performed using LMAS data, it will be necessary to re-evaluate the variables selected for the SLID implementation. For example, for the first panel in SLID, the interview is done using Computer-Assisted Interviewing. This may have an impact of the response mechanisms (for example, an interviewer effect may be present).

■ Farm Financial Survey

The Farm Financial Survey (FFS) has been a regular agricultural survey since 1980. The objective of the survey is to gather financial information on Canadian farmers. The survey collects information on revenues, expenses, assets and liabilities. Crop and livestock information is also collected to measure physical characteristics of the farms. Due to the collection of sensitive data, a low response rate has always been observed for the survey. A study was initiated on the 1992 survey data to identify the causes of nonresponse and possible solutions to reduce its impact on the estimates.

The population of interest consists of all Canadian farms active for the reference year, excluding the multi-holdings companies, the institutional farms, the community pastures, the farms on Indian Reserves, and the farms with less than $2,000 in sales. The survey population is represented by a list frame and an area frame. The 1992 list frame was a register of all of the 1986 Census farms, without the farms defined by the above exclusion rules. The list frame was stratified within each province by farm type and by farm size. The farm size was defined by the total farm assets derived on the Census.

The area frame was used to compensate for the undercoverage due to the Census itself or caused by new farms which started their activities since 1986. Basically, the area frame was a list of land segments outlined on topographic maps. Stratified replicates of segments were selected from the area frame. All farmers operating some land in the sampled segments were enumerated, and a register was created. There were 1,153 area frame farms that did not appear on the list frame that were contacted for the FFS, as for other agricultural surveys. In addition to the area frame farms, a stratified sample was selected from the list frame to obtain an overall sample of about 12,000 farms. See Britney and Poirier (1992) for more details on the 1992 FFS sample design.

Domain estimation within each stratum was performed to obtain estimates of level from both the list and area samples. The simple expansion estimator was used on the 1992 list sample. The initial weighting was done by stratum using the population size over the ***observed*** sample size, so that a nonresponse adjustment is made at the stratum level. For the area frame, the estimation was done separately by replicate. For a given replicate, the data were aggregated at the segment level, by applying to the farm data factors corresponding to the proportion of the farms within the segment. Then, the segment totals received expansion (π) weights to represent the population. When nonresponse occurred for an area farm, the respondents within the same segment were reweighted on an area basis to compensate the farm land for which data are unavailable. For both the list and area units, partial nonresponses were donor imputed and used the same way the regular respondent were. Details are given by Maranda (1989).

The nonresponse observed in the 1992 Farm Financial Survey was relatively important. The FFS questionnaire was relatively long, with many sensitive questions related to the financial balance sheet. The resulting total unit-level refusal rate of about 15% across the country was the highest of our agricultural surveys. In addition to the total refusals, the no-contacts represented another 5% of the sample. Some provinces presented higher nonresponse rates than others. In Saskatchewan, data were unavailable for almost 30% of the sampled farms.

The potential causes that were studied on the 1992 FFS data are:

- **The frame origin** -- Area frame farms vs list frame farms.
- **The farm size** -- It was evaluated using the farm assets and sales obtained from the 1986 Census of Agriculture. This size was available only for the list units.
- **Geography** -- Census divisions were used.
- **Farm type** -- The farmer's availability depends on the type of his farm. Seven categories of farm type were used to differentiate the farms.
- **Response burden** -- The overlaps with the December Stock Survey and the January Livestock Survey were both studied to verify impact on the response rates. The effect of the overlap with the previous FFS, held in 1990, was also investigated.

The independence tests, conducted with a confidence level of 5%, identified certain causes of nonresponse. First, within each province except Ontario, the farm type had a high impact on nonresponse. Also, the farm size, measured in term of assets, affected the response rates in most of the provinces, but no significant impact was due to the sales variable. The geographic location and the re-

sponse burden generated by the previous FFS survey significantly affected the probability to respond in three provinces. Finally, the frame origin and the overlap with the January Livestock Survey or the December Crops Survey seemed to not affect the response status at all.

As in the previous section, nonresponse was modelled using a logistic regression. The analysis was done separately by province. Using frame origin as an independent variable, the results confirmed the previous conclusions of no frame effect. Since some variables were not available for the area sample and since the frame origin did not seem to affect the response, the remaining analyses were performed only on the list units, which represented more than 90 percent of the whole sample. In the rest on this paper, the results apply for the list units only. The following variables were included in the model:

- Assets (1 if assets is smaller than the median, 0 otherwise),
- Sales (1 if sales is smaller than the median, 0 otherwise),
- Type i (1 if in the ith farm type, 0 otherwise),
- Area i (1 if in the ith geographic area, 0 otherwise),
- FFS (1 if in the 1990 FFS sample, 0 otherwise),
- JLS (1 if in the 1992 JLS sample, 0 otherwise),

The farm types are: crop farms, dairy farms, cattle farms, hog farms, poultry farms, sheep farms, and unknown type of farm.

The variables that were found more significant by the BACKWARD option within the provinces were kept in the model. (See Binder et al. (1994) for full discussion of this analysis.) The most commonly selected variables were the farm types and the FFS variables. The positive FFS parameters mean that farms overlapping the previous FFS tended to have higher response rates, whereas the negative sheep farm parameters imply they tended to respond less often.

Weighted regressions were also fitted to the data using the WEIGHT statement of the LOGISTIC procedure. The weighting variable was defined at the stratum level as the design weight adjusted to the overall sample size. Stratum-level adjustments were not performed. The resulting estimated parameters were very close to the first set of estimates which, as we explained in the previous section, is highly desirable.

To evaluate the nonresponse adjustment, the 1992 frame values representing farm assets were estimated from the sample. Assets levels were estimated for each province with the corresponding coefficient of variation (CV), including the nonresponding units. Then, estimates based only on respondents were produced, using the original weight, adjusted for nonresponse at the stratum level only. By comparing both set of estimates, we could derive the nonresponse bias introduced by the current method. Finally, regression-adjusted estimates were produced from the above logistic model.

Estimated level and coefficient of variation were calculated, respectively, from the full sample, and y_{adj} the corresponding adjusted estimates based only on respondents. The bias associated with the adjustment model was estimated. Again, results are presented in Binder et al. (1994). The logistic-adjusted weight generally performs better, but not consistently so. In fact, the bias increases for some provinces and for the Canada total. To improve the model, inclusion of some interaction factors -- like size and farm type or size and geography -- was tried, but they were rarely kept in the model and, when they were, the resulting effects were small and their impact was negligible. The selected model did not consistently provide the expected bias adjustment. This may be caused by a low number of factors included in the model or by the fact that significant factors were used in the frame stratification. Future work might include looking for more interactions

using the Automated Interaction Detection method mentioned in the previous section.

Summary

Nonresponse adjustment through reweighting is now in common use. We have shown that the success of this technique generally depends on having available variables that can be used as good predictors of the nonresponse behaviour. Having such variables, various models can be used to adjust the estimates based on the predicted response propensities. This seems to be the best general approach. Other approaches include using estimation methods such as regression estimators to compensate for the deficiencies of the sample. We have seen that if the regression models are valid, the nonresponse bias vanishes.

We have concentrated here on asymptotic biases. However, there are still many unresolved issues for estimation of variances and construction of confidence intervals. As well, we have not properly addressed the issue of whether or not to use the sampling weights when fitting the nonresponse models. In our examples, the weighted and unweighted versions of the estimated response models gave similar results. This is highly desirable since it confirms the validity of the model.

Nonresponse problems will not go away. A better understanding of the response mechanisms, however, will lead to better survey practices in the long run.

References

Binder, D.A., Michaud, S., Poirier, C. (1994), "Model Based Reweighting for Nonresponse Adjustment," *Seminar on New Directions in Statistical Methodology*, Statistical Working Paper 23, Federal Committee on Statistical Methodology, Office of Management and Budget, Washington D.C.

Britney, H. and Poirier, C., (1992), "1992 Farm Credit Corporation: Design Documentation," Internal Paper, Agriculture Section, Business Survey Methods Division, Statistics Canada.

Maranda, F. (1989), "Proposal for NFS Estimation," Internal Paper, Agriculture Section, Business Survey Methods Division, Statistics Canada. ■