MEASURING THE IMPACT OF ALTERNATIVE WEIGHTING SCHEMES FOR LONGITUDINAL DATA

Johane Dufour, François Gagnon, Yves Morin, Martin Renaud, and Carl-Erik Särndal, Statistics Canada.
François Gagnon, Statistics Canada, Jean-Talon Bldg, 7th floor, Section C2, Ottawa, Ontario, Canada, K1A 0T6

Key Words: Longitudinal weighting; logistic regression; segmentation modelling; nonresponse adjustment

1. INTRODUCTION

The literature on longitudinal surveys proposes several approaches to producing a set of final weights to be used in data analysis. A common feature of these approaches is that they are weight modification procedures. That is, a set of initial weights is transformed into a set of final weights, in one or more steps of varying complexity.

The weight modification procedure commonly found in the literature is performed in three steps. The first step adjusts the sampling weights to reflect the fact that the size measures used in selecting primary sampling units (PSUs) are not perfectly accurate at the time of the survey. At Statistics Canada, most of the longitudinal household surveys use only a subset of the Labour Force Survey (LFS) as their data collection vehicle. Consequently, the LFS *subweights* (which are the inverse of the selection probabilities adjusted to reflect design and allocation changes over time) have to go through a first weight adjustment to compensate for the subsampling of the LFS to obtain what is called the *initial weight*.

The second step adjusts the initial weights to reduce the potential bias that can be introduced by nonresponse. The step that produces the *final weights* consists in a poststratification, or more generally a calibration (see Deville and Särndal, 1996), to benchmark the weights to population control totals known from an external source.

A common aim of all weight modification procedures is to produce, in some sense, the "best" possible set of final weights. One hopes, by different steps, to include in the final weights any information deemed to be relevant, considering that different users may be carrying out different kinds of statistical analyses with the aid of the final weights. At the same time, it is clear that parts of the weight modification procedure may have negligible impact on the final weight system. For example, a step may be simplified without causing any significant changes in the final weights.

The main objective of this paper consists of studying and measuring the change (between initial and final weights) produced by the adopted weight modification procedure. The following section presents a general framework for longitudinal weighting. Section 3 introduces a measure of change that will be used to quantify and to understand the transition from the initial weight to the final weight. Some adjustment strategies to deal with nonresponse, found in the literature, are presented in Section 4. Section 5 gives a few empirical comparisons using two of Statistics Canada's longitudinal household surveys, namely the Survey of Labour and Income Dynamics (SLID) and the National Longitudinal Survey of Children and Youth (NLSCY). Concluding remarks are given in Section 6.

2. GENERAL FRAMEWORK FOR LONGITUDINAL WEIGHTING

In a longitudinal survey, persons who are part of the original sample and who are followed through time are usually referred to as *longitudinal persons*. It is this set of persons that will be used in this paper. More precisely, the emphasis will be put on every responding unit, $k \in r$, where r is the set of responding units. The current section presents an overview of the steps followed to modify the initial weight of the longitudinal persons into a final weight.

2.1 Initial weight

For this paper's surveys, the initial weight, w_{0k}, has the structure $w_{0k} = \pi_k^{-1} * f_k$ where π_k^{-1} is the LFS subweight and f_k is the compensation factor for subsampling. Thus, we will have a weight, w_{0k}, for all $k \in s$, where s represents the entire sample. In the absence of nonresponse, the weight system $\{w_{0k} : k \in s\}$ would yield an estimator for Y according to the formula $\Sigma_s w_{0k} y_k$. Being (essentially) unbiased for Y, its only drawback would be that it does not incorporate auxiliary information in the form of known control totals for poststrata.

2.2 Nonresponse adjustment and intermediate weight

Most, if not all, surveys face nonresponse. Despite

numerous efforts, some reference units remain without response for various reasons: refusals, special circumstances, temporary absence, etc. To compensate for this nonresponse, a frequently used method consists of proportionally adjusting the initial weight of the responding reference units by the inverse of the weighted response rate. These adjustments are usually performed within response homogeneity groups (RHG), such that each group is formed of reference units believed to have a similar probability of response.

Nonresponse reduces the sample size due to the fact that the value y_k is available only for $k \in r$. For this reduced set of data, the weights w_{0k} are too small on average and the estimator $\Sigma_r w_{0k} y_k$ is not admissible since it will systematically underestimate Y.

The nonresponse adjustment consists of multiplying w_{0k}, for each unit k belonging to the same RHG, by a factor equal to the inverse of the weighted response rate in this particular group. This operation yields an intermediate system $\{w_{1k} : k \in r\}$ that could be used to construct an admissible estimator $\Sigma_r w_{1k} y_k$. It eliminates the underestimation that characterises $\Sigma_r w_{0k} y_k$. Its principal drawback is that it also fails to incorporate the available information for the poststrata.

2.3 Poststratification and final weight

A procedure commonly employed in surveys consists of modifying weights in such a way that the sum of the final weights corresponds to known population control totals of certain auxiliary variables. This procedure insures the sample will be representative, at least with respect to a certain set of variables. The choice of such variables is often limited by the availability of control totals. Demographic variables such as age and sex or geographic variables such as province or region of residence are frequently used. Various methods can be used to calibrate the weights to the chosen control totals. The choice of one of these methods is often dictated by the chosen variables and the number of control totals to respect. Weights obtained after this step are generally considered as the final weights and are noted w_{2k}. This step produces the system $\{w_{2k} : k \in r\}$, which incorporates the auxiliary information and achieves, at the same time, consistency with respect to the control totals for the poststrata.

3. MEASURE OF CHANGE FROM INITIAL TO FINAL WEIGHTS

We assume that the initial weights have been scaled so that $\Sigma_s w_{0k} = N$. Then, the three systems of weights described in section 2 verify the following relations:

$$\sum_r w_{0k} < N, \sum_r w_{1k} \approx N, \sum_r w_{2k} = N$$

Let us define

$$\overline{w_1} = \left(\sum_r w_{1k}\right) \times \left(\sum_r w_{0k}\right)^{-1} \text{ and } \overline{w_2} = \left(\sum_r w_{2k}\right) \times \left(\sum_r w_{1k}\right)^{-1}$$

The ratio $\overline{w_1}$ measures the average change of the intermediate weight system relative to the initial system. As nonresponse increases, $\overline{w_1}$ moves further away from a value of 1, which it is equal to only in the case of full response.

The ratio $\overline{w_2}$ represents the average change of the final system relative to the intermediate system. In practice, its value is close to one. For example, for Statistics Canada's LFS, $\overline{w_2}$ is close to 1.10.

The ratios $\overline{w_1}$ and $\overline{w_2}$ measure an average weight change. To measure individual weight change, we define, for all $k \in r$,

$$r_{1k} = w_{1k}\left(w_{0k}\overline{w_1}\right)^{-1} \text{ and } \quad r_{2k} = w_{2k}\left(w_{1k}\overline{w_2}\right)^{-1}$$

These quantities vary around an average value of one. More precisely, we have the following weighted averages:

$$\frac{\sum_r w_{0k} r_{1k}}{\sum_r w_{0k}} = \frac{\sum_r w_{1k} r_{2k}}{\sum_r w_{1k}} = 1$$

The quantities r_{1k} and r_{2k} will be helpful in measuring individual weight movements according to the procedure that we now explain.

The total change that the weighting system goes through, from the initial system to the final system, via the intermediate step, can be calculated by a measure of distance. In this paper, we will use the following distance:

$$D = \left[\sum_r w_{0k}\left(\frac{w_{2k}}{w_{0k}} - 1\right)^2\right] \times \left(\sum_r w_{0k}\right)^{-1}$$

This is a weighted average of the individual weight change factors $\left(\frac{w_{2k}}{w_{0k}} - 1\right)^2 = \left(\frac{w_{2k}}{w_{1k}} \times \frac{w_{1k}}{w_{0k}} - 1\right)^2$. We have $D \geq 0$, with equality holding when both of the following conditions are satisfied:

(i) absence of nonresponse ($r=s$ and $w_{1k}=w_{0k}$ for all k)

(ii) poststratification has no effect on the intermediate weights ($w_{2k}=w_{1k}$ for all k).

Normally, D > 0. One factor that will tend to increase the value of the distance D is a high nonresponse rate since, in such a case, w_{1k} is considerably larger than w_{0k}, on average.

The distance D can be decomposed into four components according to the following equation:

$$D = R_{01} + R_{12} + R_{int} + G$$

where:

$$R_{01} = (\overline{w}_1\,\overline{w}_2)^2 \frac{\sum_r w_{0k}\,(r_{1k}-1)^2}{\sum_r w_{0k}}$$

$$R_{12} = (\overline{w}_1\,\overline{w}_2)^2 \frac{\sum_r w_{0k}\,r_{1k}^2\,(r_{2k}-1)^2}{\sum_r w_{0k}}$$

$$R_{int} = 2(\overline{w}_1\,\overline{w}_2)^2 \frac{\sum_r w_{0k}\,r_{1k}\,(r_{1k}-1)(r_{2k}-1)}{\sum_r w_{0k}}$$

$$G = (\overline{w}_1\,\overline{w}_2 - 1)^2$$

The interpretation of these terms is as follows. The R_{01} term measures the individual changes that the weights undergo between the initial and the intermediate step. In the present case, it is the change due to the nonresponse adjustment. The R_{12} term measures individual changes between the intermediate and the final step, which is the change due to poststratification in the present case. The R_{int} term measures interaction between the two types of changes. Finally, the G term measures the average weight change between the initial and the final step. A high rate of nonresponse will tend to make G larger.

For a given survey, we can draw some important conclusions by looking at the relative importance of the three terms R_{01}, R_{12} and R_{int}. If R_{01} is large, and at the same time R_{12} is small, the survey is one where the nonresponse adjustment causes important movements in the weights, while poststratification does not change the weights very much. In the inverse situation, where R_{12} is rather large while R_{01} is small, the nonresponse adjustment has little effect on the weights but the poststratification causes important changes. The sign of R_{int} will indicate if both types of change move in the same direction ($R_{int} > 0$) or in opposite directions ($R_{int} < 0$). Often, R_{int} is numerically small.

4. VARIOUS STRATEGIES FOR NONRESPONSE ADJUSTMENTS

Traditionally, two approaches have been used to compensate for nonresponse: imputation and weighting adjustments. The latter is most commonly used to compensate for total nonresponse, while imputation is often used to compensate for partial nonresponse. There are several weighting adjustment procedures that can be used to compensate for total nonresponse. A common approach is to divide the sample into a set of RHGs. The idea is to group people who have a similar probability of response, so that a uniform response mechanism can be assumed in each RHG. Then, a nonresponse adjustment factor is applied within each RHG.

RHGs are based on auxiliary information available for both respondents and nonrespondents. In many surveys, little information is available for nonrespondents, beyond the PSU and stratum membership. Here, the choice of possible RHGs is very limited, and the procedure can be applied directly using the strata as the RHGs. This assumes that the response probability is the same within a stratum. This assumption may often be questionable. Nevertheless, this procedure is the best possible one under the circumstances, given that no other information is available for nonrespondents. However, in longitudinal surveys, there is often much information available for both respondents and nonrespondents from previous waves. This extensive auxiliary information can be used to create efficient RHGs in which the assumption of a uniform response mechanism within a RHG is more likely to hold. This would result in a better nonresponse adjustment and consequently in a reduction of the risk of a nonresponse bias in the survey estimates.

4.1 Variable selection method

By definition, RHGs are formed by a set of predictors of response propensity. In longitudinal surveys, many variables could be candidates for use in a nonresponse adjustment procedure. In our case, we use the following approach. First, discussions are held with subject matter experts to determine a set of potential predictors of response propensity. Note that categorical variables are usually transformed into dichotomous ones to simplify the analysis. An initial

screening of variables is then performed using univariate tests in order to reduce the large number of variables to a more manageable set. Finally, a variable selection method is used to select the best set of predictors of response. Common variable selection methods are the logistic regression (LR) method and the segmentation modelling (SM) method.

4.1.1 Logistic regression approach

LR analysis seems a natural method for selecting the best set of dichotomous predictors. We use the response status as the dependent variable and, using standardised weights and the *stepwise* procedure, we obtain a list of the most significant predictors of response propensity. The RHGs are created using all 2^k combinations of a selected set of k predictors. This method of creating RHGs is often referred to as the symmetrical approach. Note that some constraints can be added when creating the RHGs. For example, a minimum group size of 30 and a minimum weighted response rate (RR) of 50% in each RHG can be required. This way we obtain $2^k - M$ *valid combinations*, where M is the reduction caused by collapsing of RHGs when the minimum requirements are not met. Kalton and Kasprzyk (1986) suggest these kinds of constraints to avoid a large variance in the weights and a possible loss of precision in the survey estimates.

4.1.2 Segmentation modelling approach

Another method that can be used to form the RHGs is SM, based on the CHAID (Chi-square Automatic Interaction Detection) algorithm. This method splits the sample into smaller subgroups based on their response rates. The splitting process continues until no more statistically significant predictors can be found. The final subgroups become the RHGs, in which our nonresponse adjustments will be calculated. Note that all minimum requirements (example: n>30, RR>50%) have to be met in each RHG. SM method is often referred to as the nonsymmetrical approach. It is important to mention that changing the significance level of the tests will lead to a different number of statistically significant predictors and thus to a change in the total number of RHGs created.

4.1.3 Nonresponse adjustment factors calculation

The RHGs can be created using either the LR method as described in 4.1.1 or the SM method as described in 4.1.2. Since a uniform response mechanism is assumed in each RHG, the nonresponse adjustment factor is simply given by the inverse of the weighted response rate in the RHG with w_{0k} being the weight used.

5. EMPIRICAL COMPARISONS

To compare the efficiency of the two variable selection methods, a simulation study was done using data from SLID, a longitudinal rotating panel survey selected with a complex design. This survey was conducted for the first time in 1994 with a sample size of 15,000 households, including approximately 31,000 adults (Lavigne and Michaud, 1998). SLID follows the same respondents for six years. We have also analysed data from NLSCY, another Statistics Canada longitudinal survey. The first NLSCY cycle of data collection took place from November 1994 to June 1995. The initial sample for this first wave included some 29,000 children from 0 to 11 years of age (Michaud, Morin, Clermont and Laflamme, 1998). Data collection will be repeated at two-year intervals. The children originally surveyed in the first wave will be followed over time until adulthood.

5.1 Description of the empirical study

As a first step, the probability of response was estimated for every unit in SLID's sample. This was done using a very large number of RHGs (without any constraints on the RHGs' size or RR) that included both SLID's respondents and nonrespondents. Then, along with their estimated probability of response, respondents were kept as our reference sample for the simulation. The size of this sample was about 25,000 persons. From our reference sample, nonresponse was generated using a Poisson sampling procedure. This process was repeated 100 times to create as many respondent and nonrespondent sets; the resulting response rate was 90% on average. For all 100 repetitions, the LR approach was used to create RHGs. However, for the SM approach, RHGs were created only for the first 20 repetitions since manual intervention and usage of a specific software package (such as Knowledge Seeker) are required. Different variations of these variable selection methods were studied:

a) LR_i where i indicates the approximate average number of RHGs generated with the method (i=4, 16, 40, 60). For example, LR_40 means that LR was used to identify the k=6 most important predictors of response. RHGs were then created using all the valid combinations of these k=6 predictors. In our case, because of additional constraints, M was equal to 24. Therefore, a total of $2^k - M = 2^6 - 24 = 40$ RHGs were formed. In our simulation study, LR_i with i=4, 16, 40 and 60 RHGs, corresponds to k =2, 4, 6 and 8 predictors respectively. Note that since the number of valid combinations using k predictors can vary from one repetition to another, the value of i is in fact an

average over the 100 repetitions. For each of them, the k most important predictors of response were identified using LR with the stepwise option, RHGs were created from valid combinations based on these predictors, and a set of weights was produced.

b) SM_ i where i indicates the approximate average number of RHGs generated with the method (i=16, 25, 40), i being an average number over the 20 repetitions. For example, SM_16 means that SM was used with a given significance level and 16 RHGs, on average, were created. In our simulation study, SM_i with i=16, 25 and 40 RHGs, corresponds to the following significance levels: 0.0001, 0.0005 and 0.0025, respectively. For each of the 20 repetitions, a set of RHGs was created based on Knowledge Seeker's results and a corresponding set of weights was produced.

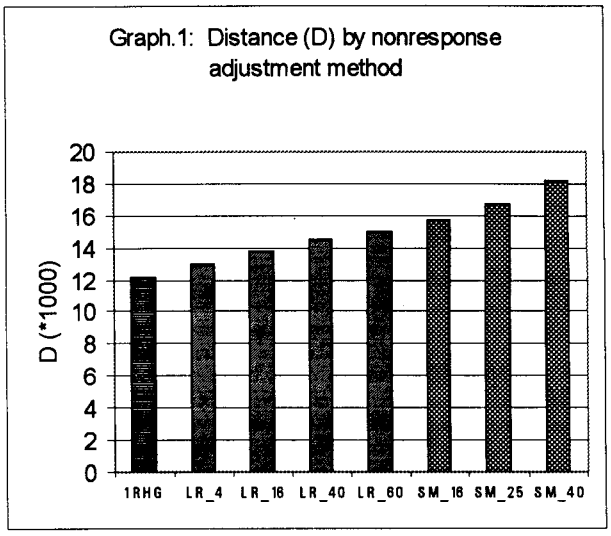
In addition to (a) and (b), a method with only one RHG was used in our simulation study for comparison purposes. This method consisted of defining the whole sample as the one and only RHG. A set of weights was produced, using this particular method, for each of the 100 repetitions. Note that this method is effective only if nonresponse is uniform within the sample.

For each repetition, eight sets of final weights were produced. Each set of weights is the result of two steps: an adjustment for nonresponse (by one of the eight methods mentioned: single RHG, LR_i with i = 4, 16, 40, 60 and SM_i with i = 16, 25, 40) and a poststratification (which is the same for all eight methods).

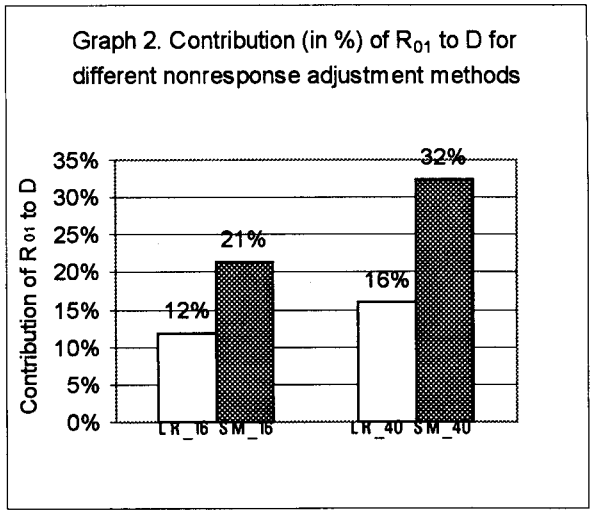
5.2 Results

For each of the eight methods discussed in the previous section, we studied each component of our measure of change D (described in section 3), as well as nonresponse bias and variance of the estimates. Estimates were produced for various variables of interest and for various domains. However, because of space limitations, only estimates of the *total number of persons living in a family whose income is below the low income cutoff for the reference year* are presented in this section. It is important to mention that given the large sample size, the low nonresponse rate (10%) and the fact that a large number of control totals were used in the poststratification, the relative bias of any of the studied methods is very small (less than 1%) when the whole population is the domain of interest. However, the relative bias can quickly increase for smaller domains.

The measure of change D (the total change the weights go through, from the initial weights to the final ones) was calculated for each set of weights. The average of D over the repetitions, is presented for each of the eight studied methods in Graph 1. The graph shows that,



within a method, as the number of RHGs increases, the measure of change, D, also increases. Also, the overall change is higher for SM than for LR. Given that, for a given sample, the term G is constant, the term R_{int} is close to zero and the term R_{12} is approximately constant, it is clear that variations in D are mainly due to variations in the R_{01} term.

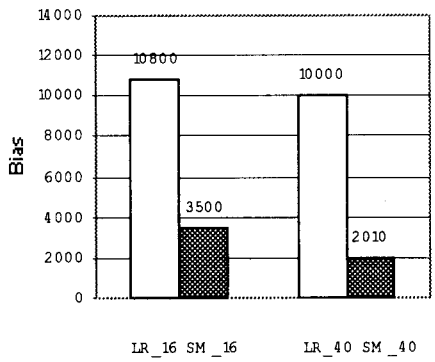


Graph 2 shows that the R_{01} term accounts for a greater percentage of the total distance D in the SM method than in the LR method, for a given number of RHGs. For example, for SM_40, R_{01} contributes to 32% of D, compared to 16% for LR_40.

This means that the individual weight changes, apart from the average change, between the initial step and

the intermediate one (which, in the present case, is the nonresponse adjustment) are greater for SM than for LR. This leads us to expect the SM method to be more effective in reducing nonresponse bias, for a fixed number of RHGs. This expectation is confirmed by Graph 3, which shows that the nonresponse bias (for our variable of interest) is smaller for the SM method than for the LR one.

Graph. 3: Bias comparison of nonresponse adjustment methods



Variance estimates were also produced by a jackknife procedure for each of the point estimates. The average, over the 100 repetitions, of the variance estimates are approximately the same for each method studied. However, there is a slight decrease as the number of RHGs increases, within both LR and SM methods. In addition, variance estimates for SM appear slightly smaller than those for LR.

It is of interest to mention that nonresponse bias has also been studied for other variables of interest, as well as for various domains, through our simulation study using SLID data. In a large proportion of the cases, SM performed better than LR in creating efficient RHGs. The same conclusions were also found using production data from NLSCY.

6. CONCLUSION AND FUTURE WORK

This paper points out that the choice of RHGs and of the method to define them depends on the following: i) the available auxiliary information; ii) the desire to reduce nonresponse bias for all survey estimates; and iii) time and operational constraints. The empirical study shows that SM is better than LR in creating efficient RHGs and, consequently in reducing nonresponse bias. The results also indicate that the proposed measure of change, D, can be a very useful tool for comparing different weighting strategies. In particular, it seems that the larger the R_{01} term, the greater the biasreduction obtained by using the corresponding RHGs as nonresponse adjustment groups. Since it is not usually possible to get precise estimates of the nonresponse bias in a sample survey, the relationship that we have found between the size of the easily computed component R_{01} and the bias reduction suggests to use R_{01} as a tool for nonresponse treatment as follows : compute R_{01} for alternative sets of RHGs; the set with the largest value of R_{01} has the potential to be more effective than the other alternatives for reducing nonresponse bias, for most variables of interest.

Future work includes carrying out further empirical studies in order to corroborate and extend the results presented in this paper. The study reported in this paper could, for example, be repeated with other variables of interest, various nonresponse rates and a larger number of repetitions. It would also be interesting to study other Statistics Canada's longitudinal surveys, both economic and social.

ACKNOWLEDGEMENTS

The authors would like to thank M. Hladky, M. Latouche, C. Nadeau and N. Tremblay for their valuable contribution to this project, as well as the many methodologists who supplied their support during the the simulation study. They would also like to thank Douglas Yeo and Wesley Yung for their comments that helped to improve the quality of the paper.

REFERENCES

Deville, J.C. and Särndal, C. E. (1992). Calibration estimators and generalized raking techniques in survey sampling. Journal of the American Statistical Association.

Kalton, G. and Kasprzyk, D. (1986). The Treatment of Missing Survey Data in Survey Methodology, 1986, Vol. 12, No. 1, 1-16.

Lavigne, M. and Michaud, S. (1998). General aspects of the Survey of Labour and Income Dynamics, SLID Research Document, Statistics Canada, catalog 98-05.

Michaud, S., Morin, Y., Clermont, Y., and Laflamme, G. (1998). Issues in the design of a survey to measure child development : The Experience of the Canadian National Longitudinal Survey of Children and Youth. To be published : ASA 1998 Proceedings.