ON REGRESSION IMPUTATION IN THE PRESENCE OF NONIGNORABLE NONRESPONSE

Jean-François Beaumont, Statistics Canada
Household Survey Methods Division, Statistics Canada, Ottawa (Ontario), Canada K1A 0T6

Key Words: Regression Imputation, Nonignorable Nonresponse, Two-phase Sampling, Taylor Linearization.

1. INTRODUCTION

Regression imputation, that includes ratio imputation as a special case, is a frequent method to reduce nonresponse bias in surveys. In practice, the generalized least squares method is often used to estimate imputation model parameters without weighting by the design weights. However, if a probabilistic response mechanism is assumed and if the response probabilities are known, or can at least be well estimated, using the design weights and the nonresponse adjustement factors lead, for large samples, to approximately unbiased estimates of the population parameters justified by the imputation model. In fact, if the response probabilities are known, we may view the response mechanism as a second phase of selection (Särndal, Swensson and Wretman, 1992, p. 558) and use the two-phase sampling theory to estimate population parameters justified by the imputation model.

The focus of this paper is only on a response mechanism that depends on the variable being imputed, referred to as a nonignorable response mechanism. Such a response mechanism can lead to severe bias in the estimates of the imputation model parameters when the standard generalized least squares method is used. This kind of response mechanism is often dealt with by simultaneously modeling and estimating the response probabilities and the variable of interest (Greenlees, Reece and Zieschang, 1982; Beaumont, 1999, 2000 among others). However, none of these papers addressed the issue of variance estimation, except the former which gave a partial solution to the problem. Although some methods for variance estimation in the presence of imputation are now well known (see Lee, Rancourt and Särndal, 2001 for a review), they usually do not apply or are more difficult to apply in the case of nonignorable nonresponse. The goal of this paper is thus to propose a practical variance estimation approach when response probabilities are estimated and when there is nonignorable nonresponse.

In section 2, regression imputation is reviewed emphasizing the case where response probabilities are not assumed to be uniform. The imputation strategy proposed in Beaumont (1999, 2000) is also developped in greater detail. In section 3, estimation of the population mean as well as variance estimation are discussed. In section 4, the estimation method in the presence of nonignorable nonresponse, developped in Beaumont (1999, 2000), is described. The variance estimation approach proposed in this paper is evaluated through a simulation study using data from the Survey of Labour and Income Dynamics (SLID) of Statistics Canada. The results are presented in section 5. Finally, a brief conclusion is provided in the last section.

2. REGRESSION IMPUTATION

In the following, the objective is to estimate the mean of variable y for a given population U. A sample s is selected from the population and the variable y is only observed for part of s. The sample of respondents is denoted by r and the sample of nonrespondents is denoted by o. It is also assumed that there is a vector of auxiliary variables, $\boldsymbol{x}$, observed for all units in the sample s and correlated with y.

The estimator of the population mean, $\bar{Y} = \sum_{k\in U} y_k / N$, where N is the population size, can be obtained by imputing missing values:

$$\bar{Y}_I^* = \frac{\sum_{k\in s} w_k y_{\bullet k}}{\sum_{k\in s} w_k}, \qquad (2.1)$$

where $w_k = 1/\pi_k$ is the sampling weight for unit k corresponding to the inverse of the selection probability π_k, $y_{\bullet k} = y_k$, for $k\in r$, $y_{\bullet k} = y_k^*$, for $k\in o$, and y_k^* is the imputed value for the nonresponding unit k. In the case of complete response, equation (2.1) yields the usual estimator:

$$\bar{Y}^* = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k}. \qquad (2.2)$$

Regression imputed values are justified by the following model:

$$y_k = \boldsymbol{x}_k' \boldsymbol{\beta} + \varepsilon_k, \qquad (2.3)$$

where $\boldsymbol{\beta}$ is an unknown vector of parameters, ε_k are mutually independent random errors, with zero mean and variance $\sigma^2 \boldsymbol{x}_k' \boldsymbol{\lambda}$, $\boldsymbol{\lambda}$ is a vector of known constants and σ^2 is an unknown parameter. The method of generalized least squares is often used to estimate $\boldsymbol{\beta}$. It consists of solving the following system of equations:

$$\sum_{k\in r} \frac{w_k}{p_k} (y_k - \boldsymbol{x}_k' \boldsymbol{\beta}) \frac{\boldsymbol{x}_k}{\boldsymbol{x}_k' \boldsymbol{\lambda}} = 0, \qquad (2.4)$$

where p_k are the true response probabilities assumed to be greater than zero and known for all units in the sample s. In practice, these response probabilities have to be estimated but, for now, they will be assumed to be known. The solution of (2.4) yields:

$$\boldsymbol{B}^* = \left(\sum_{k \in r} \frac{w_k}{p_k \boldsymbol{x}_k' \boldsymbol{\lambda}} \boldsymbol{x}_k \boldsymbol{x}_k' \right)^{-1} \times \sum_{k \in r} \frac{w_k}{p_k \boldsymbol{x}_k' \boldsymbol{\lambda}} \boldsymbol{x}_k y_k \ . \tag{2.5}$$

It should also be noted that the design weights are included in equations (2.4) and (2.5). Whether they should be included or not is far from obvious (Deville and Särndal, 1994). However, if the response mechanism is viewed as a second phase of selection, the two-phase sampling theory can be used to justify that $\boldsymbol{B}^*$ is a design-response consistent estimator of

$$\boldsymbol{B} = \left(\sum_{k \in U} \frac{1}{\boldsymbol{x}_k' \boldsymbol{\lambda}} \boldsymbol{x}_k \boldsymbol{x}_k' \right)^{-1} \times \sum_{k \in U} \frac{1}{\boldsymbol{x}_k' \boldsymbol{\lambda}} \boldsymbol{x}_k y_k \ ,$$

which is in turn model unbiased for $\boldsymbol{\beta}$. In fact, $\boldsymbol{B}$ is the vector of parameters that we would have obtained, had the entire population been observed. Moreover, we will see in the next section that the two-phase sampling theory can also be used to estimate the population mean as well as the variance of the population mean estimator. If either the design weights or the response probabilities are not taken into account to estimate $\boldsymbol{\beta}$, then we can no longer rely on the two-phase sampling theory and variance estimation becomes more complex. It is for this reason that we have chosen to include the design weights (as well as the response probabilities) in equations (2.4) and (2.5).

Once the vector of unknown parameters has been estimated, missing values are imputed. If the response mechanism is assumed to be uniform, which is a frequent assumption in practice, imputed values can be obtained by the predicted values ($y_k^* = \boldsymbol{x}_k' \boldsymbol{B}^*$). However, for other response mechanisms, this can lead to biased estimates of the population mean, especially when the nonresponse is nonignorable. For example, when nonresponse depends on the variable being imputed, it can be easily shown that the conditional model expectation of y_k given that $k \in o$ is different from the model expectation of y_k (Beaumont, 2000).

Beaumont (1999, 2000) developped an imputation strategy for a nonignorable response mechanism which could be applied more generally to all cases where the nonresponse is not assumed to be uniform and response probabilities are known or can at least be well estimated. The strategy consists of finding the imputed values y_k^* for the nonresponding units such that

$$\sum_{k \in s} w_k \frac{(y_{\bullet k} - \boldsymbol{x}_k' \boldsymbol{B}^*)^2}{\boldsymbol{x}_k' \boldsymbol{\lambda}}$$

is minimized subject to the constraints

$$\sum_{k \in s} w_k (y_{\bullet k} - \boldsymbol{x}_k' \boldsymbol{B}^*) \frac{\boldsymbol{x}_k}{\boldsymbol{x}_k' \boldsymbol{\lambda}} = 0 \ .$$

The rationale behind this imputation strategy is that the preceding constraints would have been respected, had the variable y been observed for all units in the sample s and had this variable been modeled using (2.3). Using Lagrange multipliers and some algebra, it can be shown that

$$y_k^* = \boldsymbol{x}_k' (\boldsymbol{B}^* + \boldsymbol{\Delta}^*) \ , \tag{2.6}$$

where

$$\boldsymbol{\Delta}^* = - \left(\sum_{k \in o} w_k \frac{\boldsymbol{x}_k \boldsymbol{x}_k'}{\boldsymbol{x}_k' \boldsymbol{\lambda}} \right)^{-1} \times \sum_{k \in r} w_k (y_k - \boldsymbol{x}_k' \boldsymbol{B}^*) \frac{\boldsymbol{x}_k}{\boldsymbol{x}_k' \boldsymbol{\lambda}} \ . \tag{2.7}$$

The preceding equation can be viewed as a correction to $\boldsymbol{B}^*$ (and thus to the imputed values) to take into account that the relationship between the variable of interest y and the auxiliary variables $\boldsymbol{x}$ may be different when only nonresponding units are considered as opposed to all units in the sample. It is interesting to note, from equations (2.4) and (2.5), that the second sum in (2.7) is zero when uniform nonresponse (equal response probabilities for all units) is assumed. Thus, the correction (2.7) vanishes for that type of nonresponse and the missing values can simply be imputed by the predicted values. Note that this correction also vanishes when there is no nonresponse since in that case nonresponse can be viewed as uniform with all response probabilities equal to one.

3. ESTIMATION

Since the model error variance is a linear combination of the auxiliary variables ($V_m(\varepsilon_k) = \sigma^2 \boldsymbol{x}_k' \boldsymbol{\lambda}$) and using a proof similar to Särndal, Swensson and Wretman (1992, p.231), it is easy to show that

$$\sum_{k \in s} w_k (y_{\bullet k} - \boldsymbol{x}_k' \boldsymbol{B}^*) = 0 \ . \tag{3.1}$$

Consequently, the estimator of the population mean (2.1) with the imputed values given in (2.6) can be expressed as:

$$\bar{Y}_I^* = \frac{\sum_{k\in s} w_k \boldsymbol{x}_k'}{\sum_{k\in s} w_k} \boldsymbol{B}^* . \qquad (3.2)$$

Using the same argument as in (3.1), it can also be shown that

$$\sum_{k\in r} \frac{w_k}{p_k} (y_k - \boldsymbol{x}_k' \boldsymbol{B}^*) = 0 .$$

Adding the preceding equation to the numerator of (3.2) leads to the following alternative form for $\bar{Y}_I^*$:

$$\bar{Y}_I^* = \frac{1}{\sum_{k\in s} w_k} \left[\sum_{k\in r} \frac{w_k}{p_k} y_k + \left(\sum_{k\in s} w_k \boldsymbol{x}_k' - \sum_{k\in r} \frac{w_k}{p_k} \boldsymbol{x}_k' \right) \boldsymbol{B}^* \right] \qquad (3.3)$$

It is very interesting to note that the part between brackets is exactly of the form of the Generalized REGression (GREG) estimator of a population total under two-phase sampling with auxiliary information available at the first phase level only. Here, the first and the second phase correspond to the sampling mechanism and to the response mechanism respectively. The two-phase sampling theory can thus be borrowed to estimate the variance of $\bar{Y}_I^*$. In fact, we can express the variance of $\bar{Y}_I^*$ as:

$$V(\bar{Y}_I^*) = V_p[E_q(\bar{Y}_I^* \mid s)] + E_p[V_q(\bar{Y}_I^* \mid s)] , \qquad (3.4)$$

where the subscript p represents the sampling mechanism and the subscript q represents the response mechanism. It can easily be shown that, at least for large samples, $E_q(\bar{Y}_I^* \mid s) \approx \bar{Y}^*$, where $\bar{Y}^*$ is defined in (2.2). The first term of the right side of (3.4) can be referred to as the sampling variance, $Vsam = V_p(\bar{Y}^*)$, and the second term can be referred to as the imputation variance, *Vimp*.

The sampling variance component of (3.4) can be estimated through the Taylor linearization technique described in Särndal, Swensson and Wretman (1992, p. 175):

$$Vsam^* = \left(\frac{1}{\sum_{k\in s} w_k} \right)^2 \sum_{k\in s} \sum_{l\in s} \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_{kl}} \times \frac{(y_k - \bar{Y}^*)}{\pi_k} \frac{(y_l - \bar{Y}^*)}{\pi_l} ,$$

where π_{kl} is the joint selection probability of units k and l. The preceding estimator cannot be used in the presence of nonresponse since it depends on unobserved values of the variable y. Therefore, we must estimate the unknown quantities which leads to the following estimator:

$$Vsam_I^* = \left(\frac{1}{\sum_{k\in s} w_k} \right)^2 \sum_{k\in r} \sum_{l\in r} \frac{(\pi_{kl} - \pi_k \pi_l)}{\pi_{kl} p_{kl}} \times \frac{(y_k - \bar{Y}_I^*)}{\pi_k} \frac{(y_l - \bar{Y}_I^*)}{\pi_l} ,$$

where p_{kl} is the joint response probability of units k and l. To simplify, it will be assumed in the following that units in the sample respond independently of each other and thus, $p_{kl} = p_k p_l$, for $k \neq l$, and $p_{kk} = p_k$.

The sampling variance could also be estimated by adding a residual to the imputed values, as in Greenlees, Reece and Zieschang (1982), to take into account that the imputed values are less variable than the true values. Then, standard variance estimation techniques built for the case of complete response could be used with the imputed values replacing the true values. Although this method can be useful to estimate the sampling variance, it is important to note that it does not estimate the imputation variance (Särndal, 1992).

An estimator for $V_q(\bar{Y}_I^* \mid s)$ is required to estimate the imputation variance component of (3.4). The Taylor linearization technique yields:

$$Vimp^* = \left(\frac{1}{\sum_{k\in s} w_k} \right)^2 \sum_{k\in r} w_k^2 \times \frac{(1 - p_k)}{p_k^2} (y_k - \boldsymbol{x}_k' \hat{\boldsymbol{B}})^2 .$$

The variance of $\bar{Y}_I^*$ is thus estimated by:

$$V_1^*(\bar{Y}_I^*) = Vsam_I^* + Vimp^* . \qquad (3.5)$$

However, the following alternative estimator has been empirically found to be slightly more stable:

$$V_2^*(\bar{Y}_I^*) = \left(\frac{\sum_{k\in s} w_k}{\sum_{k\in r} w_k / p_k} \right)^2 V_1^*(\bar{Y}_I^*) . \qquad (3.6)$$

One reason justifying estimator (3.6) over (3.5) happens when one or more response probabilities (or estimated response probabilities) is very small. In that case, (3.5) may become very large. In (3.6), however, the large component in the numerator, $V_1^*(\bar{Y}_I^*)$, will be compensated to some extent by the large component in the denominator, $\sum_{k\in r} w_k / p_k$. We can also justify (3.6) by using similar arguments to those making (2.2) preferred over $\bar{Y}^{**} = \sum_{k\in s} w_k y_k / N$ as a population mean

estimator (see Särndal, Swensson and Wretman, 1992, p. 183).

It may also happen that some auxiliary variables are available at the population level, a case which has not been discussed in this paper. However, if the population mean estimator as well as its variance estimator are available for the case of complete response then not much complexity is added in the presence of imputation and we can still rely on the two-phase sampling theory. For a more thorough discussion of two-phase sampling, the reader is referred to Hidiroglou and Särndal (1998).

4. NONIGNORABLE NONRESPONSE

In this section, the case where the response probabilities depend on the variable being imputed, which is a nonignorable response mechanism, is considered. Such a response mechanism is often dealt with by simultaneously modeling and estimating the response probabilities and the variable of interest y. The response probability for unit k can be modeled by some function $p_k(y_k, \boldsymbol{z}_k; \boldsymbol{\alpha})$, where $\boldsymbol{z}_k$ is a vector of auxiliary variables known for all units in the sample and $\boldsymbol{\alpha}$ is a vector of unknown parameters. The unknown vectors of parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be estimated using the maximum likelihood method or the robust estimation method described in Beaumont (1999, 2000).

The maximum likelihood method requires that model (2.3) be appropriate and that errors be normally distributed. According to two simulation studies (Beaumont, 1999, 2000), when one or both assumptions is violated, it is preferable to use the more robust estimation method described below.

If the response probabilities were known and greater than zero for all units in the sample, we could use estimating equations (2.4) to estimate $\boldsymbol{\beta}$. On the other hand, if the conditional distribution of y_k given $\boldsymbol{x}_k$ were known, we could estimate $\boldsymbol{\alpha}$ by solving estimating equations resulting from the maximum likelihood method (given that the sampling mechanism is ignorable; otherwise the pseudo-maximum likelihood method may be preferable). We obtain estimates for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by solving simultaneously both systems of estimating equations. However, this requires to work out the unknown expectation $\mathrm{E}_m[p_k(y_k, \boldsymbol{z}_k; \boldsymbol{\alpha}) \mid \boldsymbol{x}_k]$ which can be approximated by $p_k(\mathrm{E}_m[y_k \mid \boldsymbol{x}_k], \boldsymbol{z}_k; \boldsymbol{\alpha})$, where the subscript m indicates that expectations are evaluated under the model (2.3). This method is considered robust to a departure from the normality assumption because, as opposed to the maximum likelihood method, it does not require to specify the distribution of the variable y. It has also been empirically shown to be robust to a departure from the assumed model (2.3).

An interesting property of the robust estimation method described above is that estimates of the unknown parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be obtained by using the following algorithm:

1. Set initial values for the response probabilities (or for the vector of parameters $\boldsymbol{\alpha}$). For example set $p_k^{(0)} = 1$ for all responding units;
2. Set $j = 1$, where j indicates the iteration number;
3. Solve (2.4) with current values of the response probabilities, $p_k^{(j-1)}$, using a weighted linear regression procedure to obtain $\boldsymbol{\beta}^{(j)}$;
4. Impute missing values by $y_k^{(j)} = \boldsymbol{x}_k' \boldsymbol{\beta}^{(j)}$, for $k \in o$;
5. Solve the maximum likelihood equations to obtain new response probabilities $p_k^{(j)}$;
6. Stop if convergence is reached, otherwise set $j = j + 1$ and return to step 3.

In practice, this algorithm has been found to always find the solution when it exists. However, it should be noted that in our simulation study (see section 5), there were no solution for 17 out of the 5000 samples. For these samples, the algorithm alternated between two sets of values. When this happened, we chose randomly one of the two sets of values as the final solution. This situation tends to occur when the response probabilities are modeled by forming categories, especially if the number of categories is high and the number of respondents is small within each category. The algorithm usually converged in four to eight iterations in our simulation study. In two previous simulation studies, the response probabilities were rather modeled through a continuous function of an auxiliary variable and the algorithm always found the solution, although it took many iterations to reach convergence in few cases.

Once the algorithm has converged, we can use the theory of sections 2 and 3 to estimate the population mean as well as the variance of this estimate. However, in the preceding two sections, we have assumed that the response probabilities were known, which is not the case in this section. So, the response probability model should be appropriate and the estimated response probabilities should not be too much unstable in order to be able to use the two-phase sampling theory. Note that the variability of the estimated response probabilities is not taken into account in variance estimation. However, we will see in the next section that, for a relatively small sample size, this does not seem to be a too serious problem.

5. SIMULATION STUDY

In order to evaluate the approach proposed in this paper,

we performed a simulation study. We used the data from the 1997 Survey of Labour and Income Dynamics (SLID) of Statistics Canada to obtain a population. We selected all the people in the province of Alberta who are between the ages of 20 to 40 years old inclusively, who did not have a missing value for the variable wages-and-salaries (our variable of interest y) and who also did not have a missing value for the previous year wages-and-salaries (our auxiliary variable x). This resulted in a population of 654 people.

From this population, 5000 samples of size 350 were selected using simple random sampling without replacement. For each of the selected samples, nonresponse was generated, such that each person responds independently of one another with the following response probability:

$$p_k = \delta + \frac{(1 - \delta)}{1 + \exp(-\alpha_0 - \alpha_1 y_k)}, \qquad (5.1)$$

where $\delta = 0.2$, $\alpha_0 = 4$ and $\alpha_1 = -0.000125$. These parameters have been chosen to ensure that the mean overall response rate be approximately 70% and that the lowest response probability possible be 20%. The reason for the latter restriction is to avoid large nonresponse adjustment factors $1/p_k$ which yield very unstable variance estimates. Note that this is a nonignorable response mechanism, where people with a high value of the variable y have less tendency to be respondents than those with a low value of that variable. This kind of response mechanism may be realistic for such a sensitive variable.

To impute, we assumed a simple linear regression model with nonzero intercept and constant variance, which seems reasonable with the data at hand. The population squared coefficient of correlation between x and y is approximately 73%.

The estimator (3.2) (or equivalently, 3.3) is used to estimate the population mean and the estimator (3.6) is chosen for variance estimation. To estimate the unknown vector of parameters $\boldsymbol{\alpha}$ (and then the response probabilities), we considered 4 assumptions: UNIF, C6X, ROB_C6Y and KNOWN. The UNIF assumption corresponds to a uniform response mechanism where each unit in the sample has the same response probability. This assumption is included in the study for evaluation purposes and also because uniform nonresponse is frequently used in practice. A better alternative to UNIF is C6X, which divides the variable x into 6 predetermined categories and assumes uniform nonresponse within each category. An even better assumption is obtained by dividing the variable y into 6 predetermined categories and assuming uniform nonresponse within each category. This assumption is combined with the robust estimation method described in section 4 and will be denoted by ROB_C6Y. Note that the assumed response model is still not the same as the true response model (5.1), as it is most likely the case in practice. However, we will see that ROB_C6Y provides a big improvement over the simplest assumptions UNIF and C6X. Finally, we have also considered the ideal and unrealistic case for which the response probabilities are known, denoted by KNOWN.

For each of the 5000 samples of respondents, four population mean estimates and four variance estimates have been obtained. Now, let us assume that m_k^* and v_k^* are respectively the population mean estimate and the variance estimate for the k^{th} sample of respondents resulting from one of the four assumptions considered above and that m and v are the true population mean and the true variance respectively. Note that the true variance has been estimated by: $v = \sqrt{\sum_k (m_k^* - \bar{m}^*)^2 / 4999}$, where $\bar{m}^*$ is the average of the 5000 m_k^* estimates. The relative bias in percentage of a population mean estimator can be estimated by:

$$\text{RB}^* = \frac{\sum_{k=1}^{5000} (m_k^* - m)}{5000} \times \frac{1}{m} \times 100\% .$$

An estimate of the standard error of this relative bias can be given by:

$$\text{SE}^* = \frac{100}{m} \sqrt{\frac{s_{m^*}^2}{5000}},$$

where $s_{m^*}^2$ is the variance of the 5000 m_k^* estimates. Finally, an estimate of the relative root mean squared error in percentage can be expressed as:

$$\text{RRMSE}^* = \sqrt{\frac{\sum_{k=1}^{5000} (m_k^* - m)^2}{5000}} \times \frac{1}{m} \times 100\% .$$

In a similar way, we can also estimate the relative bias, the standard error of the relative bias and the relative root mean squared error of a variance estimator by replacing m_k^* by v_k^* and m by v in the preceding three equations. An estimate of the coverage rate (COVR*) in percentage has finally been calculated by taking the proportion of the 5000 samples of respondents for which the true value m was inside the interval $[m_k^* - 1.96\sqrt{v_k^*}, m_k^* + 1.96\sqrt{v_k^*}]$.

Table 1 shows the results of the simulation study. It can first be observed that ROB_C6Y is much better than

UNIF or C6X when estimating the population mean in terms of RB^* and $RRMSE^*$. Indeed, ROB_C6Y is almost as good as the ideal case KNOWN even if the response probability model is not exactly the same as the true one.

When estimating variance, we see that even the ideal case has a negative bias (and therefore too low $COVR^*$). This is not surprising since the Taylor linearization technique is well known to underestimate the variance (Särndal, Swensson and Wretman, 1992, p. 176), especially for small samples. Note that the RB^* of C6X is relatively low and positive, which is more difficult to explain. Concerning the $RRMSE^*$, the same conclusion as for the population mean estimators can be drawn.

What is more interesting is the analysis of $COVR^*$. A huge improvement can be obtained by using ROB_C6Y over the more naive assumptions UNIF and C6X, which lead to much too low $COVR^*$s. If we could add a constant to the ROB_C6Y population mean estimator such that it is unbiased, then $COVR^*$ would be 91.7%, just about 1% below the ideal case. The emphasis should thus be put on the bias of the point estimators first when choosing an estimation strategy.

6. CONCLUSION

For sensitive variables, such as income, assuming a nonignorable response mechanism that depends on the variable being imputed may be more realistic in some practical cases than the usual assumptions of uniform or ignorable response mechanism (for example, a response mechanism that depends on one or more auxiliary variables of the imputation model). If this assumption is true, the usual least squares method combined with the usual assumptions about the response mechanism will lead to bias in point estimates and poor confidence intervals. The approach proposed in this paper can thus be a useful alternative in this situation to reduce bias and obtain better confidence intervals.

ACKNOWLEDGEMENTS

I would like to thank Mike Hidiroglou, David Haziza and Eric Rancourt of Statistics Canada for their useful comments.

REFERENCES

Beaumont, J.-F. (1999). A robust estimation method in the presence of nonignorable nonresponse. *Proceedings of the Section on Survey Research Methods, American Statistical Association* (to appear).

Beaumont, J.-F. (2000). An estimation method in the presence of nonignorable nonresponse. *Survey Methodology* (to appear).

Deville, J.-C., and Särndal, C.-E. (1994). Variance estimation for the regression imputed Horvitz-Thompson estimator. *Journal of Official Statistics*, 10, 381-394.

Greenlees, J.S., Reece, W.S., and Zieschang, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, 77, 251-261.

Hidiroglou, M.A., and Särndal, C.-E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, 24, 11-20.

Lee, H., Rancourt, E., and Särndal, C.-E. (2001). Variance estimation from survey data under single value imputation. In *Survey Nonresponse*, Groves, R., Dillman, D., Eltinge, J., and Little, R. (editors), Chapter 21, New-York, John Wiley & Sons, Inc. (to appear)

Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18, 241-252.

Särndal, C.-E., Swensson, B., and Wretman, J.H. (1992). *Model Assisted Survey Sampling*. New-York, Springer-Verlag.

TABLE 1: RESULTS OF THE SIMULATION STUDY

ASSUMP-TIONS	POPULATION MEAN			VARIANCE			
	RB^*(%)	SE^*	$RRMSE^*$(%)	RB^*(%)	SE^*	$RRMSE^*$(%)	$COVR^*$(%)
UNIF	-10.3	0.05	11.0	-56.1	0.08	56.4	7.9
C6X	-9.6	0.06	10.4	5.7	0.63	44.9	34.1
ROB_C6Y	-1.5	0.05	4.1	-19.9	0.39	34.3	87.3
KNOWN	-0.1	0.05	3.9	-5.4	0.40	28.9	92.9