# EXPERIMENTS WITH VARIANCE ESTIMATION FROM SURVEY DATA WITH IMPUTED VALUES

**Hyunshik Lee and Eric Rancourt, Statistics Canada**
**Carl E. Särndal, University of Montreal**
**Hyunshik Lee, 11-M, R.H. Coats Bldg., Tunney's Pasture, Ottawa, Ontario K1A 0T6**

Key Words: Model-Assisted Approach, Regression Imputation, Multiple Imputation, Monte Carlo Study.

## 1. SUMMARY

In this report, we describe the methodology and the results of a Monte Carlo study of different variance estimators intended for six different methods of imputation.

The imputation methods considered in the study were:
1). Single imputation by regression (REG)
2). Single imputation by regression with added residual (REGRES)
3). Single imputation by regression with added standardized residual (REGRESST)
4). Single imputation by nearest neighbor (NN)
5). Multiple imputation by regression with added residual MULTREG)
6). Multiple imputation by nearest neighbor (MULTNN)

We used $M = 2$ repetitions for the multiple imputation methods 5 and 6. For each imputation method, we evaluated one or more variance estimators. A total of 10 variance estimators were included in the study.

The simulations were carried out with 12 different populations representing a variety of relationships between x (the auxiliary variable used in the imputation) and y (the study variable). For each population, three different response mechanisms were used, leading to a total of $12 \times 3 = 36$ different cases. The objective was to identify variance estimators that perform reasonably well under a variety of conditions. Ideal performance under all possible circumstances seems impossible to attain.

Some of the main conclusions are:
1). Concerning the point estimators corresponding to the six imputation methods: All imputation methods have a tolerable bias if the nonresponse is ignorable (that is, when the nonresponse occurs at random for given x; the precise definition is given in Rubin (1976)). However, all of the methods lead to a fairly substantial bias when the nonresponse is non-ignorable (that is, when the probability of nonresponse is systematically related to the variable of interest). Nearest neighbor imputation tends to produce a greater bias than regression imputation.
2). Concerning the variance estimators: None of the 10 variance estimators included in our study comes close to yielding unbiased estimates in all 36 cases. However, out of the ten variance estimators that we tested, there are a few whose overall performance can be termed acceptable. Their bias is fairly limited in all or most of the 36 cases, and they typically alternate between a mild overestimation and a mild underestimation. These methods, defined in detail in Section 2, are: REGRES-SARN, REG-RAO1 and REG-RAO2 for single regression imputation; NN-SARN for single nearest neighbor imputation; the multiple imputation variance estimators MULTREG for multiple regression imputation and MULTNN for multiple nearest neighbor imputation. Some of the variance estimators we examined may work very well under the particular conditions for which they were designed. For instance, the methods REG-RAO1 and REG-RAO2 (for single regression imputation) perform very well when the nonresponse is ignorable. The multiple imputation variance estimators are more variable than the other alternatives; consequently, the confidence intervals calculated with these methods have a more unpredictable length. This disadvantage is in addition to the heavy calculations caused by two or more imputations.

Our study shows the difficulty of identifying variance estimators that have impeccable behavior under a variety of conditions. Our study also emphasizes that the variance estimators based on "standard formulas" must not be used. The standard estimators are based on an usually invalid assumption that imputed values have the same quality as observed values. These estimators lead to a considerable underestimation of the variance.

## 2. THE SIX IMPUTATION METHODS AND THE CORRESPONDING VARIANCE ESTIMATORS

The objective is to estimate the mean $\bar{y}_U = (1/N)\sum_U y_k$ of the finite population $U = \{1, \ldots, k, \ldots, N\}$. A simple random sample without replacement (SRSWOR), s, of size n is drawn from U. Denote by r the set of responding units; let m be the size of r. The nonresponse set is $s - r$; its size is $n - m$. For every unit $k \in r$, the value $y_k$ is observed. However, for the units $k \in s - r$, the $y_k$-values are missing, and imputed values are derived with a specified imputation method. The six imputation methods studied in this paper are defined in the following.

The imputation leads to a completed data set, called the data after imputation. This data set is denoted as $\{y_{\cdot k} : k \in s\}$, where $y_{\cdot k}$ equals the observed value $y_k$ if k is a responding units, that is, if $k \in r$, and $y_{\cdot k}$ equals the imputed value if k is a nonresponding unit, that is, if $k \in s - r$.

The point estimator for the single imputation methods REG, NN, REGRES and REGRESST consists simply of the mean of the data after imputation, $\bar{y}_{\cdot s} = (1/n)\sum_s y_{\cdot k}$. The estimator formula is thus the same as the one that would be used in the case of 100% response. In other words, there is an implicit assumption that a negligible bias is caused by replacing missing data by imputations. This assumption is often violated, particularly when the nonresponse is nonignorable.

In the multiple imputation methods MULTREG and MULTNN, the point estimator is calculated as the average of the means of two sets of data after imputation. In the multiple imputation methods, too, the bias of the point estimator is considerable when the nonresponse mechanism is nonignorable.

When the nonresponse is nonignorable, any available information on the characteristics of nonrespondents should be used to reduce the bias of the point estimator of the population mean. However, we did not try to do this because our primary objective was to see how various variance estimators perform under a variety of conditions.

In all the methods studied, the imputation was carried out with the aid of an auxiliary variable, x. We assume that $x_k$, the value of x for the unit k, is positive and known for every unit $k \in s$.

1). Imputation Method REG. (Single regression imputation): If the unit k requires imputation, the value $\hat{B}x_k$ is imputed, where $\hat{B} = (\sum_r y_k)/(\sum_r x_k)$. (The method is also referred to as ratio imputation.) The data after imputation are therefore

$$y_{\cdot k} = \begin{cases} y_k, & \text{if } k \in r \\ \hat{B}x_k, & \text{if } k \in s - r. \end{cases}$$

The point estimator in this method becomes $\bar{y}_{\cdot s} = (1/n)\sum_s y_{\cdot k} = \bar{x}_s \bar{y}_r / \bar{x}_r$ where $\bar{x}_s = (1/n)\sum_s x_k$, $\bar{y}_r = (1/m)\sum_r y_k$ and $\bar{x}_r = (1/m)\sum_r x_k$. Four different variance estimators are considered with this imputation method:

REG-ORD. This method uses the ordinary variance estimator formula, but computed using the data after imputation, that is, $\hat{V} = (1/n - 1/N)S^2_{y \cdot s}$, where $S^2_{y \cdot s} = \sum_s (y_{\cdot k} - \bar{y}_{\cdot s})^2/(n-1)$. The method is known to underestimate the real variance and is included in the study only to assess the underestimation caused

by acting as if imputed data are as good as actual data.

REG-SARN. This model-assisted variance estimator, derived in Särndal (1990) and also in Deville and Särndal (1991) for more general cases, is given by

$$\hat{V} = \left(\frac{1}{n} - \frac{1}{N}\right)\{S^2_{y\cdot s} + C_0\hat{\sigma}^2\} + \left(\frac{1}{m} - \frac{1}{n}\right)C_1\hat{\sigma}^2$$

where

$$C_0 = \frac{1}{n-1}\left(\sum_{s-r} x_k - \frac{\sum_{s-r} x_k^2}{\sum_r x_k} + \frac{1}{n}\frac{\sum_{s-r} x_k \sum_s x_k}{\sum_r x_k}\right)$$

$$C_1 = \frac{\bar{x}_s \bar{x}_{s-r}}{\bar{x}_r},$$

$$\hat{\sigma}^2 = \frac{\sum_r e_k^2/(m-1)}{\bar{x}_r\{1-(\mathrm{cv}_{xr})^2/m\}},$$

with $\bar{x}_{s-r} = \sum_{s-r} x_k/(n-m)$, $e_k = y_k - \hat{B}x_k$ and $\mathrm{cv}_{xr} = S_{xr}/\bar{x}_r$, which is the coefficient of variation of x in the response set r. This variance estimator is based on the model $\xi$ stating that $y_k = \beta x_k + \epsilon_k$, for $k = 1, \ldots, N$, where $E_\xi(\epsilon_k) = 0$, $V_\xi(\epsilon_k) = \sigma^2 x_k$ and the model errors $\epsilon_k$ are independent. The REG-SARN variance estimator is therefore expected to perform particularly well when the finite population scatter $(y_k, x_k)$ agrees closely with this model. For many practical purposes, $C_0 \doteq (1 - m/n)\bar{x}_{s-r}$ and $\hat{\sigma}^2 \doteq \sum_r e_k^2/\sum_r x_k$ are good approximations of the more cumbersome exact expressions when m is large.

REG-RAO1. This variance estimator is justified by a two-phase sampling argument and was suggested by Rao (1990). It is given by

$$\hat{V} = \left(\frac{1}{n} - \frac{1}{N}\right)S^2_{yr} + \left(\frac{1}{m} - \frac{1}{n}\right)S^2_{er}$$

where $S^2_{yr} = \sum_r (y_k - \bar{y}_r)^2/(m-1)$ and $S^2_{er} = \sum_r e_k^2/(m-1)$.

REG-RAO2. This variance estimator, also suggested by Rao (1990), is given by

$$\hat{V} = \left(\frac{1}{n} - \frac{1}{N}\right)\hat{B}^2 S^2_{xs} + 2\left(\frac{1}{n} - \frac{1}{N}\right)\hat{B}S_{xer} + \left(\frac{1}{m} - \frac{1}{N}\right)S^2_{er}$$

where $S_{xer} = \sum_r e_k x_k/(m-1)$.

2). Imputation Method REGRES. (Single imputation by regression with added residual): If unit k requires imputation, the value $\hat{B}x_k + e_k^*$ is imputed, where $e_k^*$ is selected with SRSWR from the set of residuals $\{e_k = y_k - \hat{B}x_k; k \in r\}$. The data after imputation are then

$$y_{\cdot k} = \begin{cases} y_k, & \text{if } k \in r \\ \hat{B}x_k + e_k^*, & \text{if } k \in s-r. \end{cases}$$

For this method, the only variance estimator included in the study was:

REGRES-ORD. This method consists of the ordinary variance estimator, $\hat{V} = (1/n - 1/N)S^2_{y\cdot s}$, where $S^2_{y\cdot s} = \sum_s (y_{\cdot k} - \bar{y}_{\cdot s})^2/(n-1)$ is the variance of the data after imputation.

3). Imputation Method REGRESST (Single imputation by regression with added standardized residual): If unit k requires imputation, the imputed value is $\hat{B}x_k + e_k^*$, where $e_k^*$ is obtained by the following procedure: First, calculate a supply of m standardized residuals, $e_k^p = e_k/\sqrt{x_k}$, $k \in r$; then for every $k \in s-r$, calculate $e_k^o = \sqrt{x_k}\, e_k^{p*}$, where $e_k^{p*}$ is randomly drawn with SRSWR from the supply; finally, calculate $e_k^* = e_k^o - \sum_{s-r} e_k^o/(n-m)$. The $e_k^*$ are thereby centered to have zero mean. The resulting data after imputation are

$$y_{\cdot k} = \begin{cases} y_k, & \text{if } k \in r \\ \hat{B}x_k + e_k^*, & \text{if } k \in s-r. \end{cases}$$

The only variance tried for this method was:

REGRESST-ORD. This method uses the ordinary variance estimator, $(1/n - 1/N)S^2_{y\cdot s}$ computed on the data after imputation.

4). Imputation Method NN. (Single nearest neighbor imputation): If the unit k requires imputation, the imputed value, $y_{NNk}$, equals the y-value of a donor unit that is as close as possible to k, as measured by the x-variable. More specifically, the donor unit is the one for which the distance $|x_k - x_l|$ is minimum among all potential donors $l$ such that $l \in r$, $l \neq k$. The data after imputation are

$$y_{\cdot k} = \begin{cases} y_k, & \text{if } k \in r \\ y_{NNk}, & \text{if } k \in s-r. \end{cases}$$

Two variance estimators were used with this imputation method:

NN-ORD. Ordinary variance estimator, $\hat{V} = (1/n - 1/N)S^2_{y\cdot s}$ computed on the data after imputation.

NN-SARN. This variance estimator is given by the formula for $\hat{V}$ as in the REG-SARN method, but $S^2_{y\cdot s}$ is computed using the data with nearest neighbor imputation value $y_{NNk}$ (instead of $\hat{B}x_k$) for $k \in s-r$, whereas $\hat{\sigma}^2$ is computed with the aid of the residuals $e_k = y_k - \hat{B}x_k$, for $k \in r$. (Thus the residuals $y_k - y_{NNk}$ are not used because the use of them leads to overestimation of $\sigma^2$.)

5). Imputation Method MULTREG. (Multiple imputation by regression): Assuming that $y_k \sim N(\beta x_k, \sigma^2 x_k)$ and nonresponse is ignorable, the multiple imputation is carried out as follows. First, $\beta$ and $\sigma^2$ are estimated, respectively, by $\hat{B}$ and

$$\hat{\sigma}^2 = \frac{1}{m-1}\sum_r \frac{(y_k - \hat{B}x_k)^2}{x_k}.$$

Then, for each $i = 1, \ldots, M$ ($M = 2$), perform the following steps:

Step 1: Draw a $\chi^2$ random variate with $(m-1)$ degrees of freedom, say $g$, and let $\sigma_{i*}^2 = \hat{\sigma}^2(m-1)/g$;

Step 2: Draw a $N(0,1)$ random variate, say $z$, and let $\beta_{i*} = \hat{B} + \sigma_{i*} z(\sum_r x_k)^{-1/2}$;

Step 3: For each $k \in s-r$, draw a $N(0,1)$ variate independently, say $u$, and let $e_{ik}^* = u\sqrt{x_k}\,\sigma_{i*}$.

Two data sets after imputation thus obtained are,

$$y_{\cdot 1k} = \begin{cases} y_k, & \text{if } k \in r \\ \beta_{1*}x_k + e_{1k}^*, & \text{if } k \in s-r \end{cases}$$

and

$$y_{\cdot 2k} = \begin{cases} y_k, & \text{if } k \in r \\ \beta_{2*}x_k + e_{2k}^*, & \text{if } k \in s-r \end{cases}$$

A modification to the above procedure for a nonnormal case is to replace Step 3 by the following:

Step 3': For each $k \in s-r$, draw a number, say $w_k$, with replacement from the set of standardized residuals $\{(y_l - \hat{B}x_l)/\sqrt{(1-1/m)x_l\hat{\sigma}^2}; l \in r\}$ and let $e_{ik}^* = w_k\sqrt{x_k}\,\sigma_{i*}$.

For more detail, see Rubin (1987, pp.166-168). We tried both but the results for the latter method are reported here.

The point estimator of the population mean is $\bar{y}_{\cdot\cdot s} = (\bar{y}_{\cdot 1s} + \bar{y}_{\cdot 2s})/2$, where $\bar{y}_{\cdot 1s}$ and $\bar{y}_{\cdot 2s}$ are the means of the data sets $\{y_{\cdot 1k}: k \in s\}$ and $\{y_{\cdot 2k}: k \in s\}$, respectively. The corresponding variance estimator, suggested by Rubin (1986), is

$$\hat{V} = \frac{1}{M}\sum_{j=1}^{M}\left(\frac{1}{n} - \frac{1}{N}\right)S^2_{y\cdot js} + \left(1 + \frac{1}{M}\right)\left(\frac{1}{M-1}\right)\sum_{j=1}^{M}(\bar{y}_{\cdot js} - \bar{y}_{\cdot\cdot s})^2$$

with $M = 2$ where $S^2_{y\cdot js} = \sum_s (y_{\cdot jk} - \bar{y}_{\cdot js})^2/(n-1)$ is the variance calculated from the j-th completed data set $\{y_{\cdot jk}: k \in s\}$.

6). Imputation Method MULTNN. (Multiple imputation by nearest neighbor): For each nonrespondent, the two nearest neighbors are selected by the NN method based on the x-value and one of their y-values is randomly picked and assigned to the

first imputation, $y_{NN1k}$ and the remaining y-value is assigned to the second imputation, $y_{NN2k}$. Two data sets after imputation are thus obtained, namely,

$$y_{\cdot 1k} = \begin{cases} y_k, & \text{if } k \in r \\ y_{NN1k}, & \text{if } k \in s-r \end{cases}$$

and

$$y_{\cdot 2k} = \begin{cases} y_k, & \text{if } k \in r \\ y_{NN2k}, & \text{if } k \in s-r. \end{cases}$$

The point estimator of the population mean is $\bar{y}_{..s} = (\bar{y}_{.1s} + \bar{y}_{.2s})/2$, where $\bar{y}_{.1s}$ and $\bar{y}_{.2s}$ are the means of the two data sets $\{y_{.1k}: k \in s\}$ and $\{y_{.2k}: k \in s\}$, respectively. The variance estimator for MULTNN is calculated in the same way as the variance estimator for MULTREG. The only difference is that nearest neighbor imputations are used to calculate the quantities $\bar{y}_{.js}$ and $S^2_{y.js}$, $j = 1, 2$. This method is not "proper" (see Rubin, 1987, pp.118-128, for the definition of a proper multiple imputation) but was illustrated in Rubin (1986) and suggested in Rubin (1987).

## 3. THE TWELVE POPULATIONS AND THE THREE RESPONSE MECHANISMS

The performance of the different variance estimators was studied with the aid of the customary Monte Carlo summary measures: Mean, bias and variance of the variance estimators; coverage rate of the confidence interval. The performance of the different imputation methods was also investigated in terms of mean and bias of the point estimators of the population mean.

The Monte Carlo simulations were carried out using 12 different artificially generated populations. These populations were generated as follows: a set of $N = 100$ x-values was generated according to a $\Gamma$-distribution with the mean 48 and the variance 768. Then, for each fixed value of x, we generated the corresponding value of y according to a $\Gamma$-distribution with the mean $\mu(x) = \alpha + bx + cx^2$ and the variance $\sigma^2(x) = d^2x^{2g}$ with appropriately chosen constants a, b, c, d, and g. If the density of the $\Gamma$-distribution is written as

$$\frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp(-x/\beta) \quad \text{for } x > 0,$$

then the mean and the variance are, respectively, $\alpha\beta$ and $\alpha\beta^2$. We thus have the equations $\mu(x) = \alpha + bx + cx^2 = \alpha\beta$ and $\sigma^2(x) = d^2x^{2g} = \alpha\beta^2$, which imply that the constants $\alpha$ and $\beta$ used to generate the y-value associated with a given x-value are determined by

$$\alpha = \frac{\{\mu(x)\}^2}{\sigma^2(x)} = \frac{(\alpha + bx + cx^2)^2}{d^2x^{2g}}$$

$$\beta = \frac{\sigma^2(x)}{\mu(x)} = \frac{d^2x^{2g}}{\alpha + bx + cx^2}.$$

The coefficient of correlation between x and y is also a function of the five constants a, b, c, d, and g. We first specified the values for a, b, c and g, and then determined the remaining constant, d, as a consequence of the desired theoretical correlation, which we fixed at 0.75 for all populations. The values of a, b, c, d and g are given for the 12 populations in Table 1, as well as the correlation coefficient $\rho$ and the mean of y calculated from the $N = 100$ pairs of $(x_k, y_k)$, $k = 1, \cdots, 100$, that were generated by the procedure.

The constants, a, b and c are shape parameters. The populations are classified into 4 population types as coded in the above table according to the shape parameter values. Populations 1 to 3 represent a linear regression through the origin. Populations 4 to 9 have a second degree polynomial regression through the origin with slight curvature. Populations 10 to 12 are based on linear regression with a non-zero intercept.

**Table 1**
**Characteristics of the 12 Populations Used in Simulation Study.**

| Pop. | Type | a | b | c | d | g | ρ | Mean |
|---|---|---|---|---|---|---|---|---|
| 1 | RATIO | 0 | 1.50 | 0.00 | 13.78 | .25 | .773 | 70.44 |
| 2 | " | 0 | 1.50 | 0.00 | 5.13 | .50 | .775 | 73.47 |
| 3 | " | 0 | 1.50 | 0.00 | 1.84 | .75 | .755 | 72.93 |
| 4 | CONCAVE | 0 | 3.00 | -.01 | 15.04 | .25 | .760 | 112.93 |
| 5 | " | 0 | 3.00 | -.01 | 5.60 | .50 | .765 | 117.10 |
| 6 | " | 0 | 3.00 | -.01 | 2.01 | .75 | .746 | 110.31 |
| 7 | CONVEX | 0 | .25 | .01 | 13.20 | .25 | .761 | 44.77 |
| 8 | " | 0 | .25 | .01 | 4.91 | .50 | .746 | 46.51 |
| 9 | " | 0 | .25 | .01 | 0.75 | .75 | .755 | 34.93 |
| 10 | NON-RAT | 20 | 1.50 | 0.00 | 13.79 | .25 | .746 | 91.97 |
| 11 | " | 20 | 1.50 | 0.00 | 5.13 | .50 | .763 | 91.29 |
| 12 | " | 20 | 1.50 | 0.00 | 1.84 | .75 | .767 | 91.22 |

For each of the 12 populations, three different nonresponse mechanisms were used. Let $\theta_k$ denote the probability of nonresponse for the unit k. Then the three nonresponse mechanisms are as follows:

(i) $\theta_k$ decreases as y increases where $\theta_k = \exp(-c_1 y_k)$ and the constant $c_1$ is chosen so that the average nonresponse probability over the whole finite population is 0.3. (A numerical method was used to achieve this goal.) This mechanism, which is such that small y-values are under-represented among the respondents, is denoted $\downarrow$.

(ii) $\theta_k$ increases as y increases where $\theta_k = 1 - \exp(-c_2 y_k)$ and the constant $c_2$ is chosen so that the average nonresponse probability over the whole finite population is 0.3. This mechanism, which is such that small y-values are over-represented among the respondents, is denoted $\uparrow$.

(iii) $\theta_k$ is constant at 0.3 for all $k \in U$. Both large and small y-values are evenly represented among the respondents. This mechanism is denoted $\rightarrow$.

In cases (i) and (ii) the nonresponse probability depends on the value $y_k$ of the variable of interest; these nonresponse mechanisms are non-ignorable. In case (iii), the probability of nonresponse is constant throughout the population; the nonresponse mechanism is ignorable.

For each population, we drew 1,000 samples, each of size $n = 30$. For each of these samples, 50 realizations for each of the three nonresponse mechanisms were obtained by performing a Bernoulli trial on each of the 30 sample units. For each of the $12 \times 3 = 36$ different combinations of population $\times$ nonresponse mechanism, we thus obtained 50,000 realized nonresponse sets. The size of the nonresponse set $(n - m)$ is random. The expected size of the nonresponse set is $30 \times 0.3 = 9$ for each of the three mechanisms.

## 4. A SUMMARY OF THE SIMULATION RESULTS

### 4.1. Bias of the Point Estimators of the Population Mean

For the mechanism $\rightarrow$, the Monte Carlo means of all 6 point estimators agree well with the respective population means for all 12 populations. That is, all point estimators have small bias when the nonresponse is ignorable regardless population type. As expected, all point estimators are noticeably biased for the mechanism $\downarrow$ (where the bias is positive) and for the mechanism $\uparrow$ (where the bias is negative). The bias can be substantial especially for the CONVEX type populations, in which the absolute relative bias lies between 12 and 22% with the $\downarrow$ mechanism and between 30 and 37% with the $\uparrow$ mechanism. In the other cases, the absolute relative bias is less than 13%. Therefore, if the nonresponse is nonignorable, elimination of the

imputation bias in the point estimator is actually a more urgent concern than finding roughly unbiased estimators of the variance. In other words, the squared bias can be a large component of the mean square error (MSE). One should in particular be on guard against situations where a CONVEX type population is combined with the ↑ mechanism.

We noted above that the REG imputation procedure leads to the ratio estimator $\bar{y}_{.s} = (1/n)\sum_s y_{.k} = \bar{x}_s \bar{y}_r / \bar{x}_r$. For the nonignorable mechanisms ↑ and ↓, the relative bias of this estimator is considerable. This is true even for population 2, although this population is ideal for the ratio estimation in the sense that it was constructed according to a linear regression passing through the origin with the variance $\sigma^2(x) \propto x$. The bias arises because the slope estimate, $(\sum_r y_k)/(\sum_r x_k)$, is considerably biased when the nonresponse is nonignorable.

It is interesting to note that, for the nonignorable mechanisms ↑ and ↓, the nearest neighbor imputation estimators, NN and MULTNN, are more severely biased, for most of the populations, than the regression imputation estimators REG, REGRES, REGRESST and MULTREG. There are a few exceptions to this for the CONVEX type populations.

## 4.2 Variance of the Point Estimators

For a majority of the 36 cases, the regression imputation estimators, REG and REGRESST have distinctly lower variances than the nearest neighbor imputation estimators, NN and MULTNN. (MULTNN, which is formed as an average of 2 repetitions, has a slightly lower variance than NN.) However, REGRES (regression with added residual) often has a higher variance than both NN and MULTNN.

## 4.3. Bias of the Variance Estimators

Table 2 below shows overall performance of the 10 variance estimators in terms of average absolute relative bias, average relative root mean square error (MSE) and average coverage rate. Needless to say, the most important performance criterion would be the bias. In the following, we discuss in detail the performance of the 10 variance estimators based on this criterion. Figures 1-4 show graphs of relative biases of the 10 variance estimators for Populations, 2, 5, 8 and 11. The graphs are fairly similar each other for other populations within the same population type.

**Table 2**
**Overall Performance of the 10 Variance Estimators**

| Variance Estimator | Ave. Abs. Rel. Bias | Ranking | Ave. Rel. Root MSE | Ranking[1] | Ave. Cov. Rate | Ranking[1] |
|---|---|---|---|---|---|---|
| REG-ORD | 30.2 | 8 | 40.4 | - | 80.9 | - |
| NN-ORD | 38.6 | 10 | 47.0 | - | 76.9 | - |
| REGRES-ORD | 30.5 | 9 | 41.1 | - | 81.1 | - |
| REGRESST-ORD | 24.0 | 7 | 38.8 | - | 82.2 | - |
| REG-RAO1 | 17.2 | 6 | 43.8 | 4 | 85.8 | 3 |
| REG-RAO2 | 14.3 | 4 | 41.9 | 3 | 86.5 | 2 |
| REG-SARN | 11.4 | 3 | 40.2 | 2 | 87.5 | 1 |
| NN-SARN | 10.2 | 2 | 38.1 | 1 | 84.7 | 5 |
| MULTREG | 8.3 | 1 | 56.1 | 6 | 85.6 | 4 |
| MULTNN | 16.6 | 5 | 50.5 | 5 | 80.8 | 6 |

Note 1: The methods based on the ordinary variance formula are not included in the rankings because of their large absolute relative bias.

The variance estimator REG-SARN. Recall that both Population 2 and the variance estimator REG-SARN were constructed from the model $\xi$ stating that $y_k = \beta x_k + \epsilon_k$, $E_\xi(\epsilon_k) = 0$ and $V_\xi(\epsilon_k) = \sigma^2 x_k$ (that is, $g = 0.5$). Population 2 therefore represents the ideal conditions for REG-SARN, and our results confirm this. However, not only for population 2 but also populations 1 and 3 (i.e. for the RATIO type populations), REG-SARN is nearly unbiased with the absolute relative bias less than 10% for all three response mechanisms, ↑, ↓ and →. For the CONCAVE type populations 4, 5 and 6, and the NON-RAT type populations 10, 11 and 12, REG-SARN leads to minor overestimates for all three mechanisms, ↑, ↓ and →. For the CONVEX type populations 7, 8 and 9, it underestimates. The overall performance of REG-SARN w.r.t. bias is one of the best (it ranks 3rd among the 10 methods).

The variance estimators REG-RAO1 and REG-RAO2. These two variance estimators should, according to theory, work well for the mechanism → (ignorable nonresponse) regardless of population type, and this is confirmed by our study. For the nonignorable mechanisms ↑ and ↓, REG-RAO1 and REG-RAO2 will either overestimate (in the case of the mechanism ↓) or underestimate (in the case of the mechanism ↑); sometimes this bias is quite pronounced. REG-RAO2 ranks 4th of the 10 methods and has, in most of 36 cases, a smaller bias than REG-RAO1, which ranks 6th. The overall performance of REG-RAO1 and REG-RAO2 is reasonably good but somewhat less satisfactory than REG-SARN and MULTREG.

The "standard formula" variance estimators, REG-ORD, REGRES-ORD and REGRESST-ORD. All three, and especially REG-ORD and REGRES-ORD, are clearly unsatisfactory and should not be used; they yield large underestimates in virtually all of the 36 cases. One might expect that adding a residual in the REGRES-ORD method would have the effect of reducing the underestimation of the REG-ORD method. However, it is somewhat surprising to observe that the performance of REGRES-ORD is not any better than that of REG-ORD. A possible explanation for the unexpected results is that adding a residual would help properly capture the sampling variance component, $V_{sam}$, at least for the ignorable mechanism → but it would also increase the imputation variance component, $V_{imp}$. Neither REG-ORD nor REGRES-ORD contains a component aimed at estimating the imputation variance (see Särndal (1990) for the decomposition of the variance). On the other hand, REGRESST-ORD does reduce the underestimation noticeably as intended. In this method, adding standardized residual does not increase the imputation variance component which is the same as that of RES-ORD.

The variance estimators for the nearest neighbor imputation, NN-ORD and NN-SARN. Here, NN-SARN is more prone to underestimate than to overestimate (underestimates are noted in 29 of the 36 cases). The bias is fairly limited in most cases except for the cases of the CONVEX type populations with the nonresponse mechanism ↑ where it has a fairly large bias. However, the overall performance is good, ranking the 2nd among the 10 methods with an average absolute relative bias of about 10%. On the other hand, NN-ORD is inadequate and is the worst in terms of average absolute relative bias; it leads to a large underestimation, which is explained in part by the fact that this estimator contains no component for the imputation variance, $V_{imp}$, which is much bigger than the imputation variance of REG.

The multiple imputation variance estimator MULTREG. For about two thirds of 36 cases, this estimator leads to overestimation of the variance; in all 36 cases, the bias is fairly limited. It performs very well. It ranks first in term of the average relative bias among the 10 methods.

The multiple imputation variance estimator MULTNN. This method consistently underestimates in all 36 cases the imputation method is not proper. But in most cases, it is a fairly limited underestimation. The overall performance is reasonably good. It ranks 5th.

### 4.4. Variance and MSE of the Variance Estimators

A striking observation in most of the 36 cases is that the multiple imputation variance estimators, MULTREG and MULTNN, usually have a much higher variance than REG-SARN, NN-SARN, REG-RAO1 and REG-RAO2. For example, the variance of MULTREG is not uncommonly 4 to 5 times higher than the variance of REG-SARN. It follows that MULTREG and MULTNN have the drawback that the corresponding confidence intervals vary considerably in length. According to the average relative MSE criterion, NN-SARN ranks 1st and REG-SARN 2nd. MULTREG ranks 5th here due to its large variance even though its bias is the smallest in average. Here we do not discuss the variances of the "ordinary formula" variance estimators, since they are eliminated from contention due to their great bias.

### 4.5 Coverage Rate of the 95% Confidence Interval

The 95% confidence interval was constructed using a point estimate and the square root of its associated variance estimate assuming the normality of the point estimator. The coverage rate was calculated as the proportion of the number of times that the interval included the true population mean out of 50,000 samples.

As expected, the coverage rates of the ordinary variance estimators are generally much lower than the nominal value 95%. Other variance estimators achieved 85% or over on average except MULTNN. It was clearly noticed that the coverage rate is very poor for all variance estimators under the CONVEX type populations with the nonresponse mechanism $\uparrow$. For example, even for the best method, REG-SARN, the coverage rate was only between 40 and 50% as compared to the nominal 95%. The reason for this phenomenon is that the problem of severe underestimation of the point estimator is compounded by underestimation of the variance estimators in these cases. One notes that although NN-SARN ranks 2nd of the 10 methods in terms of average absolute relative bias, it ranks only 5th in terms of average coverage rate. The reason is that the point estimator is, as we noted earlier, more biased in the case of NN imputation than the case of REG imputation.

The coverage rates of the 6 good variance estimators (REG-SARN, NN-SARN, REG-RAO1, REG-RAO2, MULTREG and MULTNN) are fairly good for the nonresponse mechanism $\rightarrow$. They could have been better if the Student t reference had been used instead of the normal.

## 5. CONCLUSION

The performance of the 10 variance estimators included in the study can be summarized as follows: the 6 variance estimators, REG-SARN, NN-SARN, REG-RAO1, REG-RAO2, MULTREG and MULTNN, show reasonably good overall performance with average absolute relative bias ranging from 8 to 17%, considering all 12 populations and all three response mechanisms. They sometimes underestimate, sometimes overestimate the true variance.

At any case, all of the estimators based on the "ordinary formula", REG-ORD, REGRES-ORD, REGRESST-ORD and NN-ORD, are outright unsatisfactory and lead to gross underestimation; they should definitely not be used to estimate variance in the presence of imputation.

MULTREG, REG-SARN and NN-SARN are fairly insensitive to the nonresponse mechanism having average relative bias of less than 15% for all three nonresponse mechanisms, while MULTNN, REG-RAO1 and REG-RAO2 have average absolute relative bias of more than 20% under at least one of the nonresponse mechanisms. REG-SARN, NN-SARN, REG-RAO1 and REG-RAO2 enjoy the advantage of simplicity, since they require only a single imputation. Besides the disadvantage of storing several data sets generated by the multiple imputation, MULTREG and MULTNN often have a large variance. REG-RAO1 and REG-RAO2 perform reasonably well overall, in fact the best, as long as the nonresponse is ignorable (the mechanism $\rightarrow$). When nonresponse is nonignorable, however, their performance is less satisfactory. They still perform better than ordinary variance formulae under the $\uparrow$ mechanism but even worse than the ordinary ones under the $\downarrow$ mechanism.

The $\uparrow$ mechanism has a striking effect on the coverage rate; for all 10 variance estimators the coverage rate is <80%. On the other hand, the 6 variance estimators identified above have fairly good coverage rate under the mechanisms $\downarrow$ and $\rightarrow$.

The population type also has a big effect on the performance of the variance estimators. For all variance estimators except REG-SARN, the largest biases occur under the CONVEX type populations (7, 8 and 9) with nonignorable nonresponse mechanisms. The combination of the population type CONVEX and the $\uparrow$ mechanism has an especially harmful effect on both average absolute relative bias and coverage rate.

The type of model variance (that is, the value of g) does not seem to have much effect on the variance estimators as long as the correlation is kept constant. Its effect on the coverage rate is virtually none. However, there is reason to believe that this will not be true if the scale parameter d is kept constant instead of the correlation.

Judging from the overall performance based on the three criteria (bias, MSE and coverage rate), REG-SARN seems to be the best when the regression imputation is used and NN-SARN is the choice for the nearest neighbor imputation. If the nonresponse is believed to occur at random (the mechanism $\rightarrow$), REG-RAO2 is hard to beat for the regression imputation and NN-SARN should be used for the NN imputation. The combination of a CONVEX type population and the $\uparrow$ nonresponse mechanism creates a situation where one should be particularly careful. In this case, none of the variance estimators performs well and the confidence intervals will be grossly misleading.

Remark. The multiple imputation methodology as presented by D.B. Rubin has been developed primarily for situations where the response mechanism is ignorable, as in the case of our mechanism $\rightarrow$. In order to apply the multiple imputation methodology for nonignorable cases, such as our mechanisms $\downarrow$ and $\uparrow$, Rubin (1986) suggests to "adjust" the imputed y-values by a factor based on an assumption about the nonresponse bias. For example, in the nearest neighbor imputation, one may assume that the nonresponse bias is such that a nonrespondent will, on the average, have a y-value 20% higher than the y-value of the donor. That is, the donor y-value is multiplied by 1.2 to obtain the imputed value. Clearly, the 20% assumption is subjective; however, the success of the multiple imputation variance estimators depends on the validity of the assumption. We did not use the adjustment factor procedure in our study.


## ACKNOWLEDGEMENT

We are grateful to Y. Leblond for his participation in the early phase of the study. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada.


## REFERENCES


Deville, J.C. and Särndal, C.E. (1991). Estimation de la variance en présence de données imputées. Invited paper for the 48th Session of the International Statistical Institute, Cairo, Egypt, September, 1991.

Rao, J.N.K. (1990). Variance Estimation under Imputation for Missing Data. Manuscript, November, 1990.

Rubin, D.B. (1976). Inference and Missing Data. Biometrika, 63, 581-592.

Rubin, D.B. (1986). Basic Ideas of Multiple Imputation for Nonresponse. Survey Methodology, 12, 37-47.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: John Wiley.

Särndal, C.E. (1990). Methods for Estimating the Precision of Survey Estimates When Imputation Has Been Used. Proceedings of Statistics Canada's Symposium '90: Measurement and Improvement of Data Quality, Ottawa, October 29-31, 1990.

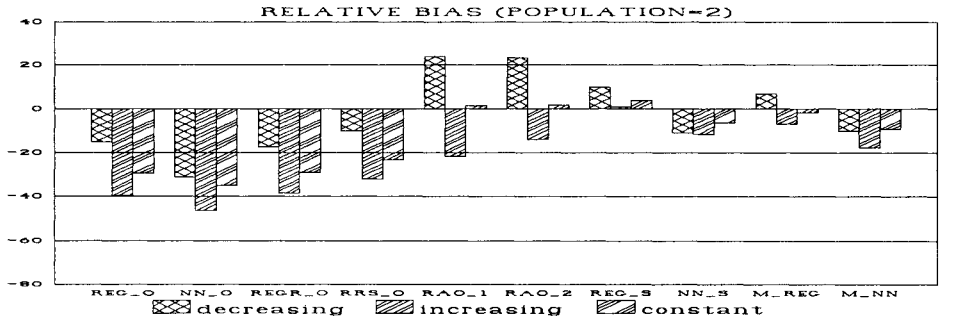


**Figure 1: Relative Biases (%) of the 10 Variance Estimators for Population 2**

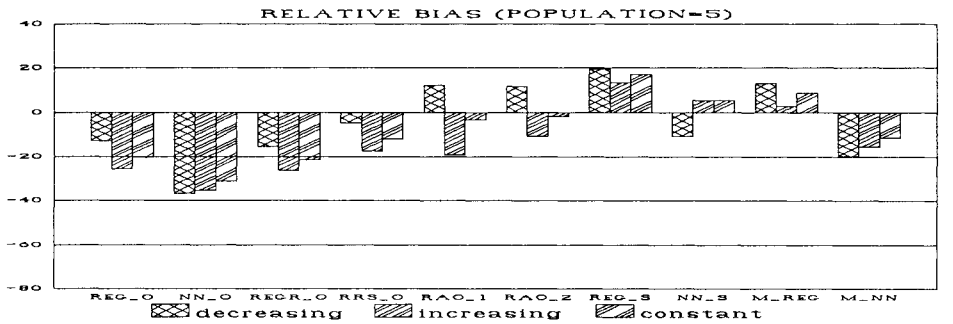


**Figure 2: Relative Biases (%) of the 10 Variance Estimators for Population 5**

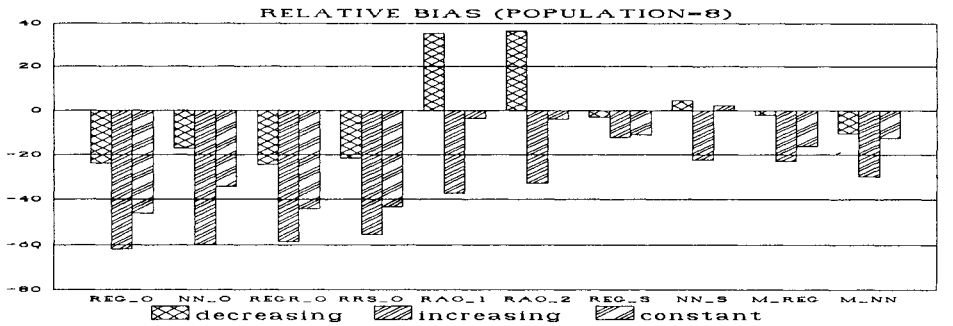


**Figure 3: Relative Biases (%) of the 10 Variance Estimators for Population 8**

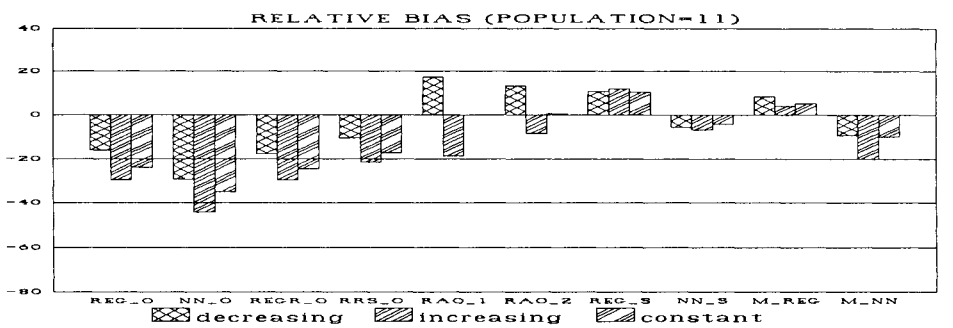


**Figure 4: Relative Biases (%) of the 10 Variance Estimators for Population 11**