ESTIMATION OF THE VARIANCE IN THE PRESENCE OF NEAREST NEIGHBOUR IMPUTATION

Eric Rancourt, Carl Särndal, and Hyunshik Lee
Hyunshik Lee, 11-Q R.H. Coats Bldg., Statistics Canada, Ottawa K1A 0T6

KEY WORDS: Imputation variance, Sampling variance, repeated donors, uniform nonresponse.

1. INTRODUCTION

The nearest neighbour (NN) imputation method is used to supply substitutes for missing data in many surveys conducted at Statistics Canada. This trend will continue since the availability of a software such as the Generalized Edit and Imputation System (GEIS) provides a relatively simple means of performing nearest neighbour imputation. Since an NN imputed value comes from a donor (one of the respondents), it is an actually occurring value, not a constructed value as in regression imputation. An NN imputed value may not be a perfect substitute, but is unlikely to be a nonsensical value. Normally, NN imputation yields point estimates with small or negligible bias, assuming that a linear relationship exists between the variable of interest y and the concomitant variable x used for nearest neighbour identification.

When the survey estimate is calculated in part from imputed values, it is not trivial matter to produce a valid estimate of its variance. It is well known that the standard complete data variance estimator severely underestimates the true variance when applied to data with imputed values. In recent years, considerable attention has been given to this problem when single value imputation is used. For example, Särndal (1990), Rao and Shao (1992), Rao and Sitter (1992), Kovar and Chen (1994), Lee, Rancourt and Särndal (1994). These attempts were very successful for regression and mean imputation but for NN imputation suggested solutions have been ad hoc. In this paper, we provide a more satisfactory solution to the variance estimation problem for NN imputation.

There are basically three approaches to variance estimation in the presence of imputation. The oldest and probably best known method is multiple imputation (Rubin, 1977, 1987). Another is the model-assisted approach (Särndal, 1990) and the third method is based on the jackknife technique (Rao, 1992). All the three approaches were tried for NN imputation by different authors with moderate success. With multiple imputation, there is some difficulty to define a "proper multiple imputation" for NN imputation and thus, the variance is underestimated (see Lee, Rancourt and Särndal, 1994). These authors also tried the model-assisted approach pretending that formulae for ratio imputation would be applicable to NN imputation as well. This worked better than the multiple imputation, but the negative bias was still present and nonnegligible (see Lee et al., 1994). The jackknife technique has been used with some success for variance estimation when the data contain imputations. However, to produce the input for the jackknife formula (the estimate recalculated after deletion of one observation), the imputed values must first be adjusted. The appropriate adjustment depends on the particular imputation method used. In particular, a difficulty with the jackknife for NN imputation has been that no entirely satisfactory adjustment has yet been found. Kovar and Chen (1994) examined the jackknife technique for NN imputation using a less than ideal adjustment, namely, with the adjustment appropriate for ratio imputation. This method substantially reduced the bias of the standard complete data variance estimator but could not eliminate it.

In this paper we develop an improved variance estimation technique for NN imputation. The method is model-assisted and gives correct variance estimation when the variable of interest y and the concomitant variable x are related with a linear regression through the origin. We obtain simple explicit estimators for the two components of the variance, that is, the sampling variance and the imputation variance. The theoretical results are presented in Section 2. In Section 3 we report the results of a Monte Carlo experiment which

confirms that the method works well for a population with regression through the origin. Section 4 presents the conclusions.

2. MODEL-ASSISTED APPROACH

Let $U=\{1, ..., k, ..., N\}$ be the index set of the population, and denoted by s a simple random sample without replacement (SRSWOR) of size n drawn from U. Let also r of size m and o of size l be respectively the sets of respondents and nonrespondents. Therefore, $s=r\cup o$. The variable of interest is denoted by y and we assume that $y_k>0$ for all $k\in U$. The parameter to estimate is the population mean of y, $\bar{y}_U=(1/N)\sum_U y_k$. In this paper we are interested in finding an estimator of $\bar{y}_U$ and a corresponding variance estimator when NN imputation is used for values that are missing because of nonresponse.

We consider single value NN imputation carried out as follows: Consider a unit $k\in o$ and suppose that

$$\min_{l\in r}|x_l-x_k|$$

occurs for $l=l(k)$. Then the value $y_{l(k)}$ is imputed for the missing value y_k. We call the $l(k)$-th unit the donor for the recipient unit k. The completed data set is thus $\{y_{\cdot k}: k\in s\}$ where

$$y_{\cdot k}=\begin{cases} y_k, & \text{if } k\in r \\ y_{l(k)}, & \text{if } k\in o \end{cases} \qquad (2.1)$$

If the survey has 100% response, then $\bar{y}_U$ is estimated by the sample mean

$$\bar{y}_s=\frac{1}{n}\sum_s y_k \qquad (2.2)$$

Its variance is estimated by

$$\hat{V}=\left(\frac{1}{n}-\frac{1}{N}\right)S_{ys}^2 \qquad (2.3)$$

where

$$S_{ys}^2=\frac{1}{n-1}\sum_s (y_k-\bar{y}_s)^2.$$

In the presence of nonresponse, the customary approach to point estimation is to take the formula for 100% response and calculate it on the completed data set. That is, from (2.2) the estimator of $\bar{y}_U$ is

$$\bar{y}_{\cdot s}=\frac{1}{n}\left(\sum_r y_k+\sum_o y_{l(k)}\right).$$

For NN imputation, the bias of $\bar{y}_{\cdot s}$ is small if the relationship between y and x is linear or approximately so.

Turning to variance estimation, the naive approach is to calculate the ordinary variance estimator (2.3) on data after imputation. This gives

$$\hat{V}_{\text{ORD}}=\left(\frac{1}{n}-\frac{1}{N}\right)S_{y^*s}^2$$

where ORD indicates "ordinary" and

$$S_{y^*s}^2=\frac{1}{n-1}\sum_s (y_{\cdot k}-\bar{y}_{\cdot s})^2$$

and $y_{\cdot k}$ is defined by (2.1). This variance estimator can be considerably off target. The objective of this paper is to present a valid approach to variance estimation when NN imputation is used for missing values.

Denote by $p(\cdot)$ the sampling design, that is, $p(s)$ is the known probability of realizing the sample s. In this paper, $p(\cdot)$ denotes the SRSWOR design. Given s, denote by $q(\cdot|s)$ the response mechanism, that is, $q(r|s)$ is the (unknown) conditional probability that the response set r is realized. We assume that $q(\cdot|s)$ is an unconfounded mechanism, that is, it may depend on the covariate values $\{x_k: k\in s\}$ but not on the values $\{y_k: k\in s\}$ of the variable of interest (see Lee, Rancourt and Särndal, 1994). The total error of $\bar{y}_{\cdot s}$ can be decomposed into sampling error and imputation error as follows:

$$\bar{y}_{\cdot s}-\bar{y}_U=\bar{y}_s-\bar{y}_U+\bar{y}_{\cdot s}-\bar{y}_s.$$

Noting that

$$E_p(\bar{y}_s)=\bar{y}_U$$
$$V_p(\bar{y}_s)=\left(\frac{1}{n}-\frac{1}{N}\right)S_{yU}^2$$

with $S_{yU}^2=\sum_U (y_k-\bar{y}_U)^2/(N-1)$, it follows easily that the bias of $\bar{y}_{\cdot s}$ is

$$B(\bar{y}_{\cdot s})=E_p\{E_q(\bar{y}_{\cdot s}-\bar{y}_s)|s\}.$$

The mean squared error (MSE) of $\bar{y}_{\cdot s}$, denoted by V, is

$$V=E_pE_q(\bar{y}_{\cdot s}-\bar{y}_U)^2=V_{\text{SAM}}+V_{\text{IMP}}+2V_{\text{MIX}}. \qquad (2.4)$$

where $V_{\text{SAM}}=(1/n-1/N)S_{yU}^2$ is the sampling variance and $V_{\text{IMP}}=E_pE_q(\bar{y}_{\cdot s}-\bar{y}_s)^2$ is the imputation variance. The V_{MIX} is the mixed term which measures the covariance between the sampling error and the imputation error. The components of (2.4) are hard to estimate unless a

model for the relationship between x and y is brought in to assist the procedure. Consider the model stating that, for $k \in U$,

$$y_k = \beta x_k + \epsilon_k \tag{2.5}$$

with $E_\xi(\epsilon_k) = 0$, $E_\xi(\epsilon_k^2) = \sigma^2 x_k$ and $E_\xi(\epsilon_k \epsilon_l) = 0$ for all $k \neq l$. The anticipated MSE (that is, the model expectation of the MSE) can be written as

$$E_\xi V = E_\xi(V_{SAM}) + E_p E_q[E_\xi\{(\bar{y}_{.s} - \bar{y}_s)^2 | s, r\}]$$
$$+ 2E_p E_q[E_\xi\{(\bar{y}_s - \bar{y}_U)(\bar{y}_{.s} - \bar{y}_s) | s, r\}]$$

The ξ-expectations appearing in the true variance components can be evaluated without difficulty, leading to expressions which depend on known x_k-values and on the unknown model parameters β^2 and σ^2. To estimate the three terms of the variance, all that we need to provide are model unbiased estimators of β^2 and σ^2 based on the data for the respondents, $\{(y_k, x_k): k \in r\}$.

1. *Estimation of* V_{SAM}

Our approach to estimating V_{SAM} is to use of the ordinary formula computed on the completed data set $\hat{V}_{ORD}$, and add a term $\hat{V}_{DIF}$ so that for all s and r,

$$E_\xi\{\hat{V}_{DIF} - (1/n - 1/N)(S_{ys}^2 - S_{y.s}^2) | s, r\} = 0 \tag{2.6}$$

We thus take

$$\hat{V}_{SAM} = \hat{V}_{ORD} + \hat{V}_{DIF} \tag{2.7}$$

The presence of $\hat{V}_{DIF}$ ensures that $\hat{V}_{SAM}$ is a correct estimator of V_{SAM} if the model holds. More explicitly, it is easy to show that $\hat{V}_{SAM}$ estimates V_{SAM} with zero anticipated bias, that is,

$$E_\xi\{E_p E_q(\hat{V}_{SAM}) - V_{SAM}\} = 0 \tag{2.8}$$

The unconfoundedness of the nonresponse mechanism ensures that the order of expectation operators E_ξ and $E_p E_q$ can be reversed and (2.8) follows from (2.6) and (2.7). (An estimator with zero anticipated bias will be called A-unbiased in the following.)

2. *Estimation of* V_{IMP}

We construct an estimator $\hat{V}_{IMP}$ satisfying

$$E_\xi\{\hat{V}_{IMP} - (\bar{y}_{.s} - \bar{y}_s)^2 | s, r\} = 0 \tag{2.9}$$

for all s and r. Then it is easy to show that $\hat{V}_{IMP}$ is A-unbiased for V_{IMP}, that is,

$$E_\xi\{E_p E_q(\hat{V}_{IMP}) - V_{IMP}\} = 0 \tag{2.10}$$

3. *Estimation of* V_{MIX}

We construct an estimator $\hat{V}_{MIX}$ satisfying

$$E_\xi\{\hat{V}_{MIX} - (\bar{y}_s - \bar{y}_U)(\bar{y}_{.s} - \bar{y}_s) | s, r\} = 0 \tag{2.11}$$

Then $\hat{V}_{MIX}$ is A-unbiased for V_{MIX}, that is,

$$E_\xi\{E_p E_q(\hat{V}_{MIX}) - V_{MIX}\} = 0 \tag{2.12}$$

The variance estimation procedure is summarized in the following result.

Result 2.1. Let $\hat{V}_{SAM}$ be defined by (2.7), where $\hat{V}_{DIF}$ satisfies (2.6). Suppose $\hat{V}_{IMP}$ and $\hat{V}_{MIX}$ satisfy (2.9) and (2.11), respectively. Then

$$\hat{V}_L = \hat{V}_{SAM} + \hat{V}_{IMP} + 2\hat{V}_{MIX} \tag{2.13}$$

is A-unbiased for V given by (2.1), that is,

$$E_\xi\{E_p E_q(\hat{V}_L) - V\} = 0$$

The proof follows easily with the aid of (2.8), (2.10) and (2.12).

This approach leads to the variance component estimators that we now describe. First define the following quantities:

$$\hat{\beta} = \sum_r y_k / \sum_r x_k$$

$$S_{xs}^2 = \frac{1}{n-1} \sum_s (x_k - \bar{x}_s)^2$$

$$S_{x.s}^2 = \frac{1}{n-1} \sum_s (x_{.k} - \bar{x}_{.s})^2$$

where $x_{.k} = x_k$ if $k \in r$ and $x_{.k} = x_{l(k)}$ if $k \in o$, $\bar{x}_{.s} = (1/n)\sum_s x_{.k}$. Finally define

$$\hat{\sigma}^2 = \frac{1}{1 - (CV_{xr})^2/m} \frac{\sum_r (y_k - \hat{\beta} x_k)^2/(m-1)}{\sum_r x_k / m}$$

and

$$\hat{\beta}^2 = (\hat{\beta})^2 - \frac{\hat{\sigma}^2}{\sum_r x_k}$$

where $CV_{xr} = S_{xr}/\bar{x}_r$ with $S_{xr}^2 = \sum_r (x_k - \bar{x}_r)^2/(m-1)$. Note that $\hat{\sigma}^2$ and $\hat{\beta}^2$ are model unbiased for σ^2 and β^2, respectively. Then we have the following expressions

for $\hat{V}_{DIF}$, $\hat{V}_{IMP}$ and $\hat{V}_{MIX}$:

$$\hat{V}_{DIF} = \left(\frac{1}{n}-\frac{1}{N}\right)\left[\hat{\beta}^2\left(S_{xs}^2 - S_{x\cdot s}^2\right) + \frac{\hat{\sigma}^2}{n(n-1)}\times \left\{2\sum_o x_k - (n-3)\sum_o d_k + \sum_r F_k(F_k-1)x_k\right\}\right]$$

$$\hat{V}_{IMP} = \frac{1}{n^2}\left[\hat{\beta}^2\left(\sum_o d_k\right)^2 + \hat{\sigma}^2\left\{2\sum_o x_k + \sum_r F_k(F_k-1)x_k + \sum_o d_k\right\}\right]$$

$$\hat{V}_{MIX} = \frac{\hat{\sigma}^2}{n^2}\left(1-\frac{n}{N}\right)\sum_o d_k$$

where F_k is the number of times that the k-th respondent is used as a donor for imputation and $d_k = x_{I(k)} - x_k$.

From a simulation study which will be reported in the next section, we noticed that $\hat{V}_{DIF}$ and $\hat{V}_{MIX}$ terms are near zero. If we drop these terms, the variance formula becomes much simpler as given below:

$$\hat{V}_S = \hat{V}_{ORD} + \hat{V}_{IMP}.$$

This estimator is subscripted by "S" to indicate that it is a short formula opposed to the long formula $\hat{V}_L$. It is interesting to note that the standard variance formula for a completed data set with NN imputation is good to estimate the sampling variance component. In a sharp contrast, this is not the case for ratio imputation.

Noting also that $\sum_o d_k$ is near zero, the short formula can be further simplified by dropping this term from $\hat{V}_{IMP}$. This further simplified variance estimator is then given by

$$\hat{V}_S = \left(\frac{1}{n}-\frac{1}{N}\right)S_{y\cdot s}^2 + \frac{\hat{\sigma}^2}{n^2}\left\{2\sum_o x_k + \sum_r F_k(F_k-1)x_k\right\} \qquad (2.14)$$

This formula was also investigated in the simulation study.

3. SIMULATION STUDY

3.1 Simulation Set-up

We carried out a simulation study to confirm that the method developed in Section 3 works. An artificial population, fairly typical of what one may encounter in practice, was generated as follows. We created $N = 400$ pairs (x_k, y_k) by first generating the x_k-values from a gamma distribution with mean 48 and variance 768. Then for each value x_k, the value y_k was generated from a gamma distribution with mean $1.5x_k$ and variance d^2x_k. The constant d was chosen in order to obtain a correlation between x and y close to 0.8. The population scatter (x_k, y_k) then follows a ratio model, that is, a linear regression through the origin, with slope close to 1.5.

From this population, a simple random sample without replacement (SRSWOR) of size 100 was drawn. Nonresponses were then randomly generated using independent Bernoulli trials with a constant parameter equal to 0.3 representing the probability of nonresponse. NN imputation was then performed for the missing values. Finally, from the completed data set, the point estimator $\bar{y}_{\cdot s}$ and the variance estimators, $\hat{V}_L$ and $\hat{V}_S$ given by (2.13) and (2.14), respectively, were calculated along with the variance components. This experiment was repeated one million times independently. A replicate sample is the resulting sample from such an experiment.

The performances of the proposed variance estimators, the long and short formulae, are assessed using the two main criteria: the relative bias and the coverage rate for the 95% confidence interval. These criteria were calculated as means and variances over the one million Monte Carlo sampling experiments. These variances and expectations are treated as true values even though they are subject to small Monte Carlo errors. We denote these Monte Carlo variance and expectation operators by V_M and E_M, respectively. Then the relative biases of $\hat{V}_L$ and $\hat{V}_S$ are given by

$$\mathrm{RB}_M = 100\times\frac{E_M(\hat{V}) - V_M(\bar{y}_{\cdot s})}{V_M(\bar{y}_{\cdot s})}$$

where $\hat{V}$ represents $\hat{V}_L$ or $\hat{V}_S$.

The 95% confidence interval was constructed using the standard normal distribution, as $\bar{y}_{\cdot s} \pm 1.96\sqrt{\hat{V}}$. The coverage rate is then given by

$$\text{COVR}_M = 100 \times \frac{t}{T}$$

where $T = 1{,}000{,}000$ and t is the number of times that the confidence interval covers the true mean.

3.2 Results

Table 1 summarizes the findings of the simulation study. The two columns headed "Sampling" refer to the repeated draws of SRSWOR samples of size $n = 100$ from $N = 400$. The column headed "Census" refers to the case $n = N = 400$, that is, the entire population is considered to be selected and nonresponse is generated by Bernoulli trials with nonresponse probability equals to 0.3 for all 400 units. Then there is no sampling variation, that is, $V_{SAM} = V_{DIF} = V_{MIX} = 0$ and imputation is the only source of variance, which is estimated by $\hat{V}_{IMP}$. The table shows in all three columns that the variance estimator developed in Section 3 works very well. The bias is virtually zero, and the coverage rate of the 95% confidence interval is very close to the nominal 95%. The short formula is an excellent approximation to the more cumbersome exact formula, the reason being that $\hat{V}_{DIF}$ and $\hat{V}_{MIX}$ are very close to zero. (It is likely to be by coincidence that the short formula actually has slightly less bias.)

4. CONCLUSIONS

The simulation study confirmed what the theory in Section 3 leads us to expect. Since the proposed estimators are obtained from the model-assisted approach, the validity of the model assumptions particularly the linearity assumption is crucial. If this assumption does not hold, then the variance estimator as well as the point estimator are biased. In this case NN imputation should not be attempted in the first place. If other imputation method is used, the variance estimator has to be changed accordingly.

If the long form of the proposed variance estimator is used, it is important to store information on the distance (denoted by d_k) between a donor and the corresponding recipient and on the number of times (denoted by F_k) that a donor is used in imputation. However, in the simplified formula, the distance information is not required. One may try to simplify the short formula even further by dropping the term containing F_k thinking that it is not important. Contrary to this, our simulation showed that this term is not negligible and thus, should not be dropped.

5. REFERENCES

Kovar, J.G., and Chen E.J. (1994). Jackknife Variance Estimation of Imputed Survey Data. *Survey Methodology*, 20, pp. 45-52.

Lee, H., Rancourt, E., and Särndal, C.E. (1994). Experiments with Variance Estimation from Survey Data with Imputed Values. *Journal of Official Statistics* (to appear).

Rao, J.N.K. and Shao, J. (1992). Jackknife Variance Estimation With Survey Data Under Hot Deck Imputation. *Biometrika*, 79, pp. 811-822.

Rao, J.N.K. and Sitter, R.R. (1992). Jackknife Variance Estimation Under Missing Survey Data. Unpublished Paper, Department of Mathematics and Statistics, Carleton University.

Rubin, D.B. (1977). Formalizing Subjective Notions about the Effect of Nonrespondents in Sample Surveys. *Journal of the American Statistical Association*, 72, pp. 538-543.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: 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*, pp. 369-380.

Table 1. Simulation Results of the Proposed Variance Estimator

	Sampling		Census
	Long Formula ($\hat{V}_L$)	Short Formula ($\hat{V}_S$)	($\hat{V}_L$)
$E_M(\bar{y}_{.s})$	70.96	70.96	70.94
$V_M(\bar{y}_{.s})$	25.91	25.91	1.57
V_{SAM}	18.99	18.99	0
V_{IMP}	6.92	6.92	1.57
$E_M(\hat{V}_L)$ or $E_M(\hat{V}_S)$	26.21	25.73	1.56
$E_M(\hat{V}_{ORD})$	18.78	18.78	0
$E_M(\hat{V}_{DIF})$	0.37	N/A	0
$E_M(\hat{V}_{IMP})$	7.08	6.95	1.56
$E_M(\hat{V}_{MIX})$	-0.01	N/A	0
Bias	0.29	-0.18	-0.01
Rel Bias	1.1%	-0.69%	-0.64%
Cov Rate	94.6%	94.4%	94.7%
Interval Length	19.99	19.81	4.88
$V_M(\hat{V}_L)$ or $V_M(\hat{V}_S)$	22.5	20.77	0.06