African Journal of Science and Technology (AJST)
Science and Engineering Series Vol. 4, No. 3, pp. 5-11

ROBUST ESTIMATION OF VARIANCE IN THE PRESENCE OF NEAREST NEIGHBOUR IMPUTATION

Charles Wafula[1], Romanus Odhiambo Otieno[2] and Mugo Maxwell Mwenda[1]

* [1]Department of Mathematics, Kenyatta University, P.O. Box 43844, Nairobi - Kenya
[2]Department of Mathematics & Statistics, Jomo Kenyatta Universityof Agriculture & Technology, P.O. Box 62000, Nairobi – KENYA

ABSTRACT:- The problem of estimating the variance of an estimator of the population total when missing values have been filled using a Nearest Neighbour (NN) imputation method is considered. The estimator is developed assuming a more general model than those considered in earlier studies. In an empirical study involving two artificial populations, the proposed estimator is found to perform better or as well as other two estimators in the current use.

INTRODUCTION

Consider a population U = {1, 2, 3, …..N}. Associated with the k – th unit of the population are two variables (x_k, y_k), k = 1, 2 … N, when $x_k > 0$, $y_k > 0$. The variable y is unknown and is the variable under study while x is the covariate assumed to be known for all units of the population. Suppose that in this sample m units respond to an item y and n – m do not. Let r denote the set of responding units and $\tilde{r}$ the set of non respondents.

In estimating the population mean $\bar{y}_u = N^{-1}\sum_{i=1}^{N} y_i$; it is usual to first fill the missing values using some imputation method. One commonly used imputation method for item response is the nearest neighbour (NN) imputation method. In what follows we consider single value NN imputation carried out as follows: Consider unit $k \in \tilde{r}$ and suppose that $\min |x - x_k|$ occurs for $l = l(k)$. Then the value $y_{l(k)}$ is imputed for the missing value y_k. The $l = l(k)$th responding unit is called the donor for the k-th missing value. The complete data set is then given by $\{y_{\cdot k} : k \in s\}$

where $y_{.k} = \begin{cases} Y_k \text{ if } k \in r \\ Y_{l(k)} \text{ if } k \in r \end{cases}$ (1)

The usual estimator of the population mean $\bar{y}_u$ is given by

$$\bar{y}_{.s} = \frac{1}{n}\sum_{k\in s} y_{.k}$$
$$= \frac{1}{n}\left\{\sum_{r} y_k + \sum_{\tilde{r}} y_{l(k)}\right\}$$
$$= \frac{1}{n}\left\{\sum_{r} y_k + \sum_{r} F_k y_k\right\} \quad (2)$$

where F_k is the number of times the K – th responding unit is used as a donor. The bias of $\bar{y}_{.s}$ is known to be small if the relationship between y and x is linear (Rancourt et al., 1994).

Let p(.) denote the sampling design; that is p(s) is the probability of obtaining the sample s. In our case P(s) is SRSWOR design. Given s, let q (./s) be the response mechanism, that is q(r/s) is the conditional probability of obtaining the response set r given s. In practice, the response mechanism is unknown. In this paper, we assume that the response mechanism may depend on the covariate values $(x_k :: k \in s)$ but not on the values $\{y_k : k \in s\}$. The total error of

$$\bar{y}_{.s} - \bar{y}_u = \left(\bar{y}_s - \bar{y}_u\right) + \left(\bar{y}_{.s} - \bar{y}_s\right) \quad (3)$$

where $\bar{y}_s$ is the sample mean when there is no non-

response. Thus bias and mean squared error of $\bar{y}_{.s}$ are given by

$$\text{Bias}\left(\bar{y}_{.s}\right) = E_p\left\{E_q\left(\left(\bar{y}_{.s} - \bar{y}_u\right)/s\right)\right\} \quad (4)$$

$$\begin{aligned}
\text{MSE}\left(\bar{y}_{.s}\right) &= E_p E_q\left(\bar{y}_{.s} - \bar{y}_u\right)^2 \\
&= E_p E_q\left[\left(\bar{y}_s - \bar{y}_u\right) + \left(\bar{y}_{.s} - \bar{y}_s\right)\right]^2 \\
&= E_p E_q\left[\left(\bar{y}_s - \bar{y}_u\right)^2 + 2\left(\bar{y}_s - \bar{y}_u\right)\left(\bar{y}_{.s} - \bar{y}_s\right) + \left(\bar{y}_{.s} - \bar{y}_s\right)^2\right] \\
&= \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{N-1}\sum_{k\in u}\left(y_k - \bar{y}_u\right)^2 + E_p E_q\left(\bar{y}_{.s} - \bar{y}_s\right)^2 \\
&+ 2E_p E_q\left(\bar{y}_{.s} - \bar{y}_s\right)\left(\bar{y}_s - \bar{y}_u\right)
\end{aligned} \quad (5)$$

$$V_{SAM} + V_{IMP} + 2V_{MIX} \quad (6)$$

The terms V_{SAM}, V_{IMP}, and V_{MIX} are defined by the terms in equation (5) respectively. The operators E_p and E_q are expectations with respect to the sampling design p(s) and response mechanism q(r/s) respectively. In this paper, we propose bias-robust estimation of MSE $\left(\bar{y}_{.s}\right)$ given by equation (5).

ESTIMATION OF MEAN SQUARED ERROR

A number of methods of estimating the mean squared error in the presence of imputed data have been proposed in the literature. A naïve method is to treat the imputed values as if they are observed values and then compute the variance estimates using standard formula. As early as the 1950s, Hansen et al., (1953) recognized that treating imputed values as observed values can lead to under estimating of variances of estimators if standard formula are used. The under estimation may become appreciable as the population of imputed values increases. Rubin (1978) introduced multiple imputation to account for inflation in the variance due to imputation. The problem with this method when applied to NN imputation is that there are some difficulties in defining a "proper multiple imputation" and, therefore, the variance is underestimated. The third method is based on the Jackknife technique (Rao and Shao, 1992). To apply this method, the imputed values must first be adjusted. The appropriate adjustment depends on the particular imputation method used. In the case of NN imputation no entirelysatisfactory adjustment has yet been found. Kovar and Chen (1994) tried the Jackknife for NN imputation using a less than ideal adjustment, that is the adjustment ideal for ratio imputation. This reduced the bias but could not eliminate it.

The fourth approach is the bootstrap (Efron, 1994; Shao and Sitter, 1996). The bootstrap can be applied to any imputation method and any estimator. However, it has two problems. First, it requires that a separate imputation be carried for every bootstrap iterate taken. Second, the properties of the bootstrap when applied to NN imputation are not known. The fifth approach is the model – assisted approach (Sarndal, 1992). Rancourt et al., (1994) applied this approach to NN imputation. These authors illustrated that the method works well under the conditions of the assumed model. However, the performance of the method when the conditions of the model do not hold was not investigated. Also the performance of the method as compared to a general method such as the bootstrap was not studied. Kaufman (1996) proposed a method for a certain variant of the NN imputation method. The method is similar to the model -assisted approach. The only difference is that in Kaufman's method a donor for a missing value is randomly chosen from the two nearest neighbours used. Therefore, the same problems associated with the model – assisted approaches are encountered in Kaufman's method. In addition, Kaufman's method introduces a donor selection variance component in the total variance. Therefore, in general Kaufman is less efficient than Sarndal's method (Fan, et al., 1998).

More recently, Montaquila and Jernigan (2002) has proposed a new approach, called the "All – cases imputation variance Estimator". The authors propose to also impute values for the respondents and then use the imputation variance for the respondents to estimate the last two terms of equation (5). The method works for any imputation method, any sampling design and any statistics used for inference. However, in their empirical study, Montaquila and Jernigan found that their estimator does not perform as well as the bootstrap estimator of Shao and Sitter (1996).

In the next section, we propose an estimator of $\text{MSE}\left(\bar{y}_{.s}\right)$ given by equation (5) using a model – assisted approach. However, the model assumed is more general that that assumed by Rancourt et al, (1994).

Consequently, the developed estimator is expected to be more robust than that suggested by Rancourt et al., (1994).

PROPOSED MEAN – SQUARED ERROR ESTIMATOR

We now turn to the problem of estimating the mean squared error given in equation (5). The development of the estimator we propose follows along the lines in the Rancourt et al., (1994). Consider the model stating that for $k \in u$,

$$E\,y_k = \beta x_k$$
$$\operatorname{cov}(y_k, y_l) = \begin{cases} \sigma(x_k), & k = l \\ 0 & k \neq l \end{cases} \qquad (7)$$

where $\sigma(x_k)$ is some smooth function of x_k. The components of equation (5) are difficult to estimate. Consequently, in the model – assisted approach, it is suggested that the anticipated mean squared error is estimated instead. Under model (7), the anticipated MSE is given by

$$E_\varsigma\left[MSE\left(\bar{y}_{.s}\right)\right] = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{N-1}E_\varsigma\sum_{k\in u}\left(y_k - \bar{y}_u\right)^2 + E_p E_q E_\xi\left(\bar{y}_{.s} - \bar{y}_s\right)^2 \qquad (8)$$

$$= E_\varsigma V_{SAM} + E_\varsigma V_{IMP} + 2E_\varsigma V_{MIX} \qquad (9)$$

Note that we have interchanged the operators E_pE_q and E_ς. This is possible because of the assumption we have made about the response mechanism. The ς -expectation appearing in the true MSE components can easily be evaluated leading to expressions which depend on β^2 and $\sigma(x_k)$ as the only unknown parameters. Therefore to estimate the three components of equation (8), all that we need to provide are the model unbiased estimators of β^2 and $\sigma(x_k)$. However, this will still not lead to an explicit estimator since we still have to find expectations of some terms with respect to the unknown response mechanism. Following Rancourt et al. (1994), we obtain A – unbiased estimators of the components of the equation (8). An estimator $\hat{\theta}$ of a parameter θ is said to be A – unbiased if its anticipated bias is zero i.e.

$$E_\varsigma E_p E_q\left(\hat{\theta} - \theta\right) = 0 \qquad (10)$$

Since the response mechanism is assumed to be uncomfounded, this is equivalent to

$$E_\varsigma E_p E_q\left(\hat{\theta} - \theta\right) = 0 \qquad (11)$$

Therefore, one way of finding an A-biased estimator of θ is to find one such that

$$E_\varsigma\left(\hat{\theta} - \theta\right) = 0 \qquad (12)$$

We now find the A-biased estimators of the components of equation (5) using condition (12)

ESTIMATION OF V_{SAM}

$$V_{SAM} = \left(1/n - 1/N\right)\frac{1}{N-1}\sum_{i\in u}\left(y_i - \bar{y}_u\right)^2$$
$$= \left(1/n - 1/N\right)\frac{1}{N-1}\left\{\sum_{i\in u} y_i^2 - N\bar{y}_u^2\right\}$$

Taking expectation of V_{SAM} with respect to model (7) we obtain

$$E_\varsigma V_{SAM} = \left(\frac{1}{n} - 1/N\right)\frac{1}{N-1}\left\{\sum_{i\in u} y_i^2 - N\ E\bar{y}_u^2\right\}$$
$$= \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{N-1}\left\{\sum_{i\in u}\left(\sigma(x_i) + \beta^2 x_i^2\right) - N\left(\frac{1}{N^2}\sum_u \sigma(x_i) + \beta^2\bar{x}_u^2\right)\right\}$$
$$= \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{N-1}\left\{\sum_{i\in u}\left(\sigma(x_i) + \beta^2 x_i^2\right) - N\left(\frac{1}{N^2}\sum_u \sigma(x_i) + \beta^2\bar{x}_u^2\right)\right\}$$
$$= \left(\frac{1}{n} - \frac{1}{N}\right)\left\{\frac{1}{N}\sum_{i\in u}\sigma(x_i) + \beta^2\frac{1}{N-1}\sum_{i\in u}\left(x_i - \bar{x}_u\right)^2\right\}$$
$$= \left(\frac{1}{n} - \frac{1}{N}\right)\left\{\frac{1}{N}\sum_{i\in u}\sigma(x_i) + \beta^2 s_x^2\right\} \qquad (13)$$

where $s_x^2 = \frac{1}{N-1}\sum_{i\in u}\left(x_i - \bar{x}_u\right)^2$ and $\bar{x}_u = \frac{1}{N}\sum_{i\in u} x_i$

We see that in $E_\varsigma V_{SAM}$ the only unknown quantities are $\sigma(x_i)$ i=1, 2, …N and β^2. Hence if $\hat{\sigma}x_i$ $\hat{\beta}^2$ are unbiased estimators of $\sigma(x_i)$ and β^2 respectively under model (7), then an A – biased estimator of V_{SAM} is given by

$$\hat{V}_{SAM} = \left(\frac{1}{n} - \frac{1}{N}\right)\left\{\frac{1}{N}\sum_{i\in u}\hat{\sigma}(x_i) + \hat{\beta}^2 s_x^2\right\} \qquad (14)$$

The estimators $\hat{\sigma}(X_i)$ and $\hat{\beta}^2$ are given in section below.

ESTIMATION OF V_{IMP}

To find an A – unbiased estimator of V_{IMP}, we need to find $E_\varsigma\left(\bar{y}_{.s} - \bar{y}_s\right)^2$ (see equation 8 above).

Now $\left(\bar{y}_{.s} - \bar{y}_s\right)^2 = \frac{1}{n^2}\left(\sum_r F_k y_k - \sum_{\bar{r}} y_k\right)^2$

$$= \frac{1}{n^2}\left\{\left(\sum_r F_k y_k\right)^2 - 2\sum_r F_k y_k \sum_{\bar{r}} y_k + \left(\sum_{\bar{r}} y_k\right)^2\right\}$$

Taking expectations under model (7), we get

$$E_\varsigma\left(\bar{y}_{.s} - \bar{y}_s\right)^2 = \frac{1}{n^2}\left\{\text{var}\sum_r F_k y_k + \left(\sum_r F_k E\, y_k\right)^2 - 2E\sum_r F_k y_k\, E\sum_{\bar{r}} y_k + \text{var}\sum_{\bar{r}} y_k + \left(E\sum_{\bar{r}} y_k\right)^2\right\}$$

$$= \frac{1}{n^2}\left\{\sum_r F_k^2\sigma(x_k) + \beta^2\left(\sum_r F_k x_k\right)^2 - 2\beta^2\sum_r F_k x_k \sum_{\bar{r}} x_k + \sum_{\bar{r}}\sigma(x_k) + \beta^2\left(\sum_{\bar{r}} x_k\right)^2\right\}$$

$$= \frac{1}{n^2}\left\{\sum_r F_k^2\sigma(x_k) + \beta^2\left[\sum_r F_k x_k - \sum_r x_k\right]^2 + \sum_r\sigma(x_k)\right\}$$

$$= \frac{1}{n^2}\left\{\sum_r F_k^2\sigma(x_k) + \beta^2\left[\sum_{\bar{r}} x_{l(k)} - \sum_{\bar{r}} x_k\right]^2 + \sum_{\bar{r}}\sigma(x_k)\right\}$$

$$= \frac{1}{n^2}\left\{\sum_r F_k^2\sigma(x_k) + \beta^2\left(\sum_{\bar{r}} d_k\right)^2 + \sum_{\bar{r}}\sigma(x_k)\right\} \qquad (15)$$

where $d_k = x_{l(k)} - x_k$. Therefore the A-unbiased estimator of V_{IMP} is given by

$$\hat{V}_{IMP} = \frac{1}{n^2}\left\{\sum_{\bar{r}} F_k^2\hat{\sigma}(x_k) + \hat{\beta}^2\left(\sum_{\bar{r}} d_k\right)^2 + \sum_{\bar{r}}\hat{\sigma}(x_k)\right\} \qquad (16)$$

ESTIMATION OF V_{MIX}

Here we need $E_\varsigma\left(\bar{y}_s - \bar{y}_u\right)\left(\bar{y}_{.s} - \bar{y}_s\right)$

$$\bar{y}_s - \bar{y}_u = \frac{1}{n}\left(\sum_r y_i + \sum_{\bar{r}} y_i\right) - \frac{1}{N}\left(\sum_r y_i + \sum_{\bar{r}} y_i + \sum_{\bar{s}} y_i\right)$$

$$= \left(\frac{1}{n} - \frac{1}{N}\right)\left(\sum_r y_i + \sum_{\bar{r}} y_i\right) - \frac{1}{N}\sum_{\bar{s}} y_i$$

The symbol $\sum_{\bar{s}}$ indicates summation over non sample values.

Hence

$$\left(\bar{y}_s - \bar{y}_u\right)\left(\bar{y}_{.s} - \bar{y}_s\right) = \frac{1}{n}\left(\frac{1}{n} - \frac{1}{N}\right)\left(\sum_r y_i + \sum_{\bar{r}} y_i\right)\left(\sum_r F_k y_k - \sum_{\bar{r}} y_k\right)$$

$$- \frac{1}{nN}\sum_{\bar{s}} y_i\left(\sum_r F_k y_k - \sum_{\bar{r}} y_k\right)$$

$$= \frac{1}{n}\left(\frac{1}{n} - \frac{1}{N}\right)\left\{\sum_r F_i y_i^2 + \sum_{i\neq k}\sum F_i y_i y_k - \sum_r y_i\sum_{\bar{r}} y_k + \sum_{\bar{r}} y_i\sum_r F_k y_k - \sum_{\bar{r}} y_i^2 - \sum_{i\neq k}\sum_{\bar{r}} y_i y_k\right\}$$

$$- \frac{1}{nN}\left\{\sum_{\bar{s}} y_i\sum_r F_k y_k - \sum_{\bar{s}} y_i\sum_{\bar{r}} y_k\right\} \qquad (17)$$

Taking expectations of equation (17) under model (7), we obtain

$$E_\varsigma\left(\bar{y}_S - \bar{y}_u\right)\left(\bar{y}_{.s} - \bar{y}_S\right) = \frac{1}{n}\left(\frac{1}{n} - \frac{1}{N}\right)\left\{\sum_r F_i\left(\sigma(x_i) + \beta^2 x_i^2\right)\right.$$
$$+ \beta^2\sum_{\substack{r\\ i\neq k}} x_i x_k - \beta^2\sum_r x_i\sum_{\bar{r}} x_k + \beta^2\sum_{\bar{r}} x_i\sum_r F_k x_k$$
$$\left. - \sum_{\bar{r}}\left(\sigma(x_i) + \beta^2 x_i^2\right) - \beta^2\sum_{i\neq k} x_i x_k\right\}$$

$$\frac{1}{nN}\left\{\beta^2\sum_s x_i\sum_r F_k x_k - \beta^2\sum_s x_i\sum_r x_k\right\} \qquad (18)$$

On simplification equation (18) becomes

$$E_s\left(\bar{y}_s - \bar{y}_u\right)\left(\bar{y}_{.s} - \bar{y}_s\right)$$
$$= \frac{1}{n^2}\left\{\left(1 - \frac{n}{N}\right)\left(\sum_r F_k\sigma(x_k) - \sum_{\bar{r}}\sigma(x_k)\right) + n\beta^2\left(\bar{x}_s - \bar{x}_u\right)\sum_{\bar{r}} d_k\right\} \qquad (19)$$

where $\bar{x}_s = \frac{1}{n}\sum_s x_i$. Therefore the A – unbiased estimator of V_{MIX} is given by

$$\hat{V}_{MIX} = \frac{1}{n^2}\left\{\left(1 - \frac{n}{N}\right)\left(\sum_r F_k\hat{\sigma}(x_k) - \sum_{\bar{r}}\hat{\sigma}(x_k)\right) + n\hat{\beta}^2\left(\bar{x}_s - \bar{x}_u\right)\sum_{\bar{r}} d_k\right\} \qquad (20)$$

ESTIMATION OF σ (.) AND β^2

A number of methods have been suggested for estimating σ (.). In this paper, we - propose to estimate σ (.) using the estimator given in Odhiambo (1995)

The estimator given in Odhiambo (1995) was developed under very general conditions. Hence it is a more robust estimator of σ (.) than any other estimator.

To estimate β^2, consider

$$G = \left(\frac{\sum_r y_i}{\sum_r x_i} \right)^2$$

$$G = \left(\frac{\sum_r y_i}{\sum_r x_i} \right)^2$$

Under model 7

$$E(G) = \beta^2 + \frac{\sum_r \sigma x_i}{\left(\sum_r x_i \right)^2}$$

Hence an unbiased estimator of β^2 is given by

$$\beta_0 = G - \frac{\sum_r \hat{\sigma}(x_i)}{\left(\sum_r x_i \right)^2}. \qquad (21)$$

Finally, from equations (14), (16), and (20), an A – biased estimator of MSE $\left(\bar{y}_{.s} \right)$ is given by

$$V_o = \hat{V}_{SAM} + 2\hat{V}_{MIX} + \hat{V}_{IMP} \qquad (22)$$

SIMULATION STUDY

SIMULATION SET-UP

We studied the performances of the estimators V_0 (see equation 22), the bootstrap estimator (V_1) and V_2 – the estimator due to Rancourt et al. (1994) in a simulation study involving the artificial population and two non response Mechanisms.

The first population 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. The value Y_k was generated from gamma distribution with mean 1.5 X_k and variance 0.25 X_k. This population was also used by Rancourt et al. (1994) in their simulation study. The estimator V_2 was developed assuming a simple linear regression model. Therefore it is expected to perform well in this population.

The second population was generated in a similar manner as the first with only one difference. The y_k value was generated from a gamma distribution with a variance 0.25 x^2_k instead of 0.25 x_k. This population represents the case when the model considered in Rancourt et al (1994) fails. It provides a good situation where the robustness of the three estimators can be tested.

Two non-response mechanisms were considered: random and unknown. Previous studies (for example Kovar and Chen, 1994) have shown that the properties of variance estimators in the presence of imputed values are more pronounced when non-response rates are high. Most of these studies considered 30% non-response rate to be a high one. We adopt the same in this study. Hence the two non-response mechanisms were generated using independent Bernoulli trials with a constant parameter equal to 0.3 representing the probability of non-response. In the unknown non-response mechanism case, we just took the last 30% of the sample values as missing.

Most of the previous simulation studies comparing variance estimators for imputed data considered negligible sampling fractions. See for example Kovar and Chen (1994), Montaquila and Jernigan (2002) and Rancourt et al. (1994). In this study we consider the case when the sampling fraction is not negligible.

From each population a simple random sample of size n=90 was taken. For each non-response mechanism, non-respondents were generated. Nearest Neighbour imputation was then performed for the missing values. Finally, from the completed data set the estimator $\bar{y}_{.s}$ and the three variance estimators V_0, V_1 and V_2 were calculated. The experiment was independently repeated 1000 times. In the case of V_1, 1000 bootstrap iterations were used while for V_2, the bandwidth parameter was that one that minimized the mean squared error and satisfied Silverman (1986)'s condition.

$\frac{\sigma}{4n^{1/5}} \le h \le \frac{3\sigma}{2n^{1/5}}$. The Epanechnikov's kernel was used in this case.

The performances of the three estimators were assessed using two criteria: the relative bias (RB) and the coverage rate for the nominal 95% confidence intervals. The relative biases of V_0, V_1 and V_2 were calculated as

$$RB = \frac{1}{1000} \frac{\sum_{i=1}^{1000} \left(v_i - MSE\left(\bar{y}_{.s} \right) \right)}{MSE\left(\bar{y}_{.s} \right)}$$

where $MSE\left(\bar{y}_{.s} \right) = \frac{1}{1000} \sum_{i=1}^{1000} \left(\bar{y}_{.si} - y_u \right)^2$

the value of $\bar{y}_{.s}$ for the ith experiment and V_i represents the value of V_0, V_1 and V_2 for the ith experiment. The 95% confidence interval was constructed using the standard normal distribution as $\bar{y}_{.s} \pm 1.96\sqrt{v_i}$. The coverage rate is then given by $CR = 100X\frac{t}{T}$ where T = 1000 and t is the number of times that the confidence interval covers the true mean.

RESULTS

Relative bias (RB) and coverage rate (CR) of V_0, V_1 and V_2.

Population	Variance Estimator	Non-response Mechanism			
		Random		Unknown	
		RB	CR	RB	CR
1	V_0	0.024	96.0	0.030	95.5
	V_1	-0.065	92.1	-0.067	92.0
	V_2	0.112	91.5	0.120	91.2
2	V_0	-0.091	96.5	-0.126	94.5
	V_1	0.140	93.5	0.177	94.0
	V_2	0.023	95.0	-0.080	94.5

As the results in the table show, V_1 underestimates the MSE in all cases. Rancourt et al. (1994) observed the same results. Generally, V_2 tends to overestimate the MSE. Clearly, between V_0 and V_1, V_0 is a better estimator in terms of relative bias. It is somewhat surprising to see the better performance of V_0 over V_1 in population 1. As has been remarked, $V_1$was developed assuming a simple linear regression model. Population 1 follows a linear regression model. Hence V_1 should perform best in this population.

Rubin (1996) emphasizes the fact that rarely is the variance of an estimator itself an estimand. That is, rarely is the sole purpose only to estimate the variance of an estimator. Rather the goal is to obtain valid inferences. The variance estimator is merely a vehicle used en route to obtaining valid inferences. Thus in comparing estimators, it is important to assess the validity of inferences obtained using the estimators.

Table 1 also compares the coverage rates for the 95% confidence intervals obtained using the three variance estimators. The proposed estimators V_0 tends to yield confidence intervals with higher coverage rates than those obtained using the other two estimator

CONCLUSION

We have developed a bias robust variance estimator, V_0, in the case where missing values are imputed using a nearest neighbour imputation method. The estimator was developed using more general condition than those used when developing V_1. Because of this V_0 is expected to be more bias – robust than V_1. Our empirical study confirms this.

The proposed estimator is not as computationally intensive as bootstrap. The bootstrap involves drawing numerous bootstrap samples, imputing independently within each bootstrap sample, and computing an estimate for each bootstrap sample. To properly implement this procedure and ensure its validity for a specified problem, the validity checks used full – sample imputation must be used for each bootstrap sample. This can be quite time consuming and labour intensive. The advantage of the bootstrap over the proposed estimator is that the bootstrap can be applied to any sampling design, any imputation method and any type of estimator while the proposed estimator is specific to the method of imputation.

In this study, we did not include the more recent estimator of Montaquila and Jernigan for comparison. The recent approach has all the advantages of the bootstrap and is not as computationally intensive as the bootstrap. We are looking at this problem at the moment.

REFERENCES

Efron, B. (1994). Missing Data, Imputation and the Bootstrap Journal of the American Statistical Association, 89, 426.

Fan, Z., Brick, M., Kaufman, S., and Walter, E. (1998). Variance estimation of imputed Survey Data. Working paper series No. 98 – 14. National Centre for Education Statistics, U.S.

Hansen, M., Hurwitz, W., and Madow, W. (1953). Sample Survey Methods and Theory, Volume 2. New York: John Wiley an Sons, Inc. PP 423 – 428.

Kaufman, S. (1996). Estimating the variance in the presence of imputation using a residual. In 1996 proceedings of the section on Survey Research methods (pp 423 428). Alexandria, VA: American Statistical Association.

Kovar, J.G., and Chen, E.J. (1994). Jackknife variance estimation of imputed Survey Data. Survey Methodology, 20 45 – 52.

Montaquila, J.M., and Jernigan, R.W. (2002) Variance estimation in the presence of imputed data. Draft Manuscript.

Odhiambo, R.O. (1995). Robust Estimation for finite population sampling. Unpublished Ph.D thesis. Kenyatta University, Kenya.

Rancourt, E., Sarndal, C.E., and Lee, H. (1994). Estimation of the variance in the presence of nearest neighbour imputation. In 1994 proceedings of the section on Survey Research Methods (pp. 888 893). Alexandria, VA: American Statistical Association.

Rao, J.N.K., and Shao, J. (1992). Jackknife Variance with Survey Data under hot deck imputation. Biometrika, 79, 811 – 822.

Rubin, D.B. (1978). Multiple imputation in sample surveys – a phenomenological Bayesian approach to non-response. In the 1978 proceedings of the section on survey research methods (pp 20 –34). Washington, D.C. American Statistical Association.

Rubin, D.B. (1996). Multiple imputation After 18+ years. Journal of the American Statistical Association, 91, 473 – 489.

Sarndal, C.E. (1996). Methods for estimating the precision of survey estimates when imputation has been used. Survey Methodology, 18. 241 – 252.

Shao, J. and Sitter, R.R. (1996) Bootstrap for imputed survey Data. Journal of the American Statistical Association, 91, 1278 – 1288.

Silverman, B. (1986). Density Estimation for statistics and Data analysis. London: Chapman and Hall.