American Journal of Theoretical and Applied Statistics
2016; 5(5): 252-259
http://www.sciencepublishinggroup.com/j/ajtas
doi: 10.11648/j.ajtas.20160505.11
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

Non-parametric Variance Estimation Using Donor Imputation Method

Hellen W. Waititu[1, *], **Edward Njenga**[2]

[1]Department of Statistics and Computer Sciences, Moi University, Nairobi, Kenya

[2]Department of Mathematics, Kenyatta University, Nairobi, Kenya

Email address:

hwaititu@yahoo.com (H. W. Waititu)

[*]Corresponding author

To cite this article:

Hellen W. Waititu, Edward Njenga. Non-parametric Variance Estimation Using Donor Imputation Method. *American Journal of Theoretical and Applied Statistics.* Vol. 5, No. 5, 2016, pp. 252-259. doi: 10.11648/j.ajtas.20160505.11

Received: July 1, 2016; **Accepted**: July 16, 2016; **Published**: August 3, 2016

Abstract: The main objective of this study is to investigate the relative performance of donor imputation method in situations that are likely to occur in practice and to carry out numerical comparative study of estimators of variance using Nadaraya-Watson kernel estimators and other estimators. Nadaraya-Watson kernel estimator can be viewed as a non-parametric imputation method as it leads to an imputed estimator with negligible bias without requiring the specification of a parametric imputation model. Simulation studies were carried out to investigate the performance of Nadaraya-Watson kernel estimators in terms of variance. From the results, it was found out that Nadaraya-Watson kernel estimator has negligible bias and its variance is small. When compared with Naïve, Jackknife and Bootstrap estimators, Nadaraya-Watson kernel estimator was found to perform better than bootstrap estimator in linear and non-linear populations.

Keywords: Hot Deck Imputation, Non-parametric, Unbiased Estimator, Donor, Recipient, Donor Imputation

1. Introduction

Donor imputation is a method in which the missing values for one or more variables of a non responding unit (recipient) are replaced by the corresponding values of a responding unit (donor) with no missing value for these variables. It is a variance estimation method which is valid even in the presence of high sampling fractions [1]. However, very few variance estimation methods that take into account donor imputation have been developed. Essentially, donor imputation is convenient and has some interesting statistical properties. Although donor imputation may not be the most efficient method in any specific scenario, it is popular in surveys due to its practical advantages. Therefore, it remains useful to develop variance estimation methods that take donor imputation into account. In this study, variance estimator after donor imputation have been investigated and compared with the Naïve estimator, Jackknife estimator and Bootstrap estimator. Variance estimation methods accounting for the effect of imputation have been studied by [11], [13] and [8], among others. Some methods of variance estimation that have been developed for use with imputed data include a model-assisted method [11], an adjusted jackknife method [11], and multiple imputations [8]. [2] considered Random Hot-Deck (RHD) imputation under more general sampling designs assuming a one-factor analysis of variance model holds. [9], [6] and [5] dealt with Nearest Neighbor Imputation (NNI). [3] considered NNI, an alternative to re-sampling variance estimation method. [10] considered NNI under simple random sampling assuming that a ratio imputation model holds. [1] dealt with general donor imputation methods including NNI and with possibly post-imputation edit rules and hierarchical imputation classes, under general sampling designs and more general imputation models. In this paper, non-parametric variance estimation using donor imputation method have been considered with estimation of parameters $\hat{\mu}(X_i)$ and $\hat{\delta}^2(X_i)$ being done using the kernel method proposed by Nadaraya (1964) and Watson (1964).

2. Estimation Procedure

Consider a population of N elements identified by a set of indices $U = \{1, 2,\ldots, N\}$. Associated with the i^{th} unit in the population are two variables (X_i, Y_i) where$X_i > 0, Y_i > 0$. The variable Y has some unknown values and it is the variable under study. The variable X is the auxiliary variable assumed to be known for all units of the population. A simple random sample without replacement (SRSWOR) of size n denoted as s is drawn from the population. Suppose that $y_1, y_2, \ldots, y_r$ are observed (respondents) and $y_{r+1}, y_{r+2}, \ldots$, y_n are missing (non-respondents). That is r units respond for y and $m = n - r$ do not respond. Therefore $s = r \cup m$. Consider a unit $i \in s$. The NNI method imputes a missing y_j by y_i where $i = 1,2,\ldots,r$ and $j = 1,2,..,m$. i is the nearest neighbor of j measured by the x variable. That is i satisfies $|x_i - x_j|$=$\min_{1\le i\le r}|x_i - x_j|$. If there are tied x values, then there may be multiple nearest neighbors of j and i is randomly selected from them. Suppose that $min|x_i - x_j|$occurs for $l = l(i)$. Then the value $y_{l(i)}$ is imputed for the missing y_j.

The completed data set is

$$\{y_i' : i \in s\} \tag{1}$$

Where $y_i' = \begin{cases} y_i, & \text{if } i \in r \\ y_{l(i)}, & \text{if } i \in m \end{cases}$. If the survey has 100% response, then the populations mean

$$\bar{Y}_U = \frac{1}{N}\sum_U Y_i \tag{2}$$

is estimated by the sample mean $\bar{y}_s = \frac{1}{n}\sum_s y_i$ and its variance is estimated by

$$\hat{V} = \left(\frac{1}{n} - \frac{1}{N}\right) s_y{}^2 \tag{3}$$

where $s_y{}^2 = \frac{1}{n-1}\sum_s (y_i - \bar{y}_s)^2$.

In the presence of non-response, the customary approach to point estimation is to take the formula for 100% response and calculate it on the completed data set. Thus from (2), the estimator of $\bar{Y}_U$ is $\bar{y}_s' = \frac{1}{n}\left(\sum_r y_i + \sum_m y_{l(i)}\right) = \frac{1}{n}[\sum_r y_i + \sum_m (F_i y_i)]$ where F_i is the number of times the i^{th} responding unit is used as a donor. For variance estimation, the naïve approach is to calculate the ordinary variance estimator, $\hat{V}_{ORD}$, to (3) on data after imputation. i.e. $\hat{V}_{ORD} = \left(\frac{1}{n} - \frac{1}{N}\right) s_{y_s'}{}^2$ where $s_{Y_s}^2 = \frac{1}{n-1}\sum_s (y_s' - \bar{y}_s')^2$ and $y_i'$is defined by (1). This variance estimator can be biased.

Let $p(\cdot)$ denote the sampling design, that is, $p(s)$ is the known probability of obtaining a sample s. In our case, $p(s)$ denote the SRSWOR design. Given s, denote the response mechanism by $q(\cdot/s)$. i.e. $q(r/s)$ is the unknown conditional probability that the response set r is obtained. We assume that $q(\cdot/s)$ may depend on the auxiliary variable $\{x_i : i \in s\}$ but not on the values $\{y_i : i \in s\}$. The total error (sum of sampling error and imputation error) of $\bar{y}_s'$ can be broken down into sampling error and imputation error as follows

$$\bar{y}_s' - \bar{Y}_U = (\bar{y}_s - \bar{Y}_U) + (\bar{y}_s' - \bar{y}_s)$$

We note that $E_p(\bar{y}_s) = \bar{Y}_U$

$V_p(\bar{y}_s) = \left(\frac{1}{n} - \frac{1}{N}\right) s_{Y_U}^2$, where $s_{Y_U}^2 = \sum_U \frac{(Y_i - \bar{Y}_U)^2}{N-1}$

Thus the bias of $\bar{y}_s'$ is $B(\bar{y}_s') = Ep[Eq(\bar{y}_s' - \bar{y}_s)\,/\,s]$ Variance of $\bar{y}_s'$ denoted by V is given by

$$V = EpEq(\bar{y}_s' - \bar{Y}_U)^2 = V_{sam} + V_{imp} + 2V_{mix} \tag{4}$$

V_{sam} is a standard variance estimator using the imputed values as if they were reported values. This is called the naïve variance estimator. [2] show that under the cell mean model and hot deck imputation, the bias of the naïve variance estimator as an estimator for V_{sam} is small when no respondent is used too often as a donor of an imputed value.

The jackknife variance estimator of $\bar{y}$ is given by $\hat{V}_j = \left(\frac{n-1}{n}\right)\sum_{j=1}^{n}\left(\bar{y}_{(j)} - \bar{y}\right)^2$[8]. In the presence of non-response to item y, the use of the above estimator may lead to serious underestimation of the variance of the estimator, especially if the non-response rate is important. [11] proposed an adjusted jackknife method that is calculated in a similar fashion as the above estimator except that, whenever a responding unit is deleted, the imputed values are adjusted. The imputed values are unchanged if a non-responding unit is deleted. Let $y_{i(j)}{}^{a*}$, denote the adjusted imputed value for unit i when unit j was deleted. For mean imputation, we have $y_{i(j)}{}^{a*} = \begin{cases} \bar{y}_{r(j)}, & \text{if } r_j = 1 \\ \bar{y}_r, & \text{if } r_j = 0 \end{cases}$ where $\bar{y}_{r(j)}$ denotes the mean of the respondents excluding unit j . The Rao-Shao jackknife variance estimator is then given by

$$\hat{V}_{jRs} = \left(\frac{n-1}{n}\right)\sum_{j=1}^{n}\left(\hat{\bar{y}}_{l(j)}^{\mathrm{a}} - \hat{\bar{y}}_l\right)^2$$

The bootstrap method is estimated by $\hat{V}_{BTS} = \frac{1}{k}\sum_{r=1}^{k}\left(\hat{\theta}_r - \hat{\theta}_{(\cdot)}\right)^2$ where $\hat{\theta}_{(\cdot)} = \frac{1}{\mathrm{k}}\sum_{\mathrm{r}=1}^{\mathrm{k}}\theta_{\mathrm{r}}$. [3] proposed a rescaling Bootstrap method in order to estimate the Variance. Their method draws bootstrap samples of size n' with replacement from the rescaled samples$s$. Note that n' may be different from n. The rescaling factor, denoted by$C$, is chosen so that the variance under re-sampling matches the usual variance estimator of the population mean.

The Rao-Wu bootstrap variance estimator is given by $\hat{V}_{Rw} = \frac{1}{B-1}\sum_{b=1}^{B}\left(\bar{z}_{(b)}{}^{*} - \bar{z}_{(\cdot)}^{*}\right)^2$ where $\bar{z}_{(\cdot)}^{*} = \sum_{b=1}^{B}{}^{\bar{z}_{(b)}^{*}}\!/_{B}$.

Applying the Rao-Wu bootstrap in the presence of missing responses and treating the missing values as true values, may lead to serious underestimation of the variance of the estimator. In the presence of imputed data, [12] proposed a bootstrap procedure for imputed survey data. The Shao-Sitter bootstrap variance estimator is given by

$$\hat{V}_{Bss} = \frac{1}{(B-1)}\sum_{b=1}^{B}\left(\bar{z}_{(b)}^{*'} - \bar{z}_{(\cdot)}^{*'}\right)^2, \text{ where } \bar{z}_{(\cdot)}^{*'} = \sum_{b=1}^{B}{}^{z_{(b)}^{*'}}\!\Big/_{B}$$

2.1. Donor Imputation

A sample s of size n is drawn from population total U according to a probability sampling design $p(s)$. In the absence of non-response, we assume SRSWOR with mean $\bar{y}_s$.

Variable y is only observed for a subset s_r of s according to a response mechanism $q(s_r|s)$. This subset of size n_r is called the set of respondents (or donors) while its complement $s_m = s - s_r$ of size $n_m = n - n_r$ is called the set of non-respondents (or recipients). To compensate for the missing y values, donor imputation is performed. This leads to the imputed estimator of the mean given by

$$\bar{y}_I = \frac{1}{n}\left[\sum_{i\in s_r} y_i + \sum_{i\in s-s_r} y_i\right] = \frac{1}{n}\left[\sum_{i\in r} y_i + \sum_{i\in r} F_i y_i\right]$$

where $F_i = \begin{cases} 0\, if i \in r \\ x_{l(i)} if i \in \bar{r} \end{cases}$

$l(i) \in s_r$ is the donor used to impute the recipient i. A variety of strategies can be considered in practice in order to find donors for imputing recipients. Usually, a vector X_i of auxiliary variables, available for all the sample units $i \in s$, is used to determine a set S_m^*, of selected donors that are "close" to the corresponding recipients in s_m.

2.2. Approach to Inference

To evaluate properties of the imputed mean estimator $\bar{y}_I$ and to make inferences, the following imputation model is used:

$$\begin{cases} E_m(y_i/X) = \mu(X_i) \\ V_m(y_i/X) = \delta^2(X_i) \\ Covv_m(y_i, y_j\,/\,X) = \begin{cases}\delta^2(X_i) if i = j \\ 0\, if i \neq j\end{cases} \end{cases} \tag{5}$$

where the subscript m indicates that the expectation, variance, and covariance are evaluated with respect to the imputation model, X is the N-row matrix containing x_i' in its i^{th} row, and $\mu(X_i)$ and $\delta^2(X_i)$ are parametric or non-parametric smooth functions of X. Note that the subscript m in $n_m, s_m, and\ s_{m,i}$ indicates missing values and should not be confused with the imputation model.

The vector X contains variables used at the imputation stage for the selection of donors. In principle, the imputer uses available variables that are associated with the y-variable. The vector X may thus contain design variables (e.g., strata and cluster indicators, size measure), the domain of interest or other auxiliary variables. It is assumed in model (5) that the imputer has appropriately chosen the vector X of auxiliary variables so that the design variables and the domain of interest do not explain further the y-variable after conditioning on X. This allows us to treat the design variables and the domain(s) of interest as being fixed under model (5).

3. Proposed Variance Estimator

Considering model (5), the total error of $\bar{y}_I$ can be broken down into sampling error and imputation error as shown in (4). The expectation appearing in the true variance component can be evaluated leading to expressions which depend on known x_i values and on the unknown model parameters $\mu(X_i)$ and $\delta^2(X_i)$. Therefore to estimate the three components of the variance, all we need to provide are the model unbiased estimators of $\mu(X_i)$ and $\delta^2(X_i)$. However, this will not completely lead to an explicit variance estimator since we still have to obtain expectations of some terms with respect to response mechanism.

3.1. Estimation of V_{SAM}

$$V_{SAM} = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{N-1}\sum_U (\bar{y}_i - \bar{Y}_U)^2\, E(V_{SAM})$$
$$= \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{N-1} E\left[\sum (y_i^2 + \bar{Y}_U^2 - 2y_i\bar{Y}_U)\right]$$
$$= \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{N-1}\left\{\left(1 - \frac{1}{N}\right)\left(\sum_U \delta^2(X_i) + \sum_U [\mu(X_i)]^2\right)\right\}$$

Hence, unbiased estimator of V_{SAM} is

$$\hat{V}_{SAM} = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{N-1}\left\{\left(1 - \frac{1}{N}\right)\left(\sum_s \hat{\delta}^2(X_i) + \sum_s [\hat{\mu}(X_i)]^2\right)\right\}$$

Where $\hat{\delta}^2(X_i)$ and $\hat{\mu}(X_i)$ are model unbiased estimators of $\delta^2(X_i)$ and $\mu(X_i)$ respectively.

3.2. Estimation of V_{IMP}

$$V_{IMP} = E(\bar{y}_I - \bar{y})^2$$

$$\bar{y}_I - \bar{y} = \frac{1}{n}\left\{\sum_r y_i + \sum_r F_i y_i - \sum_r y_i - \sum_{\bar{r}} y_i\right\}$$

$$(\bar{y}_I - \bar{y})^2 = \frac{1}{n^2}\left\{\left(\sum_r F_i y_i\right)^2 + \left(\sum_{\bar{r}} y_i\right)^2 - 2\sum_r F_i y_i \sum_{\bar{r}} y_i\right\}$$

$$E(\bar{y}_I - \bar{y})^2 = \frac{1}{n^2}\left\{Var\sum_r (F_i y_i) + \left[E\sum_r (F_i y_i)\right]^2 + Var\sum_{\bar{r}} (y_i) + \left[E\left[\sum_{\bar{r}} (y_i)\right]\right]^2 - 2\left[E\sum_r F_i y_i . E\sum_{\bar{r}} y_i\right]\right\}$$
$$= \frac{1}{n^2}\left\{\sum_r F_i^2 \delta^2(X_i) + \sum_{\bar{r}} . \delta^2(X_i) + \sum_{\bar{r}} \left[x_{l(i)} - \mu(X_i)\right]^2\right\}$$

Hence an unbiased estimator of V_{IMP} is

$$\hat{V}_{IMP}=\frac{1}{n^2}\left\{\sum_{r}F_i^2\hat{\delta}^2(X_i)+\sum_{\bar{r}}.\hat{\delta}^2(X_i)+\sum_{\bar{r}}\left[x_{l(i)}-\hat{\mu}(X_i)\right]^2\right\}$$

3.3. Estimation of V_{MIX}

$$\begin{aligned}V_{MIX}&=E[(\bar{y}-\bar{Y}_U)(\bar{y}_I-\bar{y})]^2\bar{Y}_U=\frac{1}{N}\sum_{U}y_i=\frac{1}{N}\left[\sum_{s}y_i+\sum_{\bar{s}}y_i\right]=\frac{1}{N}\left[\sum_{r}y_i+\sum_{\bar{r}}y_i+\sum_{\bar{s}}y_i\right]\bar{y}-\bar{Y}_U=\frac{1}{n}\left(\sum_{r}y_i+\sum_{\bar{r}}y_i\right)-\frac{1}{N}\left(\sum_{r}y_i+\sum_{\bar{r}}y_i+\sum_{\bar{s}}y_i\right)\\&=\left(\frac{1}{n}-\frac{1}{N}\right)\left(\sum_{r}y_i+\sum_{\bar{r}}y_i\right)-\frac{1}{N}\sum_{\bar{s}}y_i\,\bar{y}_I-\bar{y}=\frac{1}{n}\left(\sum_{r}y_i+\sum_{r}F_iy_i\right)-\frac{1}{n}\left(\sum_{r}y_i+\sum_{\bar{r}}y_i\right)\\&=\frac{1}{n}\left(\sum_{r}F_iy_i-\sum_{\bar{r}}y_i\right)(\bar{y}-\bar{Y}_U)(\bar{y}_I-\bar{y}_i)\\&=\frac{1}{n}\left(\frac{1}{n}-\frac{1}{N}\right)\sum_{r}F_iy_i\left(\sum_{r}y_i+\sum_{\bar{r}}y_i\right)-\frac{1}{nN}\sum_{r}F_iy_i\sum_{\bar{s}}y_i-\frac{1}{n}\left(\frac{1}{n}-\frac{1}{N}\right)\sum_{\bar{r}}y_i\left(\sum_{r}y_i+\sum_{\bar{r}}y_i\right)+\frac{1}{nN}\sum_{\bar{r}}y_i\sum_{\bar{s}}y_i\\&=\frac{1}{n}\left(\frac{1}{n}-\frac{1}{N}\right)\left[\sum_{r}F_iy_i^2+\sum\sum_{i\neq j}y_iy_j-\sum_{r}y_i\sum_{\bar{r}}y_i+\sum_{\bar{r}}y_i\sum_{r}F_iy_i-\sum_{\bar{r}}y_i^2-\sum\sum_{i\neq j}y_iy_j-\frac{1}{nN}\left[\sum_{\bar{s}}y_i\sum_{r}F_iy_i-\sum_{\bar{s}}y_i\sum_{\bar{r}}y_i\right]\right]\end{aligned}$$

$$\begin{aligned}&E[(\bar{y}-\bar{Y}_U)(\bar{y}_I-\bar{y})]\\&=\frac{1}{n}\left(\frac{1}{n}-\frac{1}{N}\right)\left[\sum_{r}F_iE(y_i^2)-\sum_{r}E(y_i)\sum_{\bar{r}}E(y_i)+\sum_{\bar{r}}E(y_i)\sum_{r}F_iE(y_i)-\sum_{\bar{r}}E(y_i^2)\right.\\&\left.-\frac{1}{nN}\left[\sum_{\bar{s}}E(y_i)\sum_{r}F_iE(y_i)-\sum_{\bar{s}}E(y_i)\sum_{\bar{r}}E(y_i)\right]\right]\\&=\frac{1}{n^2}\left(1-\frac{n}{N}\right)\left\{\sum_{r}Fi\delta^2(xi)-\sum_{\bar{r}}\delta^2(xi)+\sum_{\bar{r}}\mu(X_i)\left(\sum_{\bar{r}}x_{l(i)}-\sum_{r}\mu(X_i)\right)\right\}\\&-\frac{1}{nN}\left\{\sum_{\bar{s}}\mu(X_i)\sum_{\bar{r}}\left(x_{l(i)}-\mu(X_i)\right)\right\}\end{aligned}$$

It follows that the unbiased estimator of V_{MIX} is

$$\hat{V}_{MIX}=\frac{1}{n^2}\left(1-\frac{n}{N}\right)\left\{\sum_{r}Fi\hat{\delta}^2(xi)-\sum_{\bar{r}}\hat{\delta}^2(xi)+\sum_{\bar{r}}\hat{\mu}(X_i)\left(\sum_{\bar{r}}x_{l(i)}-\sum_{r}\hat{\mu}(X_i)\right)\right\}-\frac{1}{nN}\left\{\sum_{\bar{s}}\hat{\mu}(X_i)\sum_{\bar{r}}\left(x_{l(i)}-\hat{\mu}(X_i)\right)\right\}$$

The estimator for the Variance is given by
$\hat{V}$=$\hat{V}_{SAM}+\hat{V}_{IMP}+\hat{V}_{MIX}$

$$\begin{aligned}\hat{V}&=\left(\frac{1}{n}-\frac{1}{N}\right)\frac{1}{N-1}\left\{\left(1-\frac{1}{N}\right)\left(\sum_{s}\hat{\delta}^2(X_i)+\sum_{s}[\hat{\mu}(X_i)]^2\right)\right\}+\frac{1}{n^2}\left\{\sum_{r}F_i^2\hat{\delta}^2(X_i)+\sum_{\bar{r}}\hat{\delta}^2(X_i)+\sum_{\bar{r}}\left[x_{l(i)}-\hat{\mu}(X_i)\right]^2\right\}\\&+\frac{1}{n^2}\left(1-\frac{n}{N}\right)\left\{\sum_{r}Fi\hat{\delta}^2(xi)-\sum_{\bar{r}}\hat{\delta}^2(xi)+\sum_{\bar{r}}\hat{\mu}(X_i)\left(\sum_{\bar{r}}x_{l(i)}-\sum_{r}\hat{\mu}(X_i)\right)\right\}-\frac{1}{nN}\left\{\sum_{\bar{s}}\hat{\mu}(X_i)\sum_{\bar{r}}\left(x_{l(i)}-\hat{\mu}(X_i)\right)\right\}\end{aligned}$$

3.4. Estimation of $\mu(X_i)$ and $\delta^2(X_i)$

One of the most common methods in non-parametric regression is the kernel method introduced by Nadaraya-Watson (1964), which is often obtained by using a bandwidth [7]. The kernel estimators with varying bandwidths are specially used to estimate density of the long-tailed and multi-mod distributions. A kernel estimate is introduced for obtaining a non-parametric estimate of a regression function.

Smooth linear estimate of $\mu(X_i)$

A smooth linear estimate of a function $\mu(X_i)$ denoted by $\hat{\mu}(X_i)$ can be written in general form as

$\hat{\mu}(X_i)$=$\sum_{j\epsilon s}w_k(X_i,X_j)Y_j$

Where $w_k(X_i,X_j)$ denotes a smoothing function with a bandwidth parameter k. This bandwidth parameter determines the amount of smoothing to be done. The estimates proposed by Nadaraya (1964) and Watson (1964) associated with kernel functions [7] will be considered.

3.5. Nadaraya-Watson Smooth Estimate of $\mu(X_i)$

Nadaraya (1964) and Watson (1964) independently proposed the following estimate of $\mu(X_i)$.

$$\hat{\mu}_{Nw}(X_i) = \sum_s w\left[\frac{(X_i - X_j)}{k}\right] Y_j / \sum_s w\left[\frac{(X_i - X_j)}{k}\right]$$

where k denotes the bandwidth parameter. w is called the kernel function with the following properties.

1) $w(t) \geq 0 \; \forall \; t$

2) $\int_{-\infty}^{\infty} w(t)dt = 1$

3) $\int_{-\infty}^{\infty} w^2(t)dt < \infty$ [7]

3.6. Smooth Linear Estimate of $\delta^2(x_i)$

Consider $Y_i = \mu(X_i) + \varepsilon_i$ where $E(\varepsilon_i/X_i) = 0$ and $V(\varepsilon_i/X_i) = \delta^2(X_i)$

The estimate of the residual term is given by $\hat{\varepsilon}_i = Y_i + \hat{\mu}(X_i)$

The estimator for the Variance is given by

$$\hat{V} = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{N-1}\left\{\left(1 - \frac{1}{N}\right)\left(\sum_s . \hat{\delta}^2(X_i) + \sum_s [\hat{\mu}(X_i)]^2\right)\right\} + \frac{1}{n^2}\left\{\sum_r F_i^2 \hat{\delta}^2(X_i) + \sum_{\bar{r}} . \hat{\delta}^2(X_i) + \sum_{\bar{r}} \left[x_{l(i)} - \hat{\mu}(X_i)\right]^2\right\}$$
$$+ \frac{1}{n^2}\left(1 - \frac{n}{N}\right)\left\{\sum_r Fi\hat{\delta}^2(xi) - \sum_{\bar{r}} \hat{\delta}^2(xi) + \sum_{\bar{r}} \hat{\mu}(X_i)\left(\sum_{\bar{r}} x_{l(i)} - \sum_r \hat{\mu}(X_i)\right)\right\}$$
$$- \frac{1}{nN}\left\{\sum_{\bar{s}} \hat{\mu}(X_i) \sum_{\bar{r}} \left(x_{l(i)} - \hat{\mu}(X_i)\right)\right\}$$

Where $\hat{\mu}(X_i)$ and $\hat{\delta}^2(X_i)$ are as given above.

The square of the estimate of this residual term $\varepsilon_i, i \in s$ is given by

$$\hat{\varepsilon}^2{}_i = \left(Y_i + \hat{\mu}(X_i)\right)^2 \qquad (6)$$

To smooth (6), we choose a smooth function $w_h(X_i, X_j)$ with a bandwidth parameter h. Using (6), we get $\hat{\delta}^2(X_i) = \sum_{i \in s} w_h(X_i, X_j)\left(Y_i + \hat{\mu}(X_i)\right)^2$ which is a smooth estimate of $V(Y_i/X_i)$

A corresponding NW estimate of $\delta^2(X_i)$ is given by $\hat{\delta}^2_{NW}(X_i) = \frac{\sum_s \left[\frac{(X_i - X_j)}{h}\right]\left(Y_j + \hat{\mu}_{NW}(X_j)\right)^2}{\sum_s \left[\frac{(X_i - X_j)}{h}\right]}$ where h denotes the bandwidth parameter.

4. Simulation Studies

In our simulation study, the performance of the proposed donor estimator was compared with the naïve estimator, Jackknife estimator and bootstrap estimator empirically. In our comparison, two artificial population structures (linear and non-linear), one real population (linear) and two non-response mechanisms were considered. We conducted a simulation study to evaluate the performance of our variance estimator in terms of Relative Bias (RB) and Variance.

The first population (linear population) was generated as follows: 100 data points were generated according to the linear homoscedastic model;

$$y_i = 0.25x_i + l_i \text{ with } l_i \sim N(0, \delta^2) \text{ and } y_i \sim U(0,1)$$

This was done by first generating the auxiliary variables (X_k) values and then the values for Y_k. In the second population structure (non-linear population), 100 data points were generated according to the quadratic homoscedastic model;

$$y_i = 0.5 + 0.25x_i + 1.5x_i^2 + l_i \text{with } l_i \sim N(0, \delta^2) \text{ and } y_i \sim U(0,1)$$

A simple random sample of size 0.225 of the population size was taken without replacement from each population structure. We considered two non response mechanisms which are random and non random non-response.

For a random non-response mechanism, non responses were generated using independent Bernoulli trials with a constant parameter 0.3 representing the probability of non-response.

For a non random non-response mechanism, the sample values were arranged in order of magnitude using Y_k values and then the largest 30% of the values were regarded as missing.

Non responses were generated for each non-response mechanism. To compensate for the missing values, nearest neighbor imputation was performed. After imputation, the four variance estimates $\hat{V}_{JKN}, \hat{V}_{BTS}, \hat{V}_{ORD}, and\ \hat{V}_{DON}$ were calculated. The experiment was repeated 1000 times independently and the average value of each value was got. In the case of bootstrap estimator, 1000 bootstrap iterations were used. In the instance of donor estimator, we used the bandwidth parameter that minimized the mean squared error and satisfied Silver-man's (1986) condition.

$$\frac{\delta}{4n^{\frac{1}{5}}} \leq h \leq \frac{3\delta}{2n^{\frac{1}{5}}} \text{ where } \delta = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

The Epanechnikov's kernel function $\frac{3}{4}(1 - x^2)$ was used since it gives optimal solutions.

The performances of estimators were assessed using two criteria: the relative bias and the Variance. The relative bias of the estimators is calculated as follows:

$\boldsymbol{RB}=\frac{\mathbf{1}}{\mathbf{1000}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\mathbf{1000}}\left[\frac{\hat{V}_i-V(\bar{y}')}{V(\bar{y}')}\right]$ where $\boldsymbol{V}(\bar{\boldsymbol{y}}')=\frac{\mathbf{1}}{\mathbf{1000}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\mathbf{1000}}[\bar{\boldsymbol{y}}_{\boldsymbol{i}}'-\bar{\boldsymbol{y}}]^{\mathbf{2}}$, $\bar{y}_i'$ is the value of $\bar{y}'$ for the i^{th} experiment and $\hat{V}_i$ represents the value of the estimator for the i^{th} experiment.

5. Results

The results were then tabulated showing the performance of the estimators in terms of relative bias and Variance. Three populations were analyzed with each population having two tables. One table shows the case when the non-response mechanism is random while the other shows the case when the non-response mechanism is non-random.

a) Case when population is linear.

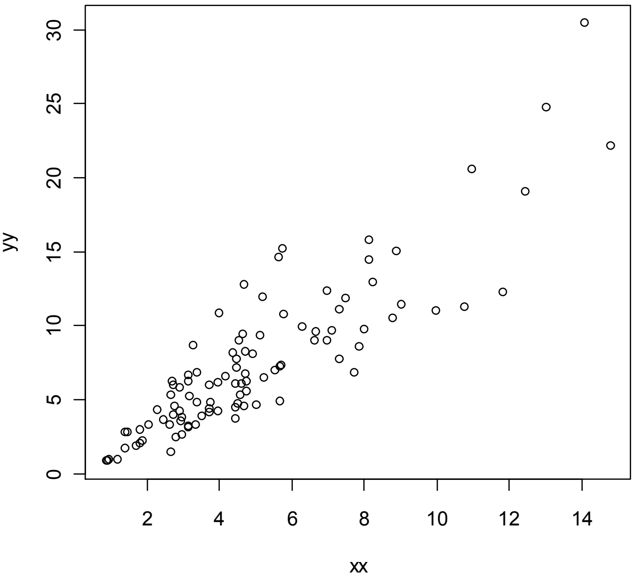

Fig. 1. Graph of Survey variables against design variables.

From Table 1, the naïve estimator has the smallest Variance followed by Jackknife while our proposed estimator performs better than Bootstrap. The proposed estimator has the highest relative bias followed by the naïve estimator while Jackknife and Bootstrap seems to do well in terms of relative bias.

b) Case when population is real

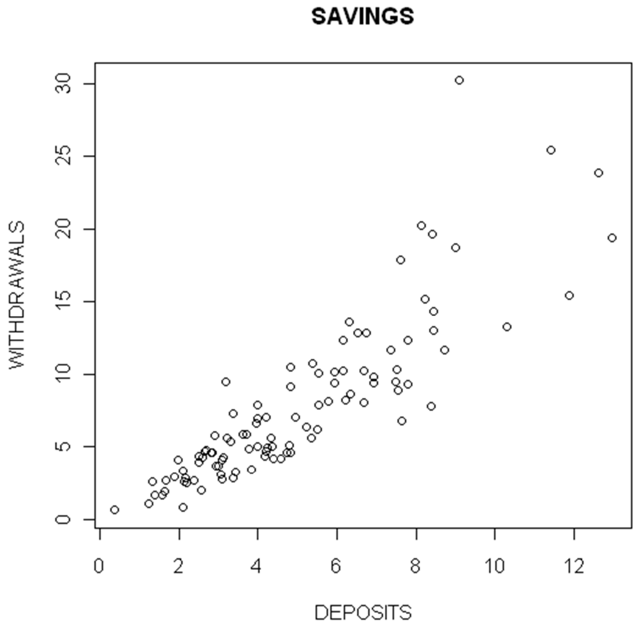

Fig. 2. Graph of withdrawals against deposits.

The results of Table 2 are similar to those of Table 1. This implies that whether the population is real or artificial, as long as it is linear, the estimators behave in the same way.

c) Case when population scatter is non- linear.

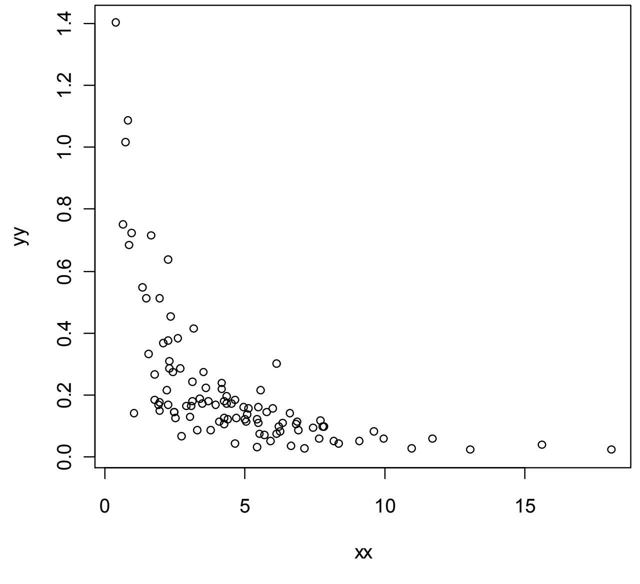

Fig. 3. Graph of Survey variables against design variables.

Table 1. Variance, Relative bias and M.S.E for the four variance estimators.

	Variance estimator	$\hat{V}_{ORD}$	$\hat{V}_{JKN}$	$\hat{V}_{BTS}$	$\hat{V}_{DON}$
	Variance	0.4439699	0.5691921	3.202729	2.293254
Random Non-Response	Relative bias	-0.2963652	-0.0979040	0.009863698	2.608374
	M.S.E	0.5318022	0.5787773	3.2028263	9.0967735
	Variance	0.6639061	0.8511616	4.67159	3.122172
Non Random Non-Response	Relative bias	-0.2690474	-0.0628812	0.01366792	2.405759
	M.S.E	0.7362926	0.8551156	4.6717768	8.909848

Table 2. Variance, Relative bias and M.S.E for the four variance estimators.

	Variance estimator	$\hat{V}_{ORD}$	$\hat{V}_{JKN}$	$\hat{V}_{BTS}$	$\hat{V}_{DON}$
Random Non-Response	Variance	0.1625649	0.2084165	6.557316	1.512266
	Relative bias	-0.7205724	-0.6417595	0.03275899	1.566941
	M.S.E	0.68178948	0.6202718	6.5583892	3.9675701
Non Random Non-Response	Variance	0.1149530	0.1473756	5.396808	1.245422
	Relative bias	-0.7506162	-0.6802772	0.01731654	1.671923
	M.S.E	0.678377679	0.610152668	5.397107863	4.040748518

Table 3. Variance, Relative bias and M.S.E for the four variance estimators.

	Variance estimator	$\hat{V}_{ORD}$	$\hat{V}_{JKN}$	$\hat{V}_{BTS}$	$\hat{V}_{DON}$
RandomNon-Response	Variance	0.0001596323	0.0002046568	0.03384162	0.001060072
	Relative bias	-0.7516874	-0.6816506	0.0002926527	0.6284259
	M.S.E	0.565193579	0.464852197	0.033841705	0.395979183
Non Random Non-Response	Variance	9.254327e-05	0.0001186452	0.010714	0.0009753563
	Relative bias	-0.7371964	-0.6630724	-0.0008681272	1.728175
	M.S.E	0.543551075	0.439783652	0.010714753	2.987564187

According to Table 3, our proposed estimator performs better than the bootstrap estimator while the naïve and Jackknife estimators have the smallest Variance. Bootstrap seems to be the best in terms of relative bias while our proposed estimator has the highest relative bias.

Discussion of the results

Considering the above three tables where we were comparing the estimators when the popuation is linear or non linear, naïve estimator seems to have the smallest Variance followed by Jackknife estimator while our proposed estimator alternates with bootstrap. In non-linear population, our proposed estimator performs better in terms of Variance than bootstrap. It is also noted that the Variance and relative bias of the four estimators have close numerical values implying that they are all valid.

It is worth noting that donor imputation may not be the most efficient imputation method in any specific scenario. Nevertheless, it is quite a popular imputation method in surveys due to its practical advantages. Therefore it is useful to develop variance estimation methods that take donor imputation into account.

6. Conclusion

The simulation study examined the performance of four variance estimators. Two population structures (linear and non-linear), and two non-response mechanisms were considered. Simulation study was conducted to evaluate the performance of the variance estimators in terms of Relative Bias (RB) and Variance. It was noted that the variance and the relative bias of the 4 estimators have very close numerical values. Hence all are valid and work well in simulation study. We have proposed a variance estimation method for any type of donor imputation. It is valid and was shown to work well in a simulation study. The variance of the proposed estimator is small and its relative bias is also small.

Thus, it is useful to develop a variance estimation method that takes donor imputation into account. Its main drawback is that it depends on the validity of an imputation model. This is also a characteristic of the methods for NN imputation. Two key issues with any variance estimation method that relies on an imputation model are the appropriate choice of auxiliary variables for donor selection and the estimation of the model mean μ_k and variance δ_k^2 given the chosen auxiliary variables. Auxiliary variables should be associated with the variable of interest so as to ensure that the conditional model bias remains small [1].

References

[1] Beaumont, J. F. and Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *Canadian Journal of Statistics, 96* (4), 917-932.

[2] Beaumont, J. F. and Bocci, C. (2007). Variance estimation when donor imputation is used to fill in missing values. Proceedings of the Third International Conference on Establishment Surveys, Montréal.

[3] F. W. Scholz (2007). The Bootstrap Small Sample Properties. *University of Washington*

[4] Brick, J. M., Kalton, G. and Kim, J. K. (2004). Variance estimation with hot deck imputation using a model. *Survey Methodology, 30,* 57-66.

[5] Chen, J. and Shao, J. (2000). Nearest neighbour imputation for survey data. *Journal of Official Statistics*, *16*, 113–131.

[6] Chen, J. and Shao, J. (2001). Jackknife variance estimation for nearest neighbor imputation. *Journal of the American statistics Association*, 96, 260-269.

[7] Fuller, A. and Kim, J. K. (2000). Hot Deck Imputation for the Response Model. Vol. 31, No. 2, pp. 139-149 Statistics Canada, Catalogue No. 12-00 Statistica Sinica 10, 1153-1169.

[8] Jae Kwang Kim (2001). Variance Estimation After Imputation. *Statistics Canada, Catalogue* No. 12001Vol. 27, No. 1, pp. 7583

[9] Njenga, E. G. (1990). Robust estimation of the regression coefficients in complex surveys. (*Doctoral dissertation, 1990).*

[10] Rao, J. N. K. and Shao, J. (1992). Jackknife Variance Estimation with Survey Data under Hot Deck Imputation. *Biometrika, 79*, 811-822.

[11] Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. *John Wiley & Sons, New York*.

[12] Shao, J., and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association, 94*, 254-265.

[13] Shao, J. and Tu, D. (1995). The Jackknife and Bootstrap. *New York: Springer-Verlag*.