Science Journal of Applied Mathematics and Statistics
2021; 9(5): 126-132
http://www.sciencepublishinggroup.com/j/sjams
doi: 10.11648/j.sjams.20210905.12
ISSN: 2376-9491 (Print); ISSN: 2376-9513 (Online)

Estimation of Finite Population Total in Presence of Missing Values in Two-Phase Sampling

Kemei Anderson Kimutai[1, *], Christopher Ouma Onyango[2], Mike Wafula[2]

[1]Department of Mathematics, Kiriri Women's University of Science and Technology, Nairobi, Kenya
[2]Department of Mathematics, Statistics & Actuarial Science, Kenyatta University, Nairobi, Kenya

Email address:
kimkm04@gmail.com (K. A. Kimutai)
[*]Corresponding author

To cite this article:
Kemei Anderson Kimutai, Christopher Ouma Onyango, Mike Wafula. Estimation of Finite Population Total in Presence of Missing Values in Two-Phase Sampling. *Science Journal of Applied Mathematics and Statistics*. Vol. 9, No. 5, 2021, pp. 126-132.
doi: 10.11648/j.sjams.20210905.12

Received: September 16, 2021; **Accepted**: November 9, 2021; **Published**: November 17, 2021

Abstract: Missing data is a real problem in many surveys. To overcome the problems caused by missing data, partial deletion and single imputation methods among others have been proposed. However, problems such as discarding usable data, inaccuracy in reproducing known population parameters and standard errors are associated with them. In ratio, regression and stochastic imputation, it is assumed that there is a variable with complete cases that can be used as a predictor in estimating missing values in the other variable(s) and the relationship between the dependent and independent variable(s) is linear. This might not always be the case. To overcome these problems accompanied to stochastic and regression estimation, two-phase sampling and nonparametric model-based estimation were employed in this research. Estimator of population total in two-phase sampling was modified. The variance of estimator developed by Hidiroglou, Haziza and Rao was used to compare the performance of the proposed non-parametric model-based imputation in reproducing well known population total and standard errors compared to mean, regression and stochastic methods of imputation. The data was simulated and analyzed using R-statistical Software. The empirical study revealed that non-parametric model-base imputation method is better in reproducing both known population total and standard error.

Keywords: Finite Population Total, Missing Values, Two-phase Sampling

1. Introduction

The main goal of all surveys is to obtain a more precise and reliable information about population characteristics under study. Both census and sampling methods can provide this information. However, missing data is associated with almost every survey and dealing with them has been a subject of research from time to time. This is because missing data creates a problem in data analysis and also makes estimates of the population characteristics under study to be biased, [1, 2]. According to [3, 4], the consequence of missing data on quantitative research can be severe, leading to biased estimates of parameters, loss of information, diminished statistical power, increased standard errors, and weakened generalizability of outcomes. Some factors that may result in missing data include non-response, dropout of a participant before the end of a longitudinal study, improper data collection, malfunctioning of equipment, bad weather condition or mistakes are made in data entry [3, 5].

Previously, various methods of dealing with missing data have been proposed. However, each of these methods has its own setbacks. As earlier mentioned, regression and stochastic imputation methods rely on the assumption that the relationship between study variable and predictor variable is linear. Nonetheless, there may be no such a relationship. It is also believed that there is a variable with complete cases that can be used in predicting missing values in the other variable(s), which might not always be the case. In this research, we provide a solution to the two cases by employing both nonparametric model-based estimation and two-phase sampling. The reason for adopting nonparametric model-based regression is because unlike simple regression and stochastic imputation, nonparametric model-based

regression does not restrict the relationship between the predictor and outcome variables to any form of an equation but it entirely leaves it to be determined by the data [2, 6, 7].

2. Methods of Dealing with Missing Data

Missing data can be dealt with in various ways. One can decide to (i) exclude the entire record from analysis if any value is missing in a given case, (ii) reduce available data so that a dataset has no missing values and estimate each element in the inter-correlation matrix using all available data, [8, 9] (iii) replace missing data with its probable value based on other existing information or (iv) employ maximum likelihood method. The first two techniques are classified as partial deletion techniques (the earlier is known as listwise deletion or complete case analysis and the latter being pairwise deletion or available case analysis). And the third method is referred to as imputation. Imputation is broadly classified into two; single imputation in which the missing values are replaced only once and multiple imputations in which missing values are replaced more than once [10, 11].

Single imputation consists of imputation methods such as, last observation carried forward, mean, ratio, regression, stochastic and pattern match. We review mean, regression, ratio and stochastic imputation methods because we shall compare their performance in terms of well reproducing known population total estimate and its variance estimate to those of the proposed nonparametric model based method of estimating missing values [12].

2.1. Mean Imputation

In mean imputation, it is assumed that the mean of the available variable values best estimates any of the missing value of that missing values with the mean of the variable. Thus, all the missing values are replaced with the mean of the available values. Mean imputation has its own advantages such as; simplicity, it does not discard any available data and does not change the mean of the variable. However, it has been cautioned that mean substitution should not be used since it has limitations of overestimating sample size, variance is underestimated, correlations are negatively biased and the distribution of new values is an incorrect representation of the population values because the shape of the distribution is distorted by adding values equal to the mean [13, 14].

2.2. Regression Imputation

As the name suggests, regression imputation replaces missing values with it estimate obtained from a regression equation.

$$y_i = \hat{a} + \hat{b}x_i \tag{1}$$

A data set is first reduced to complete cases and the parameters of a regression line are estimated. Once the model has been obtained, missing data are estimated using the model together with the other variable(s) available information. However, apart from the already mentioned drawbacks of regression estimation, the problem of underestimation of the standard variance still exists (though there is an improvement compared to mean imputation). This is because by substituting a value that is perfectly predictable from other variables, no new information is added but the sample size is increased thus the standard error is reduced [8, 15].

2.3. Stochastic Regression

Stochastic regression uses the same basic procedure as standard regression imputation, so the regression coefficients are first obtained as in the case of regression imputation. However, for stochastic imputation an additional residual term Z_i that is introduced to correct the problem of underestimation of variance in regression imputation. The equation is given by

$$y_i = \hat{a} + \hat{b}x_i + z_i \tag{2}$$

This residual term is a random value that is normally distributed with a mean of zero and a variance equal to the residual variance from the regression model, [15]. After stochastic model was obtained, missing values were estimated and substituted in the dataset and analysis was done as if there were no missing value. Since it starts with the same procedure as regression imputation, the assumption of linearity between the outcome variable and predictor variable(s) and existence of a variable with complete cases to be used as predictor variable(s) which affects regression imputation are also brought in.

3. Proposed Estimator and Its Asymptotic Properties

Let $\pi_i^{(1)} = pr(i \in A_1)$ and $\pi_{ij}^{(1)} = pr(i, j \in A_1)$ be the first order and second order inclusion probability in the first-phase sample. Also let $\pi_{i/A_1}^{(2)} = pr(i \in i/A_1)$ and $\pi_{ij/A_1}^{(2)} = pr(i, j \in A_2/A_1)$ be the conditional first order and second order inclusion probability in second-phase sample given that the unit is in the first-phase sample. Since the first-phase sample units are selected by SRSWOR,

$$\pi_i^{(1)} = {}^{n}/_{N} \text{ and } \pi_{ij}^{(1)} = \frac{n(n-1)}{N(N-1)} \tag{3}$$

Calculation of second-phase inclusion probabilities is a bit complicated. When $n' > 1$ and some values x_i are extremely large, the inclusion probabilities for those elements are greater than 1. Simply, for the large units $n'\, x_i/\sum_{i=1}^{n} x_i > 1$, may be encountered. One possibility of solving this was provided by [16]. The research stated that for large units set $\pi_i = 1$. This implies that those elements are taken with certainty and for the remaining elements, π_i is set proportional to the size of x_i. This results in a loss in precision as a result of purposively selecting some units. For less extremely large units, second-phase inclusion probabilities are given by

$$\pi_{i/A_1}^{(2)} = \frac{n\prime x_i}{\sum_{i=1}^{n} x_i} \quad (4)$$

After obtaining the first order inclusion probability into the second phase sample, the second-order inclusion probability are is estimated as [17]

$$\pi_{ij/A_1}^{(2)} = \frac{2(n\prime-1)\pi_{i/A_1}^{(2)}\pi_{j/A_1}^{(2)}}{2n\prime-\pi_{i/A_1}^{(2)}-\pi_{j/A_1}^{(2)}} \quad (5)$$

3.1. Proposed Estimator

The $\pi^* -$ estimator of population total is given in [18] as

$$\hat{Y} = \sum_{i\in A_2} \frac{y_i}{\pi_i^{(1)}\pi_{i/A_1}^{(2)}} = \sum_{i\in A_2} \frac{y_i}{\pi_i^*} \quad (6)$$

Where $\pi_i^* = \pi_i^{(1)}\pi_{i/A_1}^{(2)} = \frac{n}{N}\pi_{i/A_1}^{(2)}$ since the first phase sample is selected by SRSWOR.

Thus equation (6) becomes

$$\hat{Y} = \frac{N}{n}\sum_{i\in A_2} \frac{y_i}{\pi_{i/A_1}^{(2)}} \quad (7)$$

Suppose that there are m units whose $y-$values are available in the second-phase sample and $y_i's$ are generated from a model relating $x_i's$ to $y_i's$ in the second-phase sample is given by $y_i = m(x_i) + e_i$, then for a data point $x = x_j$ whose study variable value is missing, $y_j$values is estimated as

$$\hat{y}_j = \hat{m}(x_j) + e_j \quad (8)$$

Using above information and (8) in (7) we get

$$\hat{Y} = \frac{N}{n}\left[\sum_{i=1}^{m} \frac{y_i}{\pi_{i/A_1}^{(2)}} + \sum_{j=m+1}^{n\prime} \frac{\hat{y}_j}{\pi_{j/A_1}^{(2)}}\right] = \frac{N}{n}\left[\sum_{i=1}^{m} \frac{y_i}{\pi_{i/A_1}^{(2)}} + \sum_{j=m+1}^{n\prime} \frac{\hat{m}(x_j)+e_j}{\pi_{j/A_1}^{(2)}}\right] \quad (9)$$

This is the proposed estimator of population total in presence of missing values under two-phase sampling scheme. $m(x_i)$ is the conditional expectation of Y given X. The only property known for $m(x_i)$ is that it is smooth and continuous. e_i is error random variable which is normally distributed with mean 0 and variance $v(x)$. Also $v(x)$ is strictly positive and smooth.

Nadaraya [19] and Watson [20] estimated the mean function $m(x_j)$ as a locally weighted average using a kernel weight function. In their case, they estimated the non-sampled part of the population using model (8), available auxiliary information and the sampled units. Here, m available data for the study variable, auxiliary information (which were availed in the first phase sample), and the model (3.6) are used to estimate the missing values of the study variable. The estimate of the mean function is given by

$$\hat{m}(x_j) = \sum_{i=1}^{m} w(x_i)y_i\,, w(x_i) = \frac{k_b(x_i-x_j)}{\sum_{i=1}^{m} k_b(x_i-x_j)} \quad (10)$$

Where $w(x_i)$ is the weight of the i^{th} unit of study variable value being available in the second-phase sample. Furthermore, $\sum_{i=1}^{m} w(x_i) = 1$ and k is a kernel function with $k_b(u) = {}^1\!/_b\, k\left({}^u\!/_b\right)$ which satisfies the following properties

(a) $\int k(u)du = 1$ (b) $\int uk(u)du = 0$

(c) $\int u^2k(u)\,du = k_2 < \infty$ (d) $k(u) = k(-u)$

b is the shaping parameter called bandwidth which determines the amount of smoothing done to the data.

3.2. Asymptotic Unbiasedness

From equation (9) we have

$$E_1E_2(\hat{Y}) = \frac{N}{n}E_1E_2\left[\sum_{i=1}^{m} \frac{y_i}{\pi_{i/A_1}^{(2)}} + \sum_{i=m+1}^{n\prime} \frac{\hat{m}(x_j)+\hat{e}_j}{\pi_{j/A_1}^{(2)}}\right] \quad (11)$$

Consider

$$E\left[\hat{m}(x_j) - m(x_j)\right] = E\left[\sum_{i=1}^{m} w(x_i)m(x_i) - m(x_j)\right] = E\left[\frac{\sum_{i=1}^{m}\frac{1}{mb}k(x_i-x_j)m(x_i)}{\sum_{i=1}^{m}\frac{1}{mb}k\left(\frac{x_i-x_j}{b}\right)} - m(x_j)\right]$$

$$= E\left[\sum_{i=1}^{m}\frac{1}{mb}\left\{k\left(\frac{x_i-x_j}{b}\right)\left[m(x_i)-m(x_j)\right]\left[\hat{d}_s^{-1}(x_j)\right]\right\}\right] \quad (12)$$

Where $\hat{d}_s(x_i) = \sum_{i=1}^{m}\frac{1}{mb}k\left(\frac{x_i-x_j}{b}\right)$

The expectation in eq. (12) reduces to

$$E\left(\left[\hat{m}(x_j) - m(x_j)\right]\Big/_{x_j}\right) = \frac{b^2\beta(x_j)\frac{k_2}{2} + O_p\left(b^3+\left(\frac{b}{m}\right)^{\frac{1}{2}}\right)}{d_s(x_j) + b^2\frac{k_2}{2}d_s''(x_j) + O_p\left(b^3+\left(\frac{1}{bm}\right)^{\frac{1}{2}}\right)} \quad (13)$$

From (13) we see that as $b \to 0$ and $\to \infty$ $E\left(\hat{m}(x_j) - m(x_j)\right) \to 0$. That is, $\hat{m}(x_j)$ converges in probability to $m(x_j)$ as

$b \longrightarrow 0$ and $mb \to \infty$. Simply, $\hat{y}_j \to y_j$ for $b \longrightarrow 0$ and $mb \to \infty$. Therefore (11) becomes

$$E_1E_2(\hat{Y}) = \frac{N}{n}E_1E_2\left[\sum_{i=1}^{m}\frac{y_i}{\pi_{i/A_1}^{(2)}} + \sum_{i=m+1}^{n'}\frac{\widehat{m}(x_j)+\hat{e}_j}{\pi_{j/A_1}^{(2)}}\right] = \frac{N}{n}E_1E_2\left[\sum_{i=1}^{m}\frac{y_i}{\pi_{i/A_1}^{(2)}} + \sum_{i=m+1}^{n'}\frac{\hat{y}_j}{\pi_{j/A_1}^{(2)}}\right]$$

$$\approx \frac{N}{n}E_1E_2\left[\sum_{i=1}^{m}\frac{y_i}{\pi_{i/A_1}^{(2)}} + \sum_{i=m+1}^{n'}\frac{y_j}{\pi_{j/A_1}^{(2)}}\right] = \frac{N}{n}E_1E_2\left[\sum_{i=1}^{n'}\frac{y_i}{\pi_{i/A_1}^{(2)}}\right] = \frac{N}{n}E_1\sum_{i=1}^{n'}\frac{E_2(y_i)}{\pi_{i/A_1}^{(2)}} = \frac{N}{n}E_1\sum_{i=1}^{n'}\frac{y_i}{\pi_{i/A_1}^{(2)}}\Pr(i \in A_2/A_1)$$

$$= NE_1\left[\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{\pi_{i/A_1}^{(2)}}\pi_{i/A_1}^{(2)}\right] = \frac{N}{n}\left[\sum_{i=1}^{n}E_1(y_i)\right] = \frac{N}{n}\sum_{i=1}^{N}Y_i\Pr(i \in A_1) = \frac{N}{n}\sum_{i=1}^{N}Y_i\pi_i^{(1)} = \frac{N}{n}\sum_{i=1}^{N}Y_i\frac{n}{N} = Y$$

Hence for $b \to 0$ and $mb \to \infty, \hat{Y}$ is asymptotically unbiased estimate of the population total of the study variable in presence of missing data. This result also indicates that for $m \to \infty$ the second-phase sample size n' must be sufficiently large as well.

3.3. Variance of the Estimator

$$Var(\hat{Y}) = E\left[var(\hat{Y}_2/A_1)\right] + Var\left[E(\hat{Y}_2/A_1)\right] = E\left[var(\hat{Y}_2/A_1)\right] + Var\left[\hat{Y}_1\right] \tag{14}$$

Using Sen-Yates-Grundy variance estimator developed in [21] $var(\hat{Y}_2/A_1)$ is estimated as

$$\widehat{var}(\hat{Y}_2/A_1) = Var\left(\frac{N}{n}\sum_{i\in A_2}\frac{y_i}{\pi_{1/A_1}^{(2)}}\right) = \left(\frac{N}{n}\right)^2\left\{\sum\sum_{i<j\in A_2}\left(\frac{\pi_{i/A_1}^{(2)}\pi_{j/A_1}^{(2)}-\pi_{ij/A_1}^{(2)}}{\pi_{ij/A_1}^{(2)}}\right)\left(\frac{y_i}{\pi_{i/A_1}^{(2)}}-\frac{y_j}{\pi_{j/A_1}^{(2)}}\right)^2\right\} \tag{15}$$

But from equation (5) we have

$$\frac{\pi_{i/A_1}^{(2)}\pi_{j/A_1}^{(2)}-\pi_{ij/A_1}^{(2)}}{\pi_{ij/A_1}^{(2)}} = \frac{2-\pi_{i/A_1}^{(2)}-\pi_{j/A_1}^{(2)}}{2(n'-1)} \tag{16}$$

Hence substituting (16) in (15) we get

$$\widehat{var}(\hat{Y}_2/A_1) = \frac{N^2}{2n^2(n'-1)}\left\{\sum\sum_{i<j\in A_2}\left(2-\pi_{i/A_1}^{(2)}\pi_{j/A_1}^{(2)}\right)\left(\frac{y_i}{\pi_{ij/A_1}^{(2)}}-\frac{y_j}{\pi_{ij/A_1}^{(2)}}\right)^2\right\} \tag{17}$$

Here the dataset is made up of both available values and estimates of missing values thus $i, j = 1, 2, \ldots, n' \in A_2$, i.e. can take any value in the second-phase sample.

The above variance estimate given in equation (17) is conditionally unbiased for $var(\hat{Y}_2/A_1)$ and consequently unbiased for $E\left[var(\hat{Y}_2/A_1)\right]$. This implies that

$$E\left[var(\hat{Y}_2/A_1)\right] = \frac{N^2}{2n^2(n'-1)}\left\{\sum\sum_{i<j\in A_2}\left(2-\pi_{i/A_1}^{(2)}-\pi_{j/A_1}^{(2)}\right)\left(\frac{y_i}{\pi_{i/A_1}^{(2)}}-\frac{y_j}{\pi_{j/A_1}^{(2)}}\right)^2\right\} \tag{18}$$

Next, consider $Var\left[\hat{Y}_1\right]$ in equation (14). Also using Sen-Yates-Grundy variance estimator,

$$Var\left[\hat{Y}_1\right] = \sum\sum_{i<l\in A_1}\left(\frac{\pi_i^{(1)}\pi_j^{(1)}-\pi_{ij}^{(1)}}{\pi_{ij}^{(1)}}\right)\left(\frac{y_i}{\pi_i^{(1)}}-\frac{y_j}{\pi_j^{(1)}}\right)^2 \tag{19}$$

Since the first phase sample is selected by SRSWOR,

$$\pi_i^{(1)} = \pi_j^{(1)} = \frac{n}{N} \; and \; \pi_{ij}^{(1)} = \frac{n(n-1)}{N(N-1)} \tag{20}$$

Therefore

$$\frac{\pi_i^{(1)}\pi_j^{(1)}-\pi_{ij}^{(1)}}{\pi_{ij}^{(1)}} = \frac{N-n}{N(n-1)} \text{ and } \left(\frac{y_i}{\pi_i^{(1)}}-\frac{y_j}{\pi_j^{(1)}}\right)^2 = \left(\frac{N}{n}\right)^2(y_i-y_j)^2 \tag{21}$$

Substituting equations (20) and (21) in eq. (19) we obtain

$$Var[\hat{Y}_1] = \sum\sum_{i<j\in A_1}\left(\frac{N-n}{N(n-1)}\right)\left(\frac{N}{n}\right)^2(y_i - y_j)^2 = \frac{N(N-n)}{n^2(n-1)}\sum\sum_{i<j\in A_1}(y_i - y_j)^2 \quad (22)$$

But y values are not obtained in the first-phase sample A_1 therefore (22) is estimated using the values in the second-phase sample A_2 as

$$\widehat{Var}[\hat{Y}_1] = \frac{N(N-n)}{n^2(n-1)}\sum\sum_{i<j\in A_2}\frac{(y_i-y_j)^2}{\pi^{(2)}_{ij/A_1}} \quad (23)$$

Also $i, j = 1,2,\dots,n' \in A_2$

Combining (18) and (23), the total variance of the population total estimate is obtained as

$$\widehat{Var}(\hat{Y}) = \frac{N^2}{2n^2(n'-1)}\left\{\sum\sum_{i<j\in A_2}\left(2 - \pi^{(2)}_{i/A_1} - \pi^{(2)}_{ij/A_1}\right)\left(\frac{y_i}{\pi^{(2)}_{i/A_1}} - \frac{y_j}{\pi^{(2)}_{j/A_1}}\right)^2\right\} + \frac{N(N-n)}{n^2(n-1)}\left\{\sum\sum_{i<j\in A_2}\frac{(y_i-y_j)^2}{\pi^{(2)}_{ij/A_1}}\right\} \quad (24)$$

3.4. Mean Square Error of the Proposed Estimator

The mean square error of a point estimator $\hat{\theta}$ of a parameter θ is

$$MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right] = Var(\hat{\theta}) + \left[Bias(\hat{\theta})\right]^2 \quad (25)$$

See [22, 23].

From equation (25) the mean square error of the proposed estimator $\hat{Y}$ is

$$MSE(\hat{Y}) = Var(\hat{Y}) + \left[Bias(\hat{Y})\right]^2$$

Since $\hat{Y}$ is asymptotically unbiased for $Y, MSE(\hat{Y})$ of $\hat{Y}$ is equivalent to the variance given in equation (22).

4. Results and Discussions

Using R Statistical Software, $N = 1000$ values of both auxiliary and study variable were generated. A first-phase sample of $n = 600$ units was drawn by simple random sampling without replacement and auxiliary variable information was obtained. None of the auxiliary variable values (x_i) were overly large and thus the first-order and second-order inclusion probability proportional to size for the second-phase sample units were calculated respectively as given by equations (4) and (5).

After obtaining the first-order inclusion probabilities, second-phase sample of sizes 100, 200,300 and 400 were drawn with probabilities proportional to size x_i from the first phase sample and both x_i and y_i are measured. Using this samples, population total estimate and its variance when there were no missing values are obtained.

The second phase samples were then subjected some conditions so that a proportion of some study variable values were missing. The missing values were then imputed the using the earlier discussed methods and The population total estimate in each case. For proposed model-based imputations, Gaussian kernel defined as $k(u) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}t^2}$ was used whose optimal bandwidth as given by [24] is $b_{opt} = 1.06\hat{\sigma}\left(m^{-1/5}\right)$. $\hat{\sigma} = \min\left(S, {}^{R}/_{1.34}\right)$, where $S = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2}$ and R is the inter-quartile range of the auxiliary data. Variance of these population total estimates were then calculated as given by equations (24). The results are as shown in the tables below.

4.1. Population Mean Estimates

Table 1. *Show the population total estimates.*

Second Phase Sample Size	No Missing Values	Mean Imputation	Regression Imputation	Stochastic Imputation	Model-based Imputation
100	76,474	78,628.66	74,129.14	74,285.05	77,523.86
200	75,853.49	77,136.18	74,588.65	75,094.61	75,137.49
300	75,004.62	73,976.80	75,321.38	75,719.28	75,201.04
400	74,956.23	75,634.71	75,024.92	75,101.00	75,001.57

In order to get an insight on the best method, the difference between the population total estimates under a given imputation and when there are no missing values as obtained. The results are as shown.

Table 2. *Show the difference in population mean estimates as disused above.*

Second Phase Sample Size	Mean Imputation	Regression Imputation	Stochastic Imputation	Model-based Imputation
100	2,154.66	-2,344.86	-2,188.95	1,049.86
200	1,282.69	-1,264.84	-758.88	-716.00
300	-1,027.82	316.76	714.66	196.42
400	678.48	68.69	144.77	45.34

From table 2, the differences between the population total estimates under model-based imputation and when there were no missing values were smaller in all sample sizes compared to mean, regression and stochastic imputations. It also can be clearly seen that estimates improve as the second phase sample size increase since the difference reduce as the sample size increase.

4.2. Variance of the Population Total Estimates

Table 3. *Show the variance of the population total estimates.*

Second Phase Sample Size	No Missing Values	Mean Imputation	Regression Imputation	Stochastic Imputation	Model-based Imputation
100	287625	186109	248789	269873	270427
200	160405	107244	137380	148177	151407
300	116214	85700.6	100733	114195	114915
400	97877.5	83510.7	94739.2	96530.7	96859.4

These variances were plotted against the sample sizes as shown in figure 1. As per the figure, the variance of the estimates obtained by the model-based imputation is closer to the variance of estimates obtained when there were no missing values in all sample sizes. It is followed by stochastic, regression and mean imputation in that order. As the sample size increase, the reproduction of well-known variance improves.

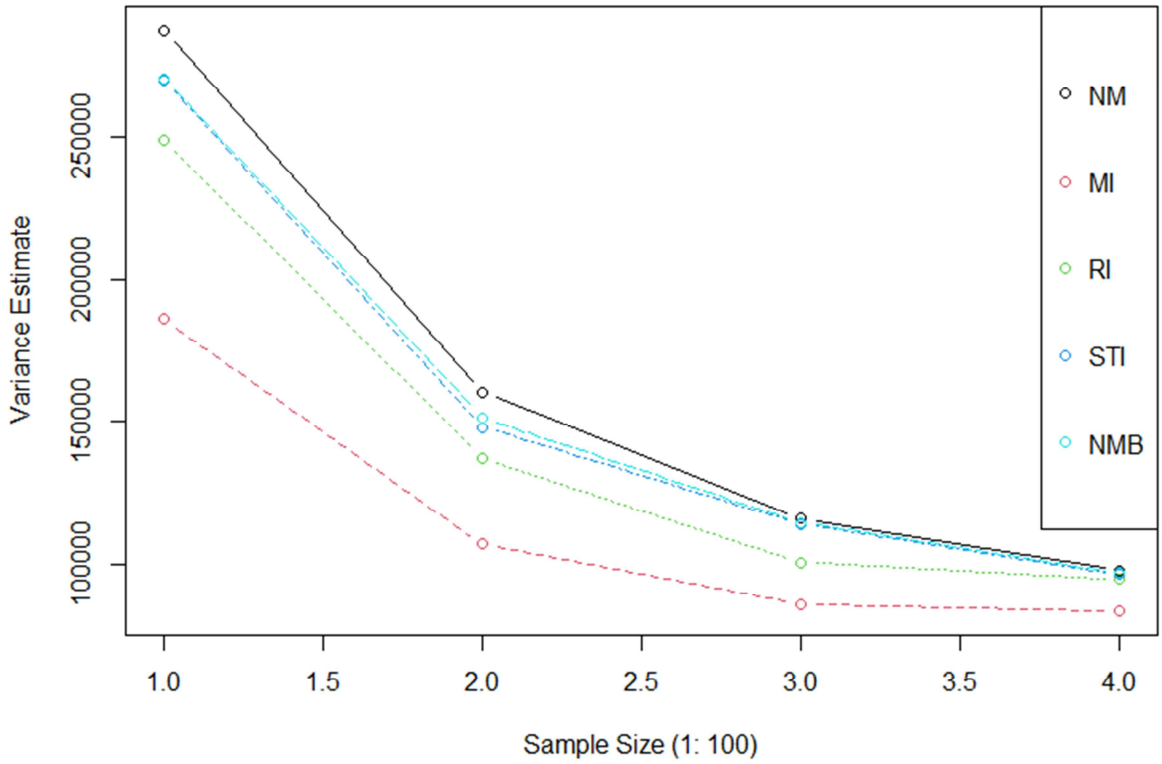

Figure 1. *Graph of Variance estimate against Sample Size.*

5. Conclusion and Recommendations

5.1. Conclusion

The empirical study compared the performance of mean, regression, stochastic and non-parametric model-based imputation in terms of reproducing population total estimates and variance that was obtained when there were no missing values. Model-based imputation was found to be best in reproducing population total estimate obtained when there were no missing values in all of the sample sizes considered compared to mean, regression and stochastic imputation. For the case of variance estimates, non-parametric model-based imputation performed better in reproducing variance estimate of population total estimate obtained when there were no missing values compared to mean, regression and stochastic imputation methods. Based on the results, non-parametric model-based imputation is recommended, especially if the relationship between variables is not linear.

5.2. Recommendations

In this paper, it is assumed that the missing data mechanism is MCAR. A case when missing data mechanism is either MNAR or MAR (see [1, 3] for additional

understanding on missing data mechanisms) may be considered. Also, a single auxiliary and a single study variable was considered but, in most case, survey variable may be determined by more than one auxiliary variable. Therefore, a case where multi-auxiliary information and non-parametric regression model are used in estimating missing values need to be investigated.

References

[1] Cali C., Rachel M. K., Richard F. and Christopher V. H. (2019). Dealing with Missing Data: A Comparative Exploration of Approaches Using the Integrated City Sustainability Database. *Urban Affairs Review,* Vol. 55 (2), 591–615.

[2] Bii N. K., Onyango C. O. and Odhiambo J. (2020). Estimation of a Finite Population Mean under Random Nonresponse Using Kernel Weights. *Journal of Probability and Statistics*, vol. 2020, 1-9.

[3] Yiran D. and Chao-Ying J. P. (2013). Principled missing data methods for researchers. *Springer Plus* 2 (1), 222-240.

[4] Adnan F. A., Jamaludin K. R., Muhamad W. Z. and Miskon S. (2021). Review of Current Publications Trend on Missing Data Imputation Over Three Decades: Direction and Future Research. https://doi.org/10.21203/rs.3.rs-996596/v1

[5] Howell, D. (2012). Treatment of Missing Data-Part 1. www.uvm.edu/dhowell/StatPages/More_Stuff/.../Missing.html

[6] Bii N. K., Onyango C. O. and Odhiambo J. (2020). Estimating a Finite Population Mean Using Transformed Data in Presence of Random Nonresponse. *International Journal of Mathematics and Mathematical Sciences* 2020(4), 1-7.

[7] Dorfman, R. (1992). Nonparametric Regression for Estimating Totals in Finite Populations. Proceedings of the Section on Survey Research Methods, *American Statistical Association,* 622–625.

[8] Enders C. K. (2010). *Applied Missing Data Analysis*. New York: Guilford Press.

[9] Brady T. W and Roderick J. A. (2013). Non-response adjustment of survey estimates based on auxiliary variables subject to error. *Journal of Royal Statistical Society*, Vol. 62 (2), 213–231.

[10] Särndal, C. E. and Lundstrom, S. (2005). *Estimation in Surveys with Nonresponse*. New York: John Wiley & Sons.

[11] Yulei, H. (2010). "Missing Data Analysis using Multiple Imputation: Getting to the Heart of the Matter" *American Heart Association*, 3, 98-105.

[12] Saunder, J. A., Morrow, N. H., Spitznagel, E., Dori, P., Enola, K. P. and Pescarino, R. (2006). "Imputing Missing Data: A Comparison of Methods for Social Work Researchers" *Social Work Research,* 30, 19-32.

[13] Little, R. J., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

[14] Chao-Ying, J. P., Harwell, M., Show-Mann, L. and Lee, H. E. (2006). "Advances in Missing Data Methods and Implications for Educational Research." In S. Sawilowsky (Ed.), *Real data analysis.* Greenwich, CT: Information Age Publishing Inc.

[15] Amanda, N. B. and Enders, C. K. (2010). "An introduction to modern missing data analyses." *Journal of School Psychology*, 48, 5–37.

[16] Lehtonen, R. and Pahkinen, E. (2004). *Practical Methods for Design and Analysis of Complex Surveys (2nd Edition).* New York: John Wiley & Sons Ltd.

[17] Overton, W. S. (1985). A Sampling Plan Tor Streams in the National Stream Survey. Statistics, Technical Report 114, Department Oregon State University, Corvallis, Oregon, 97331.

[18] Särndal, C. E., Swensson, B., Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.

[19] Nadaraya, E. A. (1964). "On Estimation Regression" *Theory of Probability and Application*, 9, 141-142.

[20] Watson, G. S. (1964). "Smoothing Regression Analysis" *Sankhya*, *Series A*, 26, 359-372.

[21] Hidiroglou, M. A., Haziza, D. and Rao, J. N. K. (2009). "Comparison of Variance Estimator in Two-phase Sampling: An Empirical Investigation" *Pak. J. of Statistics, 27, 477-492*.

[22] Cochran, W. G. (1977). *Sampling Techniques (3rd Edition)*. New York, John Wiley and Sons.

[23] Dennis, D. W., Mendenhall, R. and Schaeffer, R. L. (2008). *Mathematical Statistics with Application (7th Edition)*. Duxbury: Thomson Books/Cole.

[24] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. New York: Chapman & Hall.