Regression-Cum-Ratio Mean Imputation Class of Estimators using Non-Conventional Robust Measures

Ahmed Audu¹, Yahaya Zakari², Majeed A. Yunusa³, Ishaq O. Olawoyin⁴, Faruk Manu⁵, and Isha Muhammad⁶

Different imputation strategies have been developed by several authors to take care of missing observations during analyses. Nevertheless, the estimators involved in some of these schemes depend on known parameters of the auxiliary variable which outliers can easily influence. In this study, a new class of ratio-type imputation methods that utilize parameters that are free from outliers has been presented. The estimators of the schemes were obtained and their MSEs were derived up to first-order approximation using the Taylor series approach. Also, conditions for which the new estimators are more efficient than others considered in the study were also established. Numerical examples were conducted and the results revealed that the proposed class of estimators is more efficient.

Keywords: Imputation, Non-response, Estimator, Population Mean, Mean Squared Error (MSE).

1. Introduction

It is often assumed at the beginning of the survey that information on sampling units drawn from the population is completely available. This assumption is often violated due to non-response due to incomplete information or inaccessibility to respondents or refusal to answer questions, especially surveys in medical and social science, etc. which often involve sensitive questions. In such situations, responses of non-respondents after often imputed or estimated using imputation techniques. Imputation is the process of replacing missing data with substituted values. There are three main problems that missing data due to non-response causes. It can introduce a substantial amount of bias, make the handling and analysis of the data more arduous, and create reductions in efficiency. Missing data due to non-response creates problems of complications during data analysis. The imputation approach provides all cases by replacing missing data with an estimated value based on other available information and auxiliary variable. Once all missing values have been imputed, the data set can then be analyzed using standard techniques for complete data.

[1] were the first to consider the problem of non-response. Several authors also proposed imputation methods to deal with non-response or missing values. Among them are [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16] and [17]. Recently, [18] suggested a generalized class of imputation in which they compared the efficiency of the estimators obtained from the scheme with that of the estimators of the schemes by the previous authors and found that their estimators outperformed the estimators of the previous authors. Nonetheless, having studied the estimators by [18], it was observed that the estimators depend on the known parameters of the auxiliary variable which outliers can easily influence. In this study, new classes of ratio-type imputation methods which utilized parameters that are free from outliers have been presented.

Notations

The following notations have been used as described in [19], [20], [21], [22], [23], and [24].

Y: Study variable.
X: Auxiliary variable.
\( \bar{X}, \bar{Y} \): Population mean of the variables X and Y respectively.
r: Size of the sample
r: Number of respondents.
R: Ratio of the population mean of study variable to the population mean of auxiliary variable.
\( \bar{X}_n \): The sample mean for the sample of size n.

CaJoST, 2023, 3, 246-254 © 2023 Faculty of Science, Sokoto State University, Sokoto. | 246
\( \bar{x}_r \) : The mean of the variable \( X \) for set \( \Phi \)

\( \bar{y}_r \) : The mean of the variable \( Y \) for set \( \Phi \)

\( S^2_x, S^2_y \) : Population variance of the variables \( X \) and \( Y \)

\( S_Y, S_X \) : Population standard deviation of \( Y \) and \( X \)

\( \beta_1 \) : Population coefficient of skewness of \( X \).

\( \beta_2 \) : Population coefficient of kurtosis of \( X \).

\( \rho_{XY} \) : Population coefficient of correlation between \( Y \) and \( X \).

\( \beta_{rg} \) : Population regression coefficient.

\( C_Y, C_X \) : Population coefficient of variation of \( Y \) and \( X \).

\( G = 4 \frac{\sum_{i=1}^{N} (2i - N - 1)}{N(N-1)} X(i) \) : Gini's mean difference for \( X \).

\( D = 2\sqrt{\frac{\pi}{N-1}} \frac{N-1}{2N} (i - \frac{N+1}{2N}) X(i) \) : Downtown's method for \( X \).

\( S_{pw} = \sqrt{\frac{\pi}{N^2}} \sum_{i=1}^{N} (2i - N - 1) X(i) \) : Probability weighted moments for \( X \).

\( \sum \) : Population variance covariance matrix.

**Sample Mean and [18] Imputation Schemes**

Let \( \Phi \) denotes the set of \( r \) units response and \( \Phi^c \) denotes the set of \( n-r \) units non-response or missing out of \( n \) units sampled without replacement from the \( N \) units population. For each \( i \in \Phi \), the value of \( y_i \) is observed.

However, for unit \( i \in \Phi^c \), \( y_i \) is missing but calculated using different methods of imputation.

The mean method of imputation is defined as

\[
    y_j = \begin{cases} 
    y_i, & \text{if } i \in \Phi \\
    \bar{y}_r, & \text{if } i \in \Phi^c
    \end{cases}
\]  

(1.1)

The point estimator of scheme in (1.1) denoted by \( \hat{\mu}_0 \) is given as in (1.2)

\[
    \hat{\mu}_0 = r^{-1} \sum_{i \in \Phi} y_i
\]  

(1.2)

The variance of \( \hat{\mu}_0 \) is given by (1.3).

\[
    \text{Var}(\hat{\mu}_0) = \psi_{r,N} S^2_Y
\]  

(1.3)

where

\[
    \psi_{r,N} = r^{-1} - N^{-1}, \quad S^2_Y = (N-1)^{-1} \sum_{i=1}^{N} (y_i - \bar{Y})^2, \quad \bar{Y} = N^{-1} \sum_{i=1}^{N} y_i
\]

[18] proposed the following generalized class of imputation schemes;

\[
    y_j = \left[ \hat{\mu}_0 + \beta_i (X_i - \bar{X}) \right] \left[ \psi_i (X_i + \bar{X}) \exp \left( \frac{\psi_i (X_i - \bar{X})}{\pi_i (X_i + \bar{X}) + 2\sigma_i} \right) \right] \quad i \in \Phi
\]  

(1.4)

where \( \pi_1 \) and \( \pi_2 \) are known functions of auxiliary variables like coefficients of skewness \( \beta_1(x) \), kurtosis \( \beta_2(x) \), variation \( C_X \), standard deviation \( S_X \) etc., and \( \pi_1 \neq \pi_2, \sigma_1 = 1, \sigma_2 = 0 \) and \( \pi_1 \neq 0 \).

The point estimators of finite population mean under these methods of imputation are given by

\[
    \hat{\mu}_q^{(\psi)} = r \frac{1}{n} \hat{\mu}_0 + \left( 1 - \frac{r}{n} \right) \left( \hat{\mu}_0 + \beta_i (X_i - \bar{X}) \right) \left( \psi_i (X_i + \bar{X}) \exp \left( \frac{\psi_i (X_i - \bar{X})}{\pi_i (X_i + \bar{X}) + 2\sigma_i} \right) \right)
\]  

(1.5)

\[
    \text{MSE}(\hat{\mu}_q^{(\psi)}) = \psi_{r,N} (S^2_Y + \bar{Y}^2 S^2_Y - 2YS_{XY})
\]  

(1.6)

where, \( \psi_1 = \frac{\pi_1 \bar{X}}{\pi_1 \bar{X} + \pi_2}, \quad \psi_2 = \frac{\sigma_1 \bar{X}}{2(\sigma_1 \bar{X} + \sigma_2)}, \quad \text{and} \quad Y = \left( 1 - \frac{r}{n} \right) (R \psi_1 + \psi_2) + \beta_{rg}
\]

2. **Materials and Methods**

**Proposed Imputation Scheme**

Having studied the scheme and estimators of [18], which utilize known functions of the auxiliary variable, that are sensitive to outliers, we proposed the following scheme to obtain estimators that are not sensitive to outliers by using nonconventional robust measures of the auxiliary variable defined as

\[
    y_j = \left\{ \begin{array}{ll}
    y_i, & \text{if } i \in \Phi \\
    \bar{y}_r + \beta_{rg} (X_i - \bar{X}) / (\phi_i \bar{X} + \phi_k), & \text{if } i \in \Phi^c
    \end{array} \right.
\]  

(2.1)

The associated class of point estimators for scheme (2.1) is obtained as
\[
\hat{t}_m = \frac{r}{n} \bar{y}_r + \left(1 - \frac{r}{n}\right) \left(\bar{y}_r + \beta_{r g} (\bar{X} - \bar{x}_r)\right) \frac{(\phi_j \bar{x} + \phi_k)}{(\phi_{j r} + \phi_{k r})}
\]

Remark 1: Note that \( \phi_j \neq \phi_k \),

where,
\[
\phi = \{(G \times n), (D \times n), (S_{pw} \times n)\}, (\phi_j, \phi_k) \in \phi, t = j, k
\]
\[m = 1, 2, 3, 4, 5, 6,\]

Table 1: Some member of \( \hat{t}_i \) for different values of \( \phi_j \) and \( \phi_k \)

<table>
<thead>
<tr>
<th>( i )</th>
<th>Estimators</th>
<th>Values of Constants</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \hat{t}_1 = \frac{r}{n} \bar{y}_r + \left(1 - \frac{r}{n}\right) \left(\bar{y}<em>r + \beta</em>{r g} (\bar{X} - \bar{x}_r)\right) ((G \times n) \bar{X} + (D \times n)) ) ( (G \times n) ) ( (D \times n) )</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>( \hat{t}_2 = \frac{r}{n} \bar{y}<em>r + \left(1 - \frac{r}{n}\right) \left(\bar{y}<em>r + \beta</em>{r g} (\bar{X} - \bar{x}<em>r)\right) ((G \times n) \bar{X} + (S</em>{pw} \times n)) ) ( (G \times n) ) ( (S</em>{pw} \times n) )</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>( \hat{t}_3 = \frac{r}{n} \bar{y}_r + \left(1 - \frac{r}{n}\right) \left(\bar{y}<em>r + \beta</em>{r g} (\bar{X} - \bar{x}_r)\right) ((D \times n) \bar{X} + (G \times n)) ) ( (D \times n) ) ( (G \times n) )</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>( \hat{t}_4 = \frac{r}{n} \bar{y}<em>r + \left(1 - \frac{r}{n}\right) \left(\bar{y}<em>r + \beta</em>{r g} (\bar{X} - \bar{x}<em>r)\right) ((D \times n) \bar{X} + (S</em>{pw} \times n)) ) ( (D \times n) ) ( (S</em>{pw} \times n) )</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>( \hat{t}_5 = \frac{r}{n} \bar{y}<em>r + \left(1 - \frac{r}{n}\right) \left(\bar{y}<em>r + \beta</em>{r g} (\bar{X} - \bar{x}<em>r)\right) ((S</em>{pw} \times n) \bar{X} + (G \times n)) ) ( (S</em>{pw} \times n) ) ( (G \times n) )</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>( \hat{t}_6 = \frac{r}{n} \bar{y}<em>r + \left(1 - \frac{r}{n}\right) \left(\bar{y}<em>r + \beta</em>{r g} (\bar{X} - \bar{x}<em>r)\right) ((S</em>{pw} \times n) \bar{X} + (D \times n)) ) ( (S</em>{pw} \times n) ) ( (D \times n) )</td>
<td></td>
</tr>
</tbody>
</table>

Data for Empirical Study

For the empirical examples on the precision of proposed estimators, three sets of data obtained from [25] in Table 2 were used.

Table 2: Data used for empirical study

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Pop. 1</th>
<th>Pop. 2</th>
<th>Pop. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N )</td>
<td>34</td>
<td>34</td>
<td>80</td>
</tr>
<tr>
<td>( n )</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>( r ) (Assumed)</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>( \bar{Y} )</td>
<td>856.4117</td>
<td>856.4117</td>
<td>5182.637</td>
</tr>
<tr>
<td>( \bar{X} )</td>
<td>208.8823</td>
<td>199.4412</td>
<td>1126.463</td>
</tr>
<tr>
<td>( C_{\bar{Y}} )</td>
<td>0.8561</td>
<td>0.8561</td>
<td>0.354193</td>
</tr>
<tr>
<td>( C_{\bar{X}} )</td>
<td>0.7205</td>
<td>0.7531</td>
<td>0.7506772</td>
</tr>
<tr>
<td>( \beta_1 (x) )</td>
<td>0.9782</td>
<td>1.1823</td>
<td>1.050002</td>
</tr>
<tr>
<td>( \beta_2 (x) )</td>
<td>0.0978</td>
<td>1.0445</td>
<td>-0.063386</td>
</tr>
<tr>
<td>( \rho_{\bar{Y}X} )</td>
<td>0.4491</td>
<td>0.4453</td>
<td>0.9410</td>
</tr>
<tr>
<td>( S_{\bar{X}} )</td>
<td>150.5059</td>
<td>150.2150</td>
<td>845.610</td>
</tr>
</tbody>
</table>
Simulated Data for Empirical Study

In this section, Data of size 1000 units were generated for study populations using function defined in Table 3. Samples of size 100 units from which 60 units were selected as respondents were randomly chosen 10,000 times by method of simple random sampling without replacement (SRSWOR). The Biases and MSEs of the considered estimators were computed using (2.3) and (2.4) respectively

\[ M S E (\hat{\theta}_d) = \frac{1}{10000} \sum_{i=1}^{10000} (\hat{\theta}_d - \bar{Y})^2 , \hat{\theta}_d = \bar{y}_r, \hat{\mu}_r(\cdot) \]

Table 3: Populations used for Simulation Study

<table>
<thead>
<tr>
<th>Populations</th>
<th>Auxiliary variable (x)</th>
<th>Study variable (y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>X \sim beta(1.1,2.0)</td>
<td>\bar{Y} = 50+10X + 20X^2 + e,</td>
</tr>
<tr>
<td>II</td>
<td>X \sim gamma(10,25)</td>
<td>where, e \sim (0,4)</td>
</tr>
</tbody>
</table>

3. Results and Discussion

Properties of the Estimators suggested

The MSE of estimators for suggested imputation schemes is obtained as

\[ M S E (\hat{t}_m^1) = \Delta \Sigma^T \Delta \]

where, \( \Delta = \left( \frac{\partial \hat{t}_m^1}{\partial \bar{Y}^2} \right) \) (3.1) where \( \Delta = \left( \frac{\partial \hat{t}_m^1}{\partial \bar{Y}^2} \right) \)

By substituting (3.3) and (3.6) into (3.1), we obtain the mean square error of the estimators as
Theorem 1: the estimators \( \hat{\mu}_m \) are consistent.

Proof: Let \( f(x) \) and \( g(x) \) be continuous function, then

\[
\lim_{x \to p} \left( f(x) \pm g(x) \right) = \lim_{x \to p} f(x) \pm \lim_{x \to p} g(x), \quad p \neq \infty
\]

\[
\lim_{x \to p} \left( f(x) \times g(x) \right) = \lim_{x \to p} f(x) \times \lim_{x \to p} g(x), \quad p \neq \infty
\]

\[
\lim_{x \to p} \frac{f(x)}{g(x)} = \lim_{x \to p} \frac{f(x)}{\lim_{x \to p} g(x)}, \quad p \neq \infty, \quad \lim_{x \to p} g(x) \neq 0
\]

(3.9)

As \( r \to N, \ n = N \), using the results of (3.9), (3.10) and (3.11), we have

\[
\hat{\mu}_m(r \to N, n = N) = \frac{\sum_{i=1}^{k} \hat{\mu}_i(r \to N, n = N)}{n}
\]

(3.12)

Since \( n = N \) if \( r \to N \), then

\[
\lim_{r \to N} \hat{\mu}_i(r) = \hat{\mu}_i, \quad \lim_{r \to N} \hat{\mu}_i = \mu_i
\]

Therefore,

\[ \hat{\mu}_m, \hat{\mu}_i, i = 1, 2, ..., 17, \hat{\mu}_k, k = 0, 1, 2, ..., 6 \]

Hence, the proof.

Theoretical Efficiency Comparison

In this section, conditions for the efficiency of the new estimators over some existing related estimators were established.

Theorem 2: Estimator \( \hat{\mu}_i \) is more efficient than \( \hat{\mu}_i \) if (3.14) is satisfied.

\[
M < 2\beta_{rg}
\]

(3.14)

Proof: Minus (3.8) from (1.3), theorem 2 is proved.

Theorem 3: Estimator \( \hat{\mu}_i \) is more efficient than \( \hat{\mu}_i \) if (3.14) is satisfied.

\[
(\gamma + M) \beta_{rg} + (M^2 - \gamma^2) < 0
\]

(3.15)

Proof: Subtract (3.8) from (1.6), theorem 3 is proved.

Empirical Study using Real life Data

Empirical examples on the precision of proposed estimators using real life data in Table 2 were considered in this section

Table 4: MSE of Estimators \( \hat{\mu}_0, \hat{\mu}_i, i = 1, 2, ..., 17, \hat{\mu}_k, k = 0, 1, 2, ..., 6 \)

<table>
<thead>
<tr>
<th>Estimators</th>
<th>Pop. 1</th>
<th>Pop. 2</th>
<th>Pop. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample mean ( \hat{\mu}_0 )</td>
<td>25537.11</td>
<td>25537.11</td>
<td>217082.8</td>
</tr>
<tr>
<td>( \hat{\mu}_i )</td>
<td>20960.84</td>
<td>21232.84</td>
<td>79355.19</td>
</tr>
<tr>
<td>( \hat{\mu}_2 )</td>
<td>20953.11</td>
<td>21222.67</td>
<td>79247.78</td>
</tr>
<tr>
<td>( \hat{\mu}_3 )</td>
<td>20950.37</td>
<td>21216.94</td>
<td>79205.03</td>
</tr>
<tr>
<td>( \hat{\mu}_4 )</td>
<td>20959.78</td>
<td>21218.77</td>
<td>79364.27</td>
</tr>
<tr>
<td>( \hat{\mu}_5 )</td>
<td>20404.57</td>
<td>20514.50</td>
<td>32126.76</td>
</tr>
</tbody>
</table>
Table 4 shows the numerical results of the MSE of estimators $\hat{\mu}_i$, $\hat{\mu}_{i(i')}$, $i = 1, 2, 3, 4, ..., 17$ and $\hat{t}_m$, $m = 1, 2, 3, 4, 5, 6$, using data sets in Table 2. Of all the subjects examined, the proposed two proposals have a minimum MSE for all data sets. This means that the proposed methods have shown a high level of efficiency on others considered in the study, and can produce a better estimate of the average population in the presence of an unresponsive or missing observation on average.

**Empirical Study using Simulated Data**

In this section, simulation studies were conducted to assess the performance of the estimators of the proposed schemes with respect to [18] estimators using study populations simulated by functions defined in Table 3 and the results are presented in table 5.
### Table 5: Biases and MSEs of Proposed and Other from Simulated Data

<table>
<thead>
<tr>
<th>Estimators</th>
<th>Population I</th>
<th></th>
<th>Population II</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Biases</td>
<td>MSEs</td>
<td>Biases</td>
<td>MSEs</td>
</tr>
<tr>
<td>Sample mean $\hat{\mu}_0$</td>
<td>0.005212425</td>
<td>0.08212077</td>
<td>-0.00238708</td>
<td>0.02055153</td>
</tr>
<tr>
<td><strong>[18] Estimators</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{\mu}_1^*$</td>
<td>0.01881164</td>
<td>0.5988081</td>
<td>0.01224332</td>
<td>0.1290456</td>
</tr>
<tr>
<td>$\hat{\mu}_2^*$</td>
<td>0.006281752</td>
<td>0.1573342</td>
<td>0.006784724</td>
<td>0.05279314</td>
</tr>
<tr>
<td>$\hat{\mu}_3^*$</td>
<td>0.00771724</td>
<td>0.2157503</td>
<td>0.005456455</td>
<td>0.03706221</td>
</tr>
<tr>
<td>$\hat{\mu}_4^*$</td>
<td>0.01400318</td>
<td>0.120201</td>
<td>0.008066989</td>
<td>0.06924065</td>
</tr>
<tr>
<td>$\hat{\mu}_5^*$</td>
<td>0.009613986</td>
<td>0.2862123</td>
<td>0.00908595</td>
<td>0.08303842</td>
</tr>
<tr>
<td>$\hat{\mu}_6^*$</td>
<td>0.006427576</td>
<td>0.1636258</td>
<td>0.003340483</td>
<td>0.01579654</td>
</tr>
<tr>
<td>$\hat{\mu}_7^*$</td>
<td>0.007157499</td>
<td>0.0326743</td>
<td>-0.0003362034</td>
<td>0.03338117</td>
</tr>
<tr>
<td>$\hat{\mu}_8^*$</td>
<td>0.007912686</td>
<td>0.2232647</td>
<td>0.006160419</td>
<td>0.04521465</td>
</tr>
<tr>
<td>$\hat{\mu}_9^*$</td>
<td>0.004970689</td>
<td>0.0328937</td>
<td>0.003465998</td>
<td>0.01688858</td>
</tr>
<tr>
<td>$\hat{\mu}_{10}^*$</td>
<td>0.005403985</td>
<td>0.1630998</td>
<td>0.007531888</td>
<td>0.06224325</td>
</tr>
<tr>
<td>$\hat{\mu}_{11}^*$</td>
<td>0.006415286</td>
<td>0.2157503</td>
<td>0.005456455</td>
<td>0.03706221</td>
</tr>
<tr>
<td>$\hat{\mu}_{12}^*$</td>
<td>0.007157499</td>
<td>0.0326743</td>
<td>-0.0003362034</td>
<td>0.03338117</td>
</tr>
<tr>
<td>$\hat{\mu}_{13}^*$</td>
<td>0.007912686</td>
<td>0.2232647</td>
<td>0.006160419</td>
<td>0.04521465</td>
</tr>
<tr>
<td>$\hat{\mu}_{14}^*$</td>
<td>0.004970689</td>
<td>0.0328937</td>
<td>0.003465998</td>
<td>0.01688858</td>
</tr>
<tr>
<td>$\hat{\mu}_{15}^*$</td>
<td>0.005403985</td>
<td>0.1630998</td>
<td>0.007531888</td>
<td>0.06224325</td>
</tr>
<tr>
<td>$\hat{\mu}_{16}^*$</td>
<td>0.006415286</td>
<td>0.2157503</td>
<td>0.005456455</td>
<td>0.03706221</td>
</tr>
<tr>
<td>$\hat{\mu}_{17}^*$</td>
<td>0.007157499</td>
<td>0.0326743</td>
<td>-0.0003362034</td>
<td>0.03338117</td>
</tr>
</tbody>
</table>

Table 5 shows the results of the biases and MSEs of the estimators of the proposed schemes, using the simulated data for population 1 in Table 3. The results revealed that the estimators of the proposed scheme have minimum biases and MSEs. This implies that the estimators of the proposed scheme are more efficient than Sample mean, [18] estimators and can produce a reliable estimate closer to the true population mean of the study variable.

### 4. Conclusion

From the results of the empirical studies, it was obtained that the proposed estimator is more efficient than other estimators considered in the study and, therefore, its use is recommended to estimate the population average when certain
values of the variables of the study are missing in the study.

**Conflict of Interest**

The author declares that there is no conflict of interest.

**Acknowledgements**

The authors acknowledge the contributions of the reviewers in shaping the standard and quality of this manuscript.

**References**


