STATISTICS IN TRANSITION new series, June 2025
Vol. 26, No. 2, pp. xxx–xxx, https://doi.org/10.59139/stattrans-2025-xxx
Received – xx.xx.xxxx; accepted – xx.xx.xxxx

Ratio regression type estimators of the population mean for missing data in sample surveys

Prachi Garg[1], Namita Srivastava[2], Manoj Kumar Srivastava[33]

Abstract

In this article, new ratio regression type estimators with imputation have been proposed as means to overcome the problem of missing data relating to a studied variable in a sample survey. It has been shown that the suggested estimators are more efficient than the mean method of imputation, the ratio method of imputation, the regression method of imputation, and the estimators given by Singh and Horn (2000), Singh and Deo (2003), Singh (2009), Diana and Perri (2010) and Gira (2015). The biases and their mean square errors of the suggested estimators are derived. A comparative study is conducted using real and simulated data. The results are found to be encouraging showing improvement of all the methods discussed in this article.

Key words: imputation methods, Bias, Mean square error (MSE), Efficiency, Ratio-Regression type estimators.

1. Introduction

Missing data or missing values occur when no data value is stored for a variable in an observation. Even in a well-designed and controlled study, missing data occurs in almost all research. Missing data is commonly described as major issue in most scientific research domains that may originate from such a mishandling sample, measurement error, non-response or deleted aberrant value. To get precise estimates of population parameters we seek information on every selected unit of the sample. Imputation means replacing a missing value with other value based on a reasonable estimate. Information on the related auxiliary variable is generally used to recreate the

[1] Department of Statistics, St. John's College Agra, Dr. B.R. Ambedkar University, Agra (U.P.).India. E-mail: prachigarg2093@gmail.com. ORCID: https://orcid.org/0009-0001-8809-4464.

[2] Department of Statistics, St. John's College Agra, Dr. B.R. Ambedkar University, Agra (U.P.). India. E-mail: drnamita.sjc@gmail.com. ORCID: https://orcid.org/0000-0001-8695-9148.

[3] Shaheed Mahendra Karma Vishwavidyalaya, Bastar, Chhattisgarh, India,. E-mail: mksriv@gmail.com. ORCID: https://orcid.org/0000-0002-8256-1439.

missing values for completing datasets. Incomplete data is usually categorized into three different response mechanisms: Missing Completely At Random (MCAR); Missing At Random (MAR); and Missing Not At Random (MNAR or NMAR) (Little & Rubin,2002). In Missing Completely at Random (MCAR) missing data is randomly distributed across the variable and unrelated to other variables. In Missing at Random (MAR) the missing observations are not randomly distributed but they are accounted for by other observed variables. In Missing Not at Random (MNAR) category, the missing data systematically differ from the observed values. In the present article we are assuming MCAR response mechanism of missing data.

Auxiliary information is important for a survey practitioner as it is utilized to improve the performance of the methods in finite sample survey. At the estimation stage the auxiliary information is utilized for suggesting imputation methods which results in ratio, product and regression estimators. Many imputation methods have been proposed utilizing the auxiliary information. Several researchers (Lee, Rancourt, and Sarndal 1994,1995; Singh and Horn 2000; Singh et. al; Diana and Perri 2010; Pandey, Thakur, and Yadav 2015; Singh et. al. 2016; Bhushan and Pandey 2018; Prasad 2017, 2018, 2019; Singh and Khalid 2019; K Chodjuntug and N Lawson (2022)[4]; K Chodjuntug and N Lawson (2022)[5]; N Lawson (2023) [12]; N Lawson (2023) [13]; N Thongsak and N Lawson (2023) [14] etc.) among others assumed MCAR mechanism to develop several imputation methods and resultant estimators to estimate population mean in the case of missing data problems. The imputation methods proposed by Singh and Horn (2000), Singh and Deo (2003), Singh (2009), Diana and Perri (2010), and Gira (2015) result in different estimators, but they all lead to the same Mean Squared Error (MSE) formula, which are same as regression method of imputation. Therefore, in this paper we compared and simulated our estimator with the mean, ratio and regression estimators after proposing the new imputation strategy and the resulting estimator. The proposed estimators come out to be more efficient than the usual mean, ratio and regression (Diana & Perri's regression) method for handling missing observations to estimate the population mean.

This article proposes three ratio-regression type imputation methods to inadequate the annoyance outcome of nonresponse in survey sampling. The resulting classes of point estimators that can be used to estimate the population mean have been discussed in detail. The bias and Mean Square Error (MSE) properties of the proposed estimators have been derived. An empirical study was conducted to assess their performance in comparison with existing estimators, and the findings have been presented. These are designed as follows.. In Section 2,the sample structure and notations are considered and in Section 3, we have reviewed several imputation techniques of finite population mean under non-response that are available in the literature suggested by various authors. In Section 4, construction of the suggested alternative method of imputation

is carried out and the bias mean square error equations for this estimator is obtained. In Section 5, we have proposed a new method of imputation and obtained their bias mean square error equations for this estimator. In Section 6, we have conducted efficiency comparison of alternative method of imputation. In Section 7, we do computational study by using real and artificial populations, respectively. Section 8 summaries the main findings and conclusions.

2. Sample Structure and Notations:

Consider a finite population U = $\{U_1, U_2, \ldots\ldots, U_N\}$ of size N for which a random sample s = $\{u_1, u_2, \ldots\ldots, u_N\}$ of size n under simple random sampling without replacement scheme is drawn to estimate the population mean $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ of the study variable y. Let y_i and x_i be the values of the study variable y and auxiliary variable x , respectively for the i^{th} unit of a finite population of size N. The information on x can be available on the entire population through knowledge of x_i, ∀ iϵU, or its population mean $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$. Let s be a simple random sample without replacement (SRSWOR) of size n $(n < N)$ drawn from U to estimate Y. Let r be the number of responding units out of sampled n units. Let the set of responding units be denoted by R and that of non-responding (n-r) units be denoted by R^C. For every unit, $i \epsilon R$ the value y_i is observed. However, for the units, $i \epsilon R^C$, the y_i values are missing and imputed values are derived. Imputation is performed by employing the auxiliary variable x where values are believed to be known for each sampled unit iϵs.

The structure of the general method of imputation in the case of complete dataset under nonresponse is defined as:

$$y_{.i}=\begin{cases} y_i & if \quad i\epsilon R \\ \hat{y}_i & if \quad i\epsilon R^c \end{cases}$$

where $\hat{y}_i$ is the imputed value for the i^{th} non-responding unit. Using the above data, we get the following form of the general point estimator of the population mean ($\bar{Y}$)

$$\bar{y}_s = \frac{1}{n}\sum_{i\epsilon s_n} y_{.i} = \frac{1}{n}[\sum_{i\epsilon R} y_{.i} + \sum_{i\epsilon R^c} y_{.i}] = \frac{1}{n}[\sum_{i\epsilon R} y_i + \sum_{i\epsilon R^c} \hat{y}_i].$$

Here, $\hat{y}_i$ takes a different value for a different imputation method.

The following notations have been adopted for further use:

$\bar{X}$,$\bar{Y}$: The population mean of the auxiliary variable x and study variable y respectively,

$\bar{y}_r$: Sample mean of responding units,

$\bar{x}_n$: Sample mean of all units,

$\bar{x}_r$: Sample mean of responding units,

$\rho_{yx} = \frac{S_{yx}}{S_y S_x}$: The correlation coefficient between the variables y and x,

$S_{yx} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})(x_i - \bar{X})$: The covariance between y and x,
$S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$: The population mean square of y,
$S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{X})^2$: The population mean square of x,
$C_{y=}\frac{S_y}{\bar{Y}}$ & $C_{x=}\frac{S_x}{\bar{X}}$: The coefficients of variation of y and x, respectively,

we define,

$$\bar{y}_r = \bar{Y}\,(1+\varepsilon_o),\ \bar{x}_r = \bar{X}\,(1+\delta_o),\ \ \bar{x}_n = \bar{X}\,(1+\eta_o)$$

using the above notation, we have

$$\mathrm{E}\,(\varepsilon_o) = E(\delta_o) = E(\eta_o) = 0$$

and,

$$\mathrm{E}(\varepsilon_o^2) = \left(\frac{1}{r}-\frac{1}{N}\right)C_y^2,\ \mathrm{E}(\delta_o^2) = \left(\frac{1}{r}-\frac{1}{N}\right)C_x^2,\ \mathrm{E}(\varepsilon_o\delta_o) = \left(\frac{1}{r}-\frac{1}{N}\right)\rho_{xy}C_xC_y,$$
$$\mathrm{E}(\eta_o^2) = \left(\frac{1}{n}-\frac{1}{N}\right)C_x^2,\mathrm{E}(\delta_o\eta_o) = \left(\frac{1}{n}-\frac{1}{N}\right)C_x^2,\mathrm{E}(\varepsilon_o\eta_o) = \left(\frac{1}{n}-\frac{1}{N}\right)\rho_{xy}C_xC_y.$$

3. Review of Some existing estimators:

In this section, we discuss some of the classical and existing imputation methods for estimating the population mean in sample surveys.

3.1. Mean method of imputation:

In the mean method of imputation the form of data by Lee, Rancourt and Sarndal (1994) is treated as

$$y_{i,m} = \begin{cases} y_i & if \quad i\epsilon R \\ \bar{y}_r & if \quad i\epsilon R^c \end{cases} \tag{3.1}$$

The mean estimator under the new data (3.1) is given by

$$\bar{y}_m = \frac{1}{r}\sum_{i\epsilon R} y_i = \bar{y}_r \tag{3.2}$$

The variance of the response sample mean $\bar{y}_m$ is given by

$$V\,(\bar{y}_m) = V\,(\bar{y}_r) = \left(\frac{1}{r}-\frac{1}{n}\right)\bar{Y}^2C_y^2 \tag{3.3}$$

3.2. Ratio method of imputation:

The ratio method of imputation, based on information from the auxiliary variable x, was proposed by Lee, Rancourt, and Särndal (1994). Under this method, the imputed data are adjusted using the known relationship between the study variable and the auxiliary variable.

$$y_{i,R} = \begin{cases} y_i & if \quad i\epsilon R \\ \hat{b}x_i & if \quad i\epsilon R^c \end{cases} \tag{3.4}$$

where $\hat{b} = \frac{\sum_{i\,\epsilon\,R} y_i}{\sum_{i\epsilon R} x_i}$.

The ratio estimator in the case of imputation method (3.4), is defined as

$$t_R = \bar{y}_r \left(\frac{\bar{x}_n}{\bar{x}_r}\right) = \bar{y}_{RAT} \tag{3.5}$$

The bias and mean square error of the estimator t_R are obtained under MCAR response mechanism up to the first order approximation, and are given by

$$B(t_R) = \left(\frac{1}{r} - \frac{1}{n}\right) \bar{Y}\, \mathrm{C}_x(C_x - \rho C_y) \tag{3.6}$$

and

$$\mathrm{MSE}\,(\,\bar{y}_{RAT}) \;= V(\,\bar{y}_r) + \left(\frac{1}{r} - \frac{1}{n}\right) \bar{Y}^2 \mathrm{C}_{\boldsymbol{x}}(C_x - 2\,\rho C_y) \tag{3.7}$$

The ratio method of imputation is better to choose over the mean method of imputation whenever $(\rho C_y/C_x)$>1 / 2.

3.3. Regression method of imputation:

In this method, the data after imputation becomes

$$y_{i,REG} = \begin{cases} y_i & if \quad i\epsilon R \\ a + b_{yx} x_i & if \quad i\epsilon R^c \end{cases} \tag{3.8}$$

where, $a = \bar{y}_r - b_{yx}\,\bar{x}_r$ and $b_{yx} = \frac{S_{yx}}{S_x^2}$

The point estimator of population mean $\overline{\mathrm{Y}}$

$$\bar{y}_{REG} = \bar{y}_r + b_{yx}(\bar{x}_n - \bar{x}_r) \tag{3.9}$$

The bias and mean square error of the estimator $\bar{y}_{REG}$ are obtained under MCAR response mechanism up to the first order approximation, and are given by

$$B(\bar{y}_{REG}) = \frac{\rho_{yx} C_y}{C_x \bar{X}} \left(\frac{1}{r} - \frac{1}{n}\right) \overline{\mathrm{Y}} \left(\frac{\mu_{300}}{\mu_{200}} - \frac{\mu_{210}}{\mu_{110}}\right) \tag{3.10}$$

where, $\mu_{abc} = \sum_{i=1}^{N} (x_i - \bar{X})^a\, (y_i - \bar{Y})^b (z_i - \bar{Z})^c$

$$\mathrm{M}(\bar{y}_{REG}) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{r} - \frac{1}{n}\right) S_y^2 \rho_{yx}^2 \tag{3.11}$$

3.4. Singh and Horn (2000) Estimator

Singh and Horn (2000) introduced this method, the data after imputation becomes

$$y_{.i} = \begin{cases} (\alpha n/r) y_i + (1-\alpha)\hat{b} x_i & if \quad i\epsilon R \\ (1-\alpha)\hat{b} x_i & if \quad i\epsilon R^c \end{cases} \tag{3.12}$$

The point estimator of population mean is given as

$$\bar{y}_{comp} = \left[\alpha \bar{\mathrm{y}}_{\mathrm{r}} + (1-\alpha)\bar{\mathrm{y}}_{\mathrm{r}} \frac{\bar{\mathrm{x}}_{\mathrm{n}}}{\bar{\mathrm{x}}_{\mathrm{r}}}\right] \tag{3.13}$$

where α is an appropriate constant with optimum value $\alpha^* = 1 - \rho_{yx}\left(\frac{C_y}{C_x}\right)$: The bias of the estimator $\bar{y}_{comp}$ is given by

$$B(\bar{y}_{COMP}) = \;(1-\alpha)\left(\frac{1}{r} - \frac{1}{n}\right) \overline{\mathrm{Y}}\, \mathrm{C}_{\boldsymbol{x}}\big(C_x - \rho_{yx}\, C_y\big) \tag{3.14}$$

using α^* we get the minimum MSE of $\bar{y}_{comp}$ as

$$\mathrm{M_{min}}(\bar{\mathrm{y}}_{\mathrm{COMP}}) = \mathrm{MSE}(\bar{y}_{RAT}) - \left(\frac{1}{r}-\frac{1}{n}\right)\bar{\mathrm{Y}}^2(C_x - \rho_{yx}\, C_y)^{\,2} \tag{3.15}$$

3.5. Singh and Deo (2003) Estimator:

Singh and Deo (2003), using power transformation, this method gives the following form of the data after imputation

$$y_{.i} = \begin{cases} y_i & if \quad i\epsilon R \\ \bar{\mathrm{y}}_{\mathrm{r}}\left[\mathrm{n}\left(\frac{\bar{x}_n}{\bar{x}_r}\right)^{\alpha} - \mathrm{r}\right]\frac{x_i}{\sum_{\mathrm{i}\epsilon \mathrm{R^c}} x_i} & if \quad i\epsilon R^c \end{cases} \tag{3.16}$$

The resultant estimator of the population mean is given as

$$\bar{y}_{SD} = \bar{\mathrm{y}}_{\mathrm{r}}\left(\frac{\bar{\mathrm{x}}_{\mathrm{n}}}{\bar{\mathrm{x}}_{\mathrm{r}}}\right)^{\alpha} \tag{3.17}$$

where α is a suitably chosen constant and the optimum value α is $\alpha^*=\rho_{yx}\left(\frac{C_y}{C_x}\right)$

The bias of the estimator $(\bar{y}_{SD})$ obtained by Singh and Deo is given by

$$B(\bar{y}_{SD}) = \left(\frac{1}{r}-\frac{1}{n}\right)\bar{\mathrm{Y}}\left(\frac{\beta(\beta-1)}{2}\mathrm{C_x^2} - \rho_{yx}\, C_y C_x\right) \tag{3.18}$$

using optimum value α^*, the minimum MSE of $\bar{y}_{SD}$ is given

$$\mathrm{MSE_{min}}(\bar{\mathrm{y}}_{\mathrm{SD}}) = \mathrm{MSE}(\bar{y}_{RAT}) - \left(\frac{1}{r}-\frac{1}{n}\right)\mathrm{S_x^2}\left(\frac{S_{yx}}{\mathrm{S_x^2}} - \frac{\bar{\mathrm{Y}}}{\bar{\mathrm{X}}}\right)^{\mathbf{2}} \tag{3.19}$$

3.6. Singh (2009) Estimator:

This method of imputation is an alternative technique to estimate population mean $\bar{\mathrm{Y}}$ in the presence of non-response. The study variate after imputation takes the following form

$$y_{.i} = \begin{cases} y_i & if \quad i\epsilon R \\ \bar{y}_r\left[\frac{(n-r)\bar{x}_n + \alpha r(\bar{x}_n - \bar{x}_r)}{\alpha\bar{x}_r + (1-\alpha)\bar{x}_n}\right]\frac{x_i}{\sum_{i\epsilon R^c} x_i} & if \quad i\epsilon R^c \end{cases} \tag{3.20}$$

The point estimator of population mean as following

$$\bar{y}_{singh} = \frac{\bar{y}_r\bar{x}_n}{\alpha\bar{x}_r + (1-\alpha)\bar{x}_n} \tag{3.21}$$

where α is an appropriate constant with optimum value $\alpha^* = \rho_{yx}$

The bias of the estimator $(\bar{y}_{singh})$ is given as

$$B(\bar{y}_{singh}) = \bar{\mathrm{Y}}\left[\left(\frac{1}{n}-\frac{1}{N}\right)\rho_{yx}\, C_y C_x + \alpha^2\left(\frac{1}{r}-\frac{1}{n}\right)c_x^2 + (1-\alpha)^2\left(\frac{1}{n}-\frac{1}{N}\right)c_x^2 - \alpha\left\{\left(\frac{1}{r}-\frac{1}{n}\right)\rho_{yx}\, C_y C_x + \left(\frac{1}{n}-\frac{1}{N}\right)c_x^2\right\} + 2\alpha(\alpha-1)\left(\frac{1}{n}-\frac{1}{N}\right)c_x^2 - (1-\alpha)\left(\frac{1}{n}-\frac{1}{N}\right)\left(\rho_{yx}\, C_y C_x + c_x^2\right)\right] \tag{3.22}$$

using optimum value of α is α^*, the minimum MSE of $\bar{y}_{Singh}$ is given

$$\mathrm{MSE_{min}}(\bar{\mathrm{y}}_{\mathrm{Singh}}) = \mathrm{MSE}(\bar{y}_{RAT}) - \left(\frac{1}{r}-\frac{1}{n}\right)\mathrm{S_x^2}\left(\frac{S_{yx}}{\mathrm{S_x^2}} - \frac{\bar{\mathrm{Y}}}{\bar{\mathrm{X}}}\right)^{\mathbf{2}} \tag{3.23}$$

3.7. Diana and Perri (2010) Estimators

Diana and Perri (2010) propounded three regression-type imputation methods for missing data as

$$y_{i,DP1} = \begin{cases} \frac{n\,y_i}{r} + b(\overline{\mathrm{X}} - x_i) & if \quad i\epsilon R \\ b(\overline{\mathrm{X}} - x_i) & if \quad i\epsilon R^c \end{cases} \tag{3.24}$$

$$y_{i,DP2} = \begin{cases} \frac{n\,y_i}{r} - \mathrm{b}\,\frac{n\,x_i}{r} & if \quad i\epsilon R \\ b\,\frac{n\overline{\mathrm{X}}}{n-r} & if \quad i\epsilon R^c \end{cases} \tag{3.25}$$

$$y_{i,DP3} = \begin{cases} \frac{n\,y_i}{r} - \mathrm{b}\,\frac{n\,x_i}{r} & if \quad i\epsilon R \\ b\,\frac{n\bar{x}_n}{n-r} & if \quad i\epsilon R^c \end{cases} \tag{3.26}$$

The subsequent estimators under the imputation methods are respectively, given as

$$\bar{y}_{DP1} = \bar{y}_r + b(\overline{\mathrm{X}} - \bar{x}_n) \tag{3.27}$$

$$\bar{y}_{DP2} = \bar{y}_r + b(\overline{\mathrm{X}} - \bar{x}_r) \tag{3.28}$$

$$\bar{y}_{DP3} = \bar{y}_r + b(\bar{x}_n - \bar{x}_r) \tag{3.29}$$

and

$$\mathrm{MSE}(\bar{y}_{DP1}) = S_y^2\left[\left(\frac{1}{n} - \frac{1}{N}\right)(1 - \rho_{yx})^2 + \left(\frac{1}{r} - \frac{1}{n}\right)\right] \tag{3.30}$$

$$\mathrm{MSE}(\bar{y}_{DP2}) = S_y^2\left(\frac{1}{r} - \frac{1}{N}\right)(1 - \rho_{yx})^2 \tag{3.31}$$

$$\mathrm{MSE}(\bar{y}_{DP3}) = S_y^2\left[\left(\frac{1}{n} - \frac{1}{N}\right) + \left(\frac{1}{r} - \frac{1}{n}\right)(1 - \rho_{yx})^2\right] \tag{3.32}$$

3.8. Gira (2015) Estimator

Gira (2015) proposed a ratio type imputation procedure where the study variate after imputation becomes

$$y_{.i} = \begin{cases} y_i & if \quad i\epsilon R \\ \bar{y}_r\left[n\left(\frac{\alpha - \bar{x}_r}{\alpha - \bar{x}_n}\right)\right]\frac{x_i}{\sum_{i\epsilon R^c} x_i} & if \quad i\epsilon R^c \end{cases} \tag{3.33}$$

where α is a suitably chosen constant, such that the MSE of the resultant estimator is minimum. Note that if $\alpha = 0$ then $\bar{\mathrm{y}}_{\mathrm{Gira}} = \bar{\mathrm{y}}_{\mathrm{Ratio}}$. The resultant estimator is obtained as

$$\bar{y}_{Gira} = \bar{y}_r\,\frac{\alpha - \bar{x}_r}{\alpha - \bar{x}_n} \tag{3.34}$$

The bias of the above estimator is

$$B(\bar{y}_{Gira}) = -\frac{\bar{X}\bar{Y}}{\alpha - \bar{X}}\left(\frac{1}{r} - \frac{1}{n}\right)\rho_{yx}\,C_y C_x \tag{3.35}$$

Using the optimum value of $\alpha = \bar{X}\{C_x(\rho_{yx}\,C_y)^{-1} - 1\}$ and the optimum MSE of $\bar{y}_{Gira}$ as follows.

$$\mathrm{M}(\bar{\mathrm{y}}_{\mathrm{Gira}}) = \mathrm{V}(\bar{\mathrm{y}}_{\mathrm{m}}) - \left(\frac{1}{r} - \frac{1}{n}\right)\overline{\mathrm{Y}}^2\mathrm{C}_{\mathrm{y}}^2\rho_{yx}^2 \tag{3.36}$$

4. An Alternative Method of Imputation

The estimators rely on three different ratio-regression type methods of imputation as follows.

Case I: Auxiliary information on X is completely available, i.e., $\bar{X}$ is known and corresponding estimates $\bar{x}_n$ are used in the imputation technique

$$y_{.i1} = \begin{cases} y_i & if \quad i\epsilon R \\ \dfrac{n\bar{y}_r}{n-r}\left[\left\{2-\left(\dfrac{\bar{x}_n}{\bar{X}}\right)^k\right\}+\beta_1(\bar{X}-\bar{x}_n)-r\right] & if \quad i\epsilon R^c \end{cases}$$

The resultant estimator of population mean $\bar{Y}$ is given as

$$\bar{y}_{KB1} = \bar{y}_r\left\{2-\left(\frac{\bar{x}_n}{\bar{X}}\right)^k\right\}+\beta_1(\bar{X}-\bar{x}_n) \tag{4.1}$$

Case II: Auxiliary information on X is completely available i.e., $\bar{X}$ is known and corresponding estimates $\bar{x}_r$ are used in the imputation technique.

$$y_{.i2} = \begin{cases} y_i & if \quad i\epsilon R \\ \dfrac{n\bar{y}_r}{n-r} & \left[\left\{2-\left(\dfrac{\bar{x}_r}{\bar{X}}\right)^k\right\}+\beta_1(\bar{X}-\bar{x}_r)-\mathrm{r}\right] \quad if \quad i\epsilon R^c \end{cases}$$

The resultant estimator of population mean $\bar{Y}$ is given as

$$\bar{y}_{KB2} = \bar{y}_r\left\{2-\left(\frac{\bar{x}_r}{\bar{X}}\right)^k\right\}+\beta_1(\bar{X}-\bar{x}_r) \tag{4.2}$$

Case III: Auxiliary information on X is not available at population level, i.e., $\bar{X}$ is not known and we use corresponding estimates $\bar{x}_n$, $\bar{x}_r$ is used in the imputation technique.

$$y_{.i3} = \begin{cases} y_i & if \quad i\epsilon R \\ \dfrac{n\bar{y}_r}{n-r}\left[\left\{2-\left(\dfrac{\bar{x}_r}{\bar{x}_n}\right)^k\right\}+\beta_1(\bar{x}_n-\bar{x}_r)-r\right] & if \quad i\epsilon R^c \end{cases}$$

The resultant estimator of population mean $\bar{Y}$ is given as

$$\bar{y}_{KB3} = \bar{y}_r\left\{2-\left(\frac{\bar{x}_r}{\bar{x}_n}\right)^k\right\}+\beta_1(\bar{x}_n-\bar{x}_r) \tag{4.3}$$

Therefore, the expression of Bias and Mean Squared Error (MSE) of proposed estimator $(\bar{y}_{KBi,(i=1,2\&3)})$ discussed as follows.

Theorem (4.1): The bias of the proposed ratio regression type estimators $\bar{y}_{Mi}$, $i = 1,2\ and\ 3$ is given by:

$$\text{Bias}\,(\bar{y}_{KBi}) = \bar{Y}f_i C_x^2\frac{k}{2}\left[1-k-2\rho_{xy}\frac{C_y}{C_x}\right] \tag{4.4}$$

where $k = \frac{\rho_{xy}S_y-\beta_1 S_x}{RS_x}$

and $f_1 = f_n, f_2 = f_r,\ f_3 = f_{rn}$.

***Proof*:** Proof is given in Appendix-1.

Theorem (4.2): The minimum mean square error of the proposed estimators T_{KB_i}, $i = 1,2,3$ is given by

$$MSE(T_{KBi}) = \left[f_r S_y^2 + f_i \left(S_x^2(kR + \beta_1)^2 - 2\rho_{xy} S_x S_y (kR + \beta_1)\right)\right]$$
$$\text{i} = 1,2,3 \ldots\ldots \tag{4.5}$$

For the optimum value k given by

$$k = \frac{\rho_{xy} S_y - \beta_1 S_x}{R S_x}$$

where,

$$R = \frac{\bar{Y}}{\bar{X}}\ , \beta_1 = \frac{S_x^2}{S_x S_y} \text{ and}$$

$$f_n = \left(\frac{1}{n} - \frac{1}{N}\right), f_r = \left(\frac{1}{r} - \frac{1}{N}\right) \ \ \& \ \ f_{rn} = \left(\frac{1}{r} - \frac{1}{n}\right).$$

The minimum MSE of the proposed estimator is given by

$$\text{Min MSE}(\bar{y}_{KBi}) = S_y^2\left[f_r - f_i * {\rho_{xy}}^2\right].$$

***Proof*:** Proof is given in Appendix-1.

5. A New Method of Imputation

The estimators rely on three different ratio-regression type methods of imputation as follows.

Case I: Auxiliary information on X is completely available, i.e., $\bar{X}$ is known and corresponding estimates $\bar{x}_n$ are used in the imputation technique.

$$y_{.i1} = \begin{cases} y_i & if \quad i \epsilon R \\ \dfrac{1}{n-r}\left[\left[n\gamma_1 \bar{y}_r \left\{2 - \left(\dfrac{\bar{x}_n}{\bar{X}}\right)^k\right\}\right] + \beta_1(\bar{X} - \bar{x}_n) - \bar{y}_r\, r\right] & if \quad i \epsilon R^c \end{cases}$$

The resultant estimator of population mean $\bar{Y}$ is given as

$$\bar{y}_{KN1} = \left[\gamma_1 \bar{y}_r \left\{2 - \left(\frac{\bar{x}_n}{\bar{X}}\right)^k\right\}\right] + \beta_1(\bar{X} - \bar{x}_n) \tag{5.1}$$

Case II: Auxiliary information on X is completely available i.e., $\bar{X}$ is known and corresponding estimates $\bar{x}_r$ are used in the imputation technique.

$$y_{.i2} = \begin{cases} y_i & if \quad i \epsilon R \\ \dfrac{1}{n-r} & \left[n\gamma_1 \bar{y}_r \left\{2 - \left(\dfrac{\bar{x}_r}{\bar{X}}\right)^k\right\} + \beta_1(\bar{X} - \bar{x}_r) - \bar{y}_r\, \text{r}\right] \quad if \quad i \epsilon R^c \end{cases}$$

The resultant estimator of population mean $\bar{Y}$ is given as

$$\bar{y}_{KN2} = \left[\gamma_1 \bar{y}_r \left\{2 - \left(\frac{\bar{x}_r}{\bar{x}}\right)^k\right\}\right] + \beta_1(\bar{X} - \bar{x}_r) \tag{5.2}$$

Case III: Auxiliary information on X is not available at the population level, i.e., $\bar{X}$ is not known and we use corresponding estimates $\bar{x}_n$, $\bar{x}_r$ in the imputation technique.

$$y_{.i3} = \begin{cases} y_i & if \quad i\epsilon R \\ \dfrac{1}{n-r}\left[n\gamma_1 \bar{y}_r \left\{2 - \left(\dfrac{\bar{x}_r}{\bar{x}_n}\right)^k\right\} + \beta_1(\bar{x}_n - \bar{x}_r) - \bar{y}_r r\right] & if \quad i\epsilon R^c \end{cases}$$

The resultant estimator of population mean $\bar{Y}$ is given as

$$\bar{y}_{KN3} = \left[\gamma_1 \bar{y}_r \left\{2 - \left(\frac{\bar{x}_r}{\bar{x}_n}\right)^k\right\}\right] + \beta_1(\bar{x}_n - \bar{x}_r) \tag{5.3}$$

Therefore, under the above situations, the properties of imputation methods discussed are as follow.

Theorem (5.1): The bias of the proposed ratio regression type estimators $\bar{y}_{KNi}$, $i = 1,2\ and\ 3$ is given by:

$$\text{Bias}\ (\bar{y}_{KNi}) = \left[(\gamma_1 - 1) - k f_i \left\{C_y^2 - \frac{(k-1)}{2} C_x^2\right\}\right] \tag{5.4}$$

where $k = \frac{\rho_{xy} S_y - \beta_1 S_x}{R S_x}$

and $f_1 = f_n, f_2 = f_r,\ f_3 = f_{rn}$.

Proof: Proof is given in Appendix-2.

Theorem (5.2): The minimum mean square error of the proposed ratio regression type estimators $\bar{y}_{KNi}\ i = 1,2$ and 3 is given by

$$\text{MSE}\ (\bar{y}_{KNi}) = \bar{Y}^2(\gamma_1^2 A_i - 2\gamma_1 B_i + C_i) \qquad \text{i} = 1,2,3 \ldots\ldots \tag{5.5}$$

For the optimum value k given by

$$k = \frac{\rho_{xy} S_y - \beta_1 S_x}{R S_x}$$

where, $A_i = 1 + f_r C_y^2 + f_i(k C_x^2 - 4k\rho_{xy} C_x C_y)$

$$B_i = 1 - f_i\left(k\rho_{xy} C_x C_y - k\beta_1 \frac{\bar{X}}{\bar{Y}} C_x^2 + \beta_1 \frac{\bar{X}}{\bar{Y}} \rho_{xy} C_x C_y - \frac{k(k-1)}{2} C_x^2\right)$$

$$C_i = 1 + \beta_1^2 \frac{\bar{X}^2}{\bar{Y}^2} f_i C_x^2$$

and $\gamma_{1opt} = \frac{B_i}{A_i}$

The minimum MSE of the proposed estimator is given by

$$Min\ MSE(\bar{y}_{KNi}) = \bar{Y}^2\left(C_i - \frac{B_i^2}{A_i}\right).$$

Proof: Proof is given in Appendix-2.

6. Efficiency Comparison

The following conditions are derived for the theoretical comparison of the Mean Squared Error (MSE) of the proposed estimator with other existing estimators.

Strartegy I:

$$V(\bar{y}_r) - MSE(\bar{y}_{KB1}) = \left(\frac{1}{n} - \frac{1}{N}\right)\rho_{xy}^2 S_y^2 \geq 0$$
$$MSE(\bar{y}_{REG}) - MSE(\bar{y}_{KB1}) = \left(\frac{2}{n} - \frac{1}{r} - \frac{1}{N}\right)\rho_{xy}^2 S_y^2 \geq 0$$
$$MSE(\bar{y}_{DP1}) - MSE(\bar{y}_{KB1}) = 0$$
$$MSE(\bar{y}_{DP2}) - MSE(\bar{y}_{KB1}) = \left(\frac{1}{n} - \frac{1}{r}\right)\rho_{xy}^2 S_y^2 \geq 0$$
$$MSE(\bar{y}_{DP3}) - MSE(\bar{y}_{KB1}) = \left(\frac{2}{n} - \frac{1}{r} - \frac{1}{N}\right)\rho_{xy}^2 S_y^2 \geq 0$$
$$MSE(\bar{y}_{GIRA}) - MSE(\bar{y}_{KB1}) = \left(\frac{2}{n} - \frac{1}{r} - \frac{1}{N}\right)\rho_{xy}^2 S_y^2 \geq 0$$

Strategy II:

$$V(\bar{y}_r) - MSE(\bar{y}_{KB2}) = \left(\frac{1}{r} - \frac{1}{N}\right)\rho_{xy}^2 S_y^2 \geq 0$$
$$MSE(\bar{y}_{REG}) - MSE(\bar{y}_{KB2}) = \left(\frac{1}{n} - \frac{1}{N}\right)\rho_{xy}^2 S_y^2 \geq 0$$
$$MSE(\bar{y}_{DP1}) - MSE(\bar{y}_{KB2}) = \left(\frac{1}{r} - \frac{1}{n}\right)\rho_{xy}^2 S_y^2 \geq 0$$
$$MSE(\bar{y}_{DP2}) - MSE(\bar{y}_{KB2}) = 0$$
$$MSE(\bar{y}_{DP3}) - MSE(\bar{y}_{KB2}) = \left(\frac{1}{n} - \frac{1}{N}\right)\rho_{xy}^2 S_y^2 \geq 0$$
$$MSE(\bar{y}_{GIRA}) - MSE(\bar{y}_{KB2}) = \left(\frac{1}{n} - \frac{1}{N}\right)\rho_{xy}^2 S_y^2 \geq 0$$

Strategy III:

$$V(\bar{y}_r) - MSE(\bar{y}_{KB3}) = \left(\frac{1}{r} - \frac{1}{n}\right)\rho_{xy}^2 S_y^2 \geq 0$$
$$MSE(\bar{y}_{REG}) - MSE(\bar{y}_{KB3}) = 0$$
$$MSE(\bar{y}_{DP1}) - MSE(\bar{y}_{KB3}) = \left(\frac{1}{r} + \frac{1}{N} - \frac{2}{n}\right)\rho_{xy}^2 S_y^2 \geq 0$$
$$MSE(\bar{y}_{DP2}) - MSE(\bar{y}_{KB3}) = \left(\frac{1}{N} - \frac{1}{n}\right)\rho_{xy}^2 S_y^2 \geq 0$$
$$MSE(\bar{y}_{DP3}) - MSE(\bar{y}_{KB3}) = 0$$
$$MSE(\bar{y}_{GIRA}) - MSE(\bar{y}_{KB3}) = 0$$

Comparing the proposed estimators, even if they involve different source of information, after simple algebra we note that:

$$MSE(\bar{y}_{KB1}) - MSE(\bar{y}_{KB2}) = \left(\frac{1}{r} - \frac{1}{N}\right)\rho_{xy}^2 S_y^2 \geq 0$$
$$MSE(\bar{y}_{KB3}) - MSE(\bar{y}_{KB2}) = \left(\frac{1}{n} - \frac{1}{N}\right)\rho_{xy}^2 S_y^2 \geq 0$$
$$MSE(\bar{y}_{KB1}) - MSE(\bar{y}_{KB3}) = \left(\frac{1}{r} + \frac{1}{N} - \frac{2}{n}\right)\rho_{xy}^2 S_y^2 \geq 0$$

This means that $\bar{y}_{KB2}$ is always more efficient than both $\bar{y}_{KB1}$ and $\bar{y}_{KB3}$, whereas $\bar{y}_{KB3}$ performs better than $\bar{y}_{KB1}$ if the condition $\boldsymbol{r < \frac{nN}{2N-n}}$ is satisfied. The results are

valuable because they highlight the role of the auxiliary information in improving the estimates and afford sampling practitioners a useful indication on a profitable collection of auxiliary information in the case of missing data. The choice among competing estimators can be certainly facilitated by awareness of the information at hand.

7. Computational Study

We have divided the computations into two categories, namely, with real data and artifically generated data.

7.1. Empirical study using real data

In this section, an empirical study is performed in the presence of auxiliary variable where the performance of the proposed methods of imputation is compared with competing methods based on MSE and PRE. This study is carried out on six real data sets. We have computed and reported MSEs and percentage relative efficiency (PREs) of the proposed imputation methods with respect to the conventional methods to compare the proposed imputation methods with that of the existing imputation method that utilizes auxiliary information.

$$PRE(\bar{y}_{KBi}, \bar{y}_r) = \frac{MSE(\bar{y}_r)}{MSE(\bar{y}_{KBi})} \times 100 \quad and \quad PRE(\bar{y}_{KNi}, \bar{y}_r) = \frac{MSE(\bar{y}_r)}{MSE(\bar{y}_{KNi})} \times 100$$

Six different real data sets have been considered in the present empirical study. Data set 1 is taken from, Kadilar & Cingi (2008) with details on y as the level of apple production, and x as the number of apple trees.Data set 2 is taken from Diana & Perri (2010) with information on the Survey of Households Income and Wealth conducted by the Bank of Italy for the year 2002, y as the household's net disposable income, x as the number of household income earners. Data set 3 is taken from Source: [7] Page 228. The source of data set 4 is Singh (2009). The data set 5 is taken from Srivastava et. al. (1989) pp. 3922): y of weight of children, x as the skull circumference of children. Data set 6 is taken from ICMR, Department of Pediatrics, BHU, during 1983-84 of school children with study variable y as height (in kg) of the children, x, variable related to weight. The required values of the parameters for all six data sets are given in table 1.

Table 1: Population Parameters of Six Different Real Population.

Parameter	Population 1	Population 2	Population 3	Population 4	Population 5	Population 6
N	19	8011	80	3055	82	95
n	10	400	20	611	43	35
r	8	250	16	520	25	10
$\bar{Y}$	575	28229.43	51.8264	308582.4	11.90	115.9526
$\bar{X}$	13573.68	1.69	2.8513	56.5	39.80	19.4968
S_y	858.36	22216.56	18.3569	425312.8	0.5792685	5.966921
S_X	12945.38	0.78	2.7041	72.3	0.8581212	3.27346
ρ_{xy}	0.88	0.46	0.9150	0.677	0.009	0.713

These population have varying amount of correlation between study variate(y) and auxiliary variate(x) as shown in the table 1.

Table 2: Mean Square Errors of the Existing and Suggested Estimators

Estimator	Population 1	Population 2	Population 3	Population 4	Population 5	Population 6
Case I						
$\bar{y}_r$	53319.74	191268.93	16.85	288655815.71	0.00932998	3.185634
$\bar{y}_{RAT1}$	47051.53	1683371.53	76.31	290511133.40	0.01066795	3.471515
$\bar{y}_{REG1}$	45934.86	1664625.61	14.17	238539244.96	0.00929992	3.095766
$\bar{y}_{SH1}$	45934.86	1664625.61	14.17	238539244.96	0.00929992	3.095766
$\bar{y}_{SD1}$	45934.86	1664625.61	14.17	238539244.96	0.00929992	3.095766
$\bar{y}_{SINGH1}$	45934.86	1664625.61	14.17	238539244.96	0.00929992	3.095766
$\bar{y}_{DP1}$	45934.86	1664625.61	14.17	238539244.96	0.00929992	3.095766
$\bar{y}_{GIRA1}$	45934.86	1664625.61	14.17	238539244.96	0.00929992	3.095766
$\bar{y}_{KB1(proposed)}$	45934.86	1664625.61	14.17	238539244.96	0.00929992	3.095766
$\bar{y}_{KN1(proposed)}$	28448.79	1209049.49	8.84	190771840.36	0.00890512	2.949472
Case II						
$\bar{y}_r$	53319.74	191268.93	16.85	288655815.71	0.009329982	3.185634
$\bar{y}_{RAT2}$	43743.31	1538549.33	96.14	290916984.15	0.01269343	4.603127
$\bar{y}_{REG2}$	42037.28	1507964.74	13.28	227576245.10	0.00925441	2.74004
$\bar{y}_{SH2}$	42037.28	1507964.74	13.28	227576245.10	0.00925441	2.74004
$\bar{y}_{SD2}$	42037.28	1507964.74	13.28	227576245.10	0.00925441	2.74004
$\bar{y}_{SINGH2}$	42037.28	1507964.74	13.28	227576245.10	0.00925441	2.74004
$\bar{y}_{DP2}$	42037.28	1507964.74	13.28	227576245.10	0.00925441	2.74004
$\bar{y}_{GIRA2}$	42037.28	1507964.74	13.28	227576245.10	0.00925441	2.74004
$\bar{y}_{KB2(proposed)}$	42037.28	1507964.74	13.28	227576245.10	0.00925441	2.74004
$\bar{y}_{KN2(proposed)}$	18811.33	767278.91	6.20	169535367.50	0.008266157	2.017615
Case III						
$\bar{y}_r$	53319.74	191268.93	16.85	288655815.71	0.009329982	3.185634
$\bar{y}_{RAT3}$	50011.52	1767867.72	36.67	289061666.45	0.01135546	4.317246
$\bar{y}_{REG3}$	49422.17	1756029.05	15.96	277692815.85	0.009284472	2.829908
$\bar{y}_{SH3}$	49422.17	1756029.05	15.96	277692815.85	0.009284472	2.829908
$\bar{y}_{SD3}$	49422.17	1756029.05	15.96	277692815.85	0.009284472	2.829908
$\bar{y}_{SINGH3}$	49422.17	1756029.05	15.96	277692815.85	0.009284472	2.829908
$\bar{y}_{DP3}$	49422.17	1756029.05	15.96	277692815.85	0.009284472	2.829908
$\bar{y}_{GIRA3}$	49422.17	1756029.05	15.96	277692815.85	0.009284472	2.829908
$\bar{y}_{KB3(proposed)}$	49422.17	1756029.05	15.96	277692815.85	0.009284472	2.829908
$\bar{y}_{KN3(proposed)}$	36826.3	1466695.66	14.11	266571886.52	0.008688242	2.253037

Table 3: Percentage Relative Efficiency of the Considered Estimators under Six Different populations

Estimator	Population 1	Population 2	Population 3	Population 4	Population 5	Population 6
Case I						
$\bar{y}_r$	100	100	100	100	100	100
$\bar{y}_{RAT1}$	152.7992	113.6226	22.07836	99.36136	87.45805	91.76495
$\bar{y}_{REG1}$	154.1536	114.9021	118.8637	121.0098	100.3233	102.9029
$\bar{y}_{SH1}$	154.1536	114.9021	118.8637	121.0098	100.3233	102.9029
$\bar{y}_{SD1}$	154.1536	114.9021	118.8637	121.0098	100.3233	102.9029
$\bar{y}_{SINGH1}$	154.1536	114.9021	118.8637	121.0098	100.3233	102.9029
$\bar{y}_{DP1}$	154.1536	114.9021	118.8637	121.0098	100.3233	102.9029
$\bar{y}_{GIRA1}$	154.1536	114.9021	118.8637	121.0098	100.3233	102.9029
$\bar{y}_{KB1(proposed)}$	154.1536	114.9021	118.8637	121.0098	100.3233	102.9029
$\bar{y}_{KN1(proposed)}$	216.947	158.1978	190.683	151.3094	104.7709	108.0069
Case II						
$\bar{y}_r$	100	100	100	100	100	100
$\bar{y}_{REG2}$	215.8442	126.8392	126.8391	126.8392	100.8166	112.8697
$\bar{y}_{SH2}$	215.8442	126.8392	126.8391	126.8392	100.8166	112.8697
$\bar{y}_{SD2}$	215.8442	126.8392	126.8391	126.8392	100.8166	112.8697
$\bar{y}_{SINGH2}$	215.8442	126.8392	126.8391	126.8392	100.8166	112.8697
$\bar{y}_{DP2}$	215.8442	126.8392	126.8391	126.8392	100.8166	112.8697
$\bar{y}_{GIRA2}$	215.8442	126.8392	126.8391	126.8392	100.8166	112.8697
$\bar{y}_{KB2(proposed)}$	215.8442	126.8392	126.8391	126.8392	100.8166	112.8697
$\bar{y}_{KN2(proposed)}$	435.1715	249.2822	271.9538	170.2629	112.8697	157.8911
Case III						
$\bar{y}_r$	100	100	100	100	100	100
$\bar{y}_{RAT3}$	122.305	108.1919	45.94658	99.8596	82.16294	73.78856
$\bar{y}_{REG3}$	122.76	108.9213	105.5855	103.9479	100.4902	112.5702
$\bar{y}_{SH3}$	122.76	108.9213	105.5855	103.9479	100.4902	112.5702
$\bar{y}_{SD3}$	122.76	108.9213	105.5855	103.9479	100.4902	112.5702
$\bar{y}_{SINGH3}$	122.76	108.9213	105.5855	103.9479	100.4902	112.5702
$\bar{y}_{DP3}$	122.76	108.9213	105.5855	103.9479	100.4902	112.5702
$\bar{y}_{GIRA3}$	122.76	108.9213	105.5855	103.9479	100.4902	112.5702
$\bar{y}_{KB3(proposed)}$	122.76	108.9213	105.5855	103.9479	100.4902	112.5702
$\bar{y}_{KN3(proposed)}$	152.4202	130.4081	119.4088	108.2844	107.3863	141.3929

7.2. Artificial population

The artificial population has been generated as described below

Population1. A population of size $N = 500$ with one study variable y and one auxiliary variable x is generated from the bivariate normal distribution where study variable y is correlated with auxiliary variables with various amount of $\rho_{yx} = 0.6, 0.7, 0.8$ and 0.9. The variables (y, x) are generated using MVNORM package in R software. A sample of size $n = 50$ is drawn from the population, with the number of responding units assumed to be $r = 30$.

Population 2. An artificial population is generated of size $N = 200$ which involves one study variable y and auxiliary variable x. Study variable y is correlated with auxiliary variables with various amount of $\rho_{yx} = 0.6, 0.7, 0.8$ and 0.9 The variables (y, x) is generated using MVNORM package in R software. From this population we draw sample of size $n = 26$, responding units are $r = 21$.

The percentage relative efficiencies (PRE) of the proposed estimators are computed through 50,000 repeated samples of size n as per imputation technique. In which we (i) draw a random sample of size n from population size N, (ii) from each selected sample $(n - r)$ units are dropped randomly, and (iii) the estimators and their MSE's are calculated for each sample and then averaged over all 50,000 samples .

The mean square error and percent relative efficiencies are given by

$$MSE(T_j) = \frac{1}{50000} \sum_{i=1}^{50000} \left(T_j(s_i) - \bar{Y}\right)^2 \quad j = 0, 1, 2, 3$$

$$PRE(T_j) = \frac{MSE(T_0)}{MSE(T_j)} \times 100 \quad j = 1, 2, 3$$

based on 50,000 repeated samples.

Table 4: Mean square error and percentage relative efficiency based on

Population 1 (Artificially generated normal population)								
(N=500, n=50, r=30)								
Correlation	0.9		0.8		0.7		0.6	
Estimator	MSE	PRE	MSE	PRE	MSE	PRE	MSE	PRE
Mean per unit($\bar{y}_r$)	3.1962	100	3.0434	100	3.23365	100	2.62266	100
Ratio method ($\bar{y}_{RAT}$)	2.3856	133.978	2.4749	122.973	2.83633	114.008	2.34778	111.7081
Regression Method ($\bar{y}_{REG}$)	2.2519	141.9321	2.4259	125.455	2.85815	113.138	2.35706	111.2686
Proposed imputation ($\bar{y}_{KNi}$)	2.1520	148.5223	2.2456	135.529	2.44280	132.375	2.13541	122.8176

Table 4: Mean square error and percentage relative efficiency based on (cont.)

Population 2 (Artificially generated normal population)								
(N=200, n=26, r=21)								
Correlation	0.9		0.8		0.7		0.6	
Estimator	MSE	PRE	MSE	PRE	MSE	PRE	MSE	PRE
Mean per unit($\bar{y}_r$)	4.7080	100	4.5813	100	4.5712	100	4.0428	100
Ratio method ($\bar{y}_{RAT}$)	4.1007	114.810	4.1094	111.484	4.1587	109.9177	3.8485	105.049
Regression Method ($\bar{y}_{REG}$)	4.0259	116.943	4.0721	112.504	4.1552	110.0103	3.8514	104.972
Proposed imputation ($\bar{y}_{KNi}$)	3.7094	126.921	3.5807	127.944	3.7578	121.6439	3.4201	118.208

8. Interpretations of the Computational Results

In this article, it is clear that MSE of the proposed alternative estimator is more efficient than the mean estimator. In addition, the proposed estimator is always more efficient than the usual ratio estimator. We note that the proposed method is free from the assumptions of a model for the ratio method of imputation. In addition, MSE is similar to the other mentioned estimators MSE ($\bar{y}_{SH}$) = MSE ($\bar{y}_{SD}$) = MSE ($\bar{y}_{Singh}$) = MSE($\bar{y}_{DPi}$) = MSE($\bar{y}_{GIRA}$) = MSE($\bar{y}_{KBi}$). A new method of imputation is introduced that remains more efficient than conventional and existing imputation methods in the presence of auxiliary variables. The following interpretations are made based on empirical results summarized in Table 3.

1. We introduce an alternative method of imputation and the resultant estimator in the presence of non-response. The performance of the proposed estimator is justified theoretically and numerically. Table 2 & 3 expressed that the relative efficiency of the proposed estimator ($\bar{\boldsymbol{y}}_{\boldsymbol{KB1}}, \bar{\boldsymbol{y}}_{\boldsymbol{KB2}}$ *and* $\bar{\boldsymbol{y}}_{\boldsymbol{KB3}}$) performs better than the mean and the ratio estimators and is equivalent to other mentioned estimators.
2. For all the populations 1 to 6, Table 2 & 3 exhibit the superiority of the proposed imputation method ($\bar{\boldsymbol{y}}_{\boldsymbol{KN1}}, \bar{\boldsymbol{y}}_{\boldsymbol{KN2}}$ *and* $\bar{\boldsymbol{y}}_{\boldsymbol{KN3}}$) over the mean and ratio type imputation method. Also, the proposed method of imputation ($\bar{\boldsymbol{y}}_{\boldsymbol{KN1}}, \bar{\boldsymbol{y}}_{\boldsymbol{KN2}}$ *and* $\bar{\boldsymbol{Y}}_{\boldsymbol{KN3}}$) is superior to the Diana and Perri regression type imputation method and the proposed alternative method ($\bar{\boldsymbol{y}}_{\boldsymbol{M1}}, \bar{\boldsymbol{y}}_{\boldsymbol{M2}}$ *and* $\bar{\boldsymbol{y}}_{\boldsymbol{M3}}$) of imputation.
3. The proposed new imputation method ($\bar{\boldsymbol{y}}_{\boldsymbol{KN1}}, \bar{\boldsymbol{y}}_{\boldsymbol{KN2}}$ *and* $\bar{\boldsymbol{y}}_{\boldsymbol{KN3}}$), in all the populations 1 to 6, as shown in Table 2 & 3, has achieved considerable gain

in performance over the conventional imputation method for all the three cases, namely $\bar{y}_{KN1}$ is considerably better than $\bar{y}_{SH}$(I), $\bar{y}_{SD}$(I), and $\bar{y}_{DP}$(I), in case I. Similarly, $\bar{y}_{KN2}$ is considerably superior to $\bar{y}_{SH}$(II), $\bar{y}_{SD}$(II), and $\bar{y}_{DP}$(II), imputation methods in case II and $\bar{y}_{KN3}$ is considerably superior to $\bar{y}_{SH}$(III), $\bar{y}_{SD}$(III), and $\bar{y}_{DP}$(III), in case III.

4. It is important to note that Table 2 & 3 exhibit that the proposed imputation method ($\bar{\boldsymbol{y}}_{\boldsymbol{KN1}}, \bar{\boldsymbol{y}}_{\boldsymbol{KN2}}$ *and* $\bar{\boldsymbol{y}}_{\boldsymbol{KN3}}$) when applied to six real data sets, and compared with conventional and recent imputation methods, attains considerable gain in efficiency over competing for imputation methods for the case II, and gives better results over other cases, namely, case I and case III.
5. This empirical study confirms the superiority of the proposed imputation method over Diana & Perri's (2010) imputation method and other landmark imputation methods that use auxiliary information.

It can be noted that the proposed method of imputation ($\bar{\boldsymbol{y}}_{\boldsymbol{KN1}}, \bar{\boldsymbol{y}}_{\boldsymbol{KN2}}$ *and* $\bar{\boldsymbol{y}}_{\boldsymbol{KN3}}$) is easy to use and rewarding in terms of efficiency and deals with the problem of non-response. Survey practitioners can use the proposed imputation method to deal with the problem of nonresponse and get a high gain in efficiency when one has access to auxiliary information.

References

Kadilar, C., Cingi, H., (2008). Estimators for the Population Mean in the Case of Missing Data. *Communications in Statistics—Theory and Methods*, 37, pp. 2226–2236. http://dx.doi.org/10.1080/03610920701855020.

Diana, G., Perri, P. F., (2010). Improved Estimators of the Population Mean for Missing Data, *Communications in Statistics—Theory and Methods*, 39, pp. 3245–3251. http://dx.doi.org/10.1080/03610920903009400.

Gira, Abdeltawab, A., (2015). Estimation of population mean with a new imputation methods. *Applied Mathematical Sciences*, 9(34), pp. 1663–1672.

Chodjuntug K., Lawson, N., (2022). Imputation for estimating the population mean in the presence of nonresponse, with application to fine particle density in Bangkok. *Mathematical Population Studies*, 29(4), pp. 204–225.

Chodjuntug K., Lawson, N., (2022). A chain regression exponential type imputation method for mean estimation in the presence of missing data. *Songklanakarin Journal of Science and Technology*, 44 (4), pp. 1109–1118.

Heitzan, D. F., Basu, S., (1996). Distinguishing 'Missing at Random' and 'Missing Completely At Random'. *The American Statistician,* 50, pp. 207–213.

Kalton, G., Kasprzyk, D., (1982). Imputing for missing survey responses, In: Proceedings of the section on survey research method. *American Statistical Association*, pp. 22–31.

Kalton, G., Kasprzyk, D. and Santos, R., (1981). Issues of nonresponse and imputation in the survey of income and program participation, In: Krewski D, Platek R, Rao JNK (eds) Current topics in survey sampling. *Academic Press, New York*, pp. 455–480.

Lee, H., Rancourt, E. and Särndal, C. E., (1994). Experiments with Variance Estimation from Survey Data with Imputed Values. *Journal of Official Statistics,* 10, pp. 231–243.

Lee, H., Rancourt, E., Sarndal, C. E., (1995). Variance estimation in the presence of imputed data for the generalized estimation system. *In: Proceedings of the section on survey research methods*, American Statistical Association.

Murthy, M. N., (1967). Sampling theory and methods. *Statistical publishing Society*, Calcutta, India.

Lawson, N., (2023). New imputation method for estimating population mean in the presence of missing data. *Lobachevskii Journal of Mathematics*, 44(9), pp. 3740–3748.

Lawson, N., (2023). A class of population mean estimators in the presence of missing data with applications to air pollution in Chiang Mai, Thailand. *Lobachevskii Journal of Mathematics*, 44(9), pp. 3749–3757.

Thongsak, N., Lawson, N., (2023). A new imputation method for population mean in the presence of missing data based on a transformed variable with applications to air pollution data in Chiang Mai, Thailand. *Journal of Air Pollution and Health*, 8(3), pp. 285–298.

Rao, J. N. K, Sitter, R. R., (1995). Variance Estimation under Two-Phase Sampling with Application to Imputation for Missing Data. *Biometrika*, 82, pp. 453–460. http://dx.doi.org/10.1093/biomet/82.2.453

Rubin, R. B., (1976). Inference and missing data. *Biometrika*, 63, pp. 581–592. http://dx.doi.org/10.1093/biomet/63.3.581

Singh, S., (2009). A new method of imputation in survey sampling. Statistics: *A Journal of Theoretical and Applied Statistics*, 43, pp. 499–511. http://dx.doi.org/ 10.1080/ 02331880802605114.

Singh, S., Deo, B., (2003). Imputation by Power Transformation. *Statistical Papers*, 44, pp. 555–579. http://dx.doi.org/10.1007/bf02926010.

Singh, S., Horn, S., (2000). Compromised Imputation in Survey Sampling. *Metrika*, 51, pp. 267–276. http://dx.doi.org/10.1007/s001840000054.

Srivenkataramana, T., Tracy, D. S., (1980). An Alternative to Ratio Method in Sample Surveys. *Annals of the Institute of Statistical Mathematics*, 32, pp. 111–120. http://dx.doi.org/10.1007/ bf02480317.

Appendix-1

Proof: proof of theorem 4.1(Bias of the proposed estimator). The line of proof here is worked out for the estimator defined under case 1, The estimator $\bar{y}_{KN1}$ can be written be as follows.

$$\begin{aligned}\bar{y}_{KN1} &= \bar{Y}(1+\varepsilon_o)[2-(1+\eta_o)^K]-\beta_1\bar{X}\eta_o \\ &= \bar{Y}(1+\varepsilon_o)\left[2-(1+K\eta_o+\tfrac{K(K-1)}{2}\eta_o^2+\cdots)\right]-\beta_1\bar{X}\eta_o \\ &= \bar{Y}(1+\varepsilon_o)\left(1-K\eta_o-\tfrac{K(K-1)}{2}\eta_o^2+\cdots\right)-\beta_1\bar{X}\eta_o \qquad (A) \\ &= \bar{Y}\left(1+\varepsilon_o-K\eta_o-K\varepsilon_o\eta_o-\tfrac{K(K-1)}{2}\eta_o^2-\beta_1\tfrac{\bar{X}}{\bar{Y}}\eta_o+O(\varepsilon_o^2)\right)\end{aligned}$$

Neglecting the higher order of approximation, the bias

$$\begin{aligned}B(\bar{y}_{KN1}) &= E(\bar{y}_{KN1}-\bar{Y}) \\ &= \bar{Y}E\left(\varepsilon_o-K\eta_o-K\varepsilon_o\eta_o-\tfrac{K(K-1)}{2}\eta_o^2-\beta_1\tfrac{\bar{X}}{\bar{Y}}\eta_o\right) \qquad (B)\end{aligned}$$

Taking the expectations of (B), we get (4.4) for i = 1, which prove theorem (4.1).

The derivation of other estimators $\bar{y}_{KNi}$ $(i = 2\ \&\ 3)$ can be carried out in a similar way.

Proof: proof of theorem 4.2

The MSE of $\bar{y}_{M1}$ can be found up to the first order of approximation by rewriting as follow:

$$\begin{aligned}MSE\ (\bar{y}_{KN1}) &= E(\bar{y}_{KN1}-\bar{Y})^2 \\ &= \bar{Y}^2E\left[\varepsilon_o-K\eta_o-K\varepsilon_o\eta_o-\tfrac{K(K-1)}{2}\eta_o^2-\beta_1\tfrac{\bar{X}}{\bar{Y}}\eta_o\right]^2 \\ &= \bar{Y}^2E\left[\varepsilon_o^2+K^2\eta_o^2+\beta_1^2\tfrac{\bar{X}^2}{\bar{Y}^2}\eta_o^2-2K\varepsilon_o\eta_o+2K\beta_1\tfrac{\bar{X}}{\bar{Y}}\eta_o^2-2\beta_1\tfrac{\bar{X}}{\bar{Y}}\varepsilon_o\eta_o\right] \\ &= \left[\left(\tfrac{1}{r}-\tfrac{1}{N}\right)S_y^2+\left(\tfrac{1}{n}-\tfrac{1}{N}\right)\left(S_x^2(KR+\beta_1)^2-2\rho_{xy}S_xS_y(KR+\beta_1)\right)\right] \quad \text{.....(C)}\end{aligned}$$

where, R = $\frac{\bar{Y}}{\bar{X}}$

Differentiating equation (C) with respect to K and equating to zero, we get

$$K=\frac{\rho_{xy}S_y-\beta_1S_x}{RS_x}=K(optimum)$$

then substitute the value of optimum K in equation (4.8), thus the resulting minimum mean square error of $\bar{y}_{M1}$ is given by

$$\text{Min MSE}(\bar{y}_{KN1})=S_y^2\left[f_r-f_n*{\rho_{xy}}^2\right]$$

Appendix-2

Proof: Proof of theorem 5.1

The estimator $\bar{y}_{KB1}$ can be written be as follows.

$$\begin{aligned}\bar{y}_{KB1} &= \gamma_1\bar{Y}(1+\varepsilon_o)[2-(1+\eta_o)^K]-\beta_1\bar{X}\eta_o \\ &= \gamma_1\bar{Y}(1+\varepsilon_o)\left[2-(1+K\eta_o+\tfrac{K(K-1)}{2}\eta_o^2+\cdots)\right]-\beta_1\bar{X}\eta_o \\ &= \gamma_1\bar{Y}(1+\varepsilon_o)\left(1-K\eta_o-\tfrac{K(K-1)}{2}\eta_o^2+\cdots\right)-\beta_1\bar{X}\eta_o \qquad \text{(D)} \\ &= \bar{Y}\left(\gamma_1+\gamma_1\varepsilon_o-\gamma_1K\eta_o-\gamma_1K\varepsilon_o\eta_o-\gamma_1\tfrac{K(K-1)}{2}\eta_o^2-\beta_1\tfrac{\bar{X}}{\bar{Y}}\eta_o+O(\varepsilon_o^2)\right)\end{aligned}$$

Neglecting the higher order of approximation, the bias

$$\begin{aligned}B(\bar{y}_{KB1}) &= (\bar{y}_{KB1}-\bar{Y}) \\ &= \bar{Y}E\left(\gamma_1+\gamma_1\varepsilon_o-\gamma_1K\eta_o-\gamma_1K\varepsilon_o\eta_o-\gamma_1\tfrac{K(K-1)}{2}\eta_o^2-\beta_1\tfrac{\bar{X}}{\bar{Y}}\eta_o-1\right) \qquad \text{(E)}\end{aligned}$$

Taking the expectations of (E), we get the (5.4) for $i=1$, which proves theorem (5.1)

The derivation of other estimators $\bar{y}_{KBi}$ $(i=2\ \&\ 3)$ can be carried out in a similar way

Theorem (5.2): The minimum mean square error of the proposed ratio regression type estimators $\bar{y}_{KB1}$ up to the first order approximation is given by

$$\text{MinMSE}\,(\bar{y}_{KB1}) = \bar{Y}^2\left[C_1-\left(\frac{B_1^2}{A_1}\right)\right] \qquad (5.7)$$

For the optimum value K given by

$$K=\frac{\rho_{xy}S_y-\beta_1S_x}{RS_x}$$

Proof: The MSE of $\bar{y}_{M11}$ can be found up to the first order of approximation by rewriting as follows:

$$MSE\,(\bar{y}_{KB1}) = E(\bar{y}_{KB1}-\bar{Y})^2$$

$$= \bar{Y}^2E\left[\gamma_1-\gamma_1K\eta_o-\gamma_1\tfrac{K(K-1)}{2}\eta_o^2+\gamma_1\varepsilon_o-\gamma_1K\varepsilon_o\eta_o-\beta_1\tfrac{\bar{X}}{\bar{Y}}\eta_o-1\right]^2$$

$$= \bar{Y}^2E\left[1+{\gamma_1}^2\left(1+\varepsilon_o^2+K\eta_o^2-4K\varepsilon_o\eta_o\right)-2\gamma_1\left(1-K\varepsilon_o\eta_o-K\beta_1\tfrac{\bar{X}}{\bar{Y}}\eta_o^2-\tfrac{K(K-1)}{2}\eta_o^2+\beta_1\tfrac{\bar{X}}{\bar{Y}}\varepsilon_o\eta_o\right)+\beta_1^2\tfrac{\bar{X}^2}{\bar{Y}^2}\eta_o^2\right]$$

$$= \bar{Y}^2\left[\begin{matrix}1+{\gamma_1}^2\{1+f_rC_y^2+f_n(KC_x^2-4K\rho_{xy}C_xC_y)\}-2\gamma_1 \\ \left\{1-f_n\left(K\rho_{xy}C_xC_y-K\beta_1\tfrac{\bar{X}}{\bar{Y}}C_x^2+\beta_1\tfrac{\bar{X}}{\bar{Y}}\rho_{xy}C_xC_y-\tfrac{K(K-1)}{2}C_x^2\right)\right\}+\beta_1^2\tfrac{\bar{X}^2}{\bar{Y}^2}f_nC_x^2\end{matrix}\right]..(F)$$

Differentiate equation (F) with respect to K when $\gamma_1$equating to 1, we get

$$K = \frac{\rho_{xy}S_y - \beta_1 S_x}{RS_x}$$

For optimum value of γ_1 differentiating the equation(G) with respect to and equating to zero, we get $\gamma_{1opt} = \frac{B_1}{A_1}$

Substituting the optimum value of γ_{1opt} in equation (G), we get minimum MSE

$$\text{Min. MSE}(\bar{y}_{K1}) = \bar{Y}^2 \left(C_1 - \frac{B_1^2}{A_1}\right)$$

we get the (5.7) for $i = 1$ that prove theorem (5.2)

The derivation of other estimators $T_{KBi\,(i=2\,\&\,3)}$ can be drive on similar lines.

In general, we have

$$MSE\ (\bar{y}_{KBi}) = \bar{Y}^2(\gamma_1^2 A_i - 2\gamma_1 B_i + C_i)$$