EFFICIENCY OF RATIO ESTIMATOR WITH RANDOM MISSING VALUES IN STRATIFIED TWO-STAGE SAMPLING USING DOUBLE SAMPLING FOR AUXILIARY INFORMATION

Iorlaha, P. I. [a*], Uba, T. [a], Nwaosu, S. C. [a] and Ikughur, A. J. [a]

[a] *Department of Statistics Joseph Saawuan Tarka University Makurdi-Nigeria*

Correspondence: iorlahapatrick@gmail.com

Abstract

Accurate estimation of population parameters in stratified two-stage sampling is often challenged by missing data, which can reduce the efficiency of estimators. Existing approaches, such as the Bahl and Saini (2011) ratio and difference estimators, do not explicitly account for missing observations, limiting their applicability in practical survey contexts. To address this gap, this study formulated and evaluated a ratio-type estimator of population mean in stratified two-stage sampling. Two populations were considered: a synthetically generated dataset from an exponential distribution and a field survey on school attendance and pupils' mathematics scores. Sample sizes of 25, 40, 70, and 100 were drawn to assess estimator's performance. To investigate efficiency under missing data, 20% of the observations were assumed Missing Completely at Random (MCAR) so that regression and ratio imputation methods were implemented. The developed estimator compared well with other results with efficiency assessed in terms of variance, standard error, coefficient of variation, and confidence intervals. Findings showed that the developed estimator consistently outperformed other estimators across all sample sizes, achieving lower coefficients of variation and narrower confidence intervals. Efficiency improved steadily with increasing sample size, and regression imputation provided superior recovery of efficiency compared to ratio imputation, particularly when strong correlations between auxiliary and study variables were present. The developed estimator demonstrated efficiency in both controlled and real-world survey conditions, making it a reliable alternative to existing estimators, especially in the presence of missing data, and highlighting its strong potential for application in practical survey contexts.

Keywords: Stratified sampling, ratio estimator, missing data, imputation, MCAR, complex surveys, hybrid imputation

1.0 Introduction

Survey sampling has evolved significantly since its early developments in the 20th century. Neyman (1934) laid the foundation for stratified sampling, demonstrating its superiority over simple random sampling in reducing variance. As large-scale surveys became more common, statisticians sought methods to further improve precision while minimizing costs. Stratified two-stage sampling emerged as a key strategy, allowing researchers to divide the population into homogeneous stratum before selecting primary sampling units (PSUs) and subsequently secondary sampling units (SSUs) (Cochran, 1977). This design has been instrumental in large-scale demographic, agricultural, and socio-economic surveys, where direct enumeration is impractical due to cost and logistical constraints (Lohr, 2021).

A significant breakthrough in estimation techniques came with the introduction of auxiliary information to enhance precision. The use of ratio and regression estimators gained traction in the mid-20th century as methods to improve population mean estimates when an auxiliary variable

ISSN NUMBER: 1116-249X

was available. Hansen *et al*. (1951) first demonstrated the effectiveness of ratio estimation when the study variable and auxiliary variable exhibit a proportional relationship, highlighting its ability to reduce variance. Over time, researchers expanded the application of ratio estimators to complex designs, including stratified two-stage sampling, where auxiliary data can significantly improve estimation accuracy (Singh and Mangat, 2013; Mukhopadhyay, 2018; Kim and Rao, 2012; Rao and Fuller, 2017).

Missing data on the other hand has been a persistent issue in survey sampling, with early treatments relying on simplistic approaches such as complete-case analysis, mean imputation which often leads to biased estimates when data are not missing completely at random (Robin 2020). Although considerable progress has been made in developing estimation procedures for stratified two-stage sampling, the incorporation of the estimators with imputation techniques remains relatively underexplored. The traditional Bahl and Saini (2011) ratio and difference estimators utilized double sampling to incorporate auxiliary variables at the first and secondary sampling unit level, and their findings illustrated substantial gains in estimator efficiency under the assumption of complete data. However, a critical challenge of their approach is the lack of consideration for incomplete observations, which are frequently encountered in applied survey research. Their estimators were not designed to accommodate missing data, nor do they incorporate any corrective mechanisms such as imputation. Consequently, the applicability of their methods in practical survey contexts where nonresponse is inevitable is limited. This highlights a pressing need to extend these methods by embedding imputation techniques within a design-based estimation framework that more accurately reflects the complexities of real world survey data.

This study seeks to address these gaps by developing a ratio-type estimator in a stratified two-stage sampling framework using double sampling for auxiliary information and also incorporates imputation techniques to address the inefficiency associated with nonresponse.

2.0 Some Existing Estimators

Several estimators have been proposed in survey sampling to enhance efficiency using auxiliary information. These estimators form the basis for comparison in this study, among which is the Bahl and Saini (2011) estimators are notable.

2.1 Bahl and Saini (2011) Estimator of Population Total in Two Stage Design

Bahl and Saini (2011) suggested following families of Estimator of Population Total in Two Stage Design with PPS sampling and using Double sampling for auxiliary information is given as;

$$\widehat{T}_2 = \sum_{k=1}^{p} a_k t_k \qquad (1)$$

Where $\widehat{T}_2 =$ is the estimated population total, $P =$ is the number of variables of interest such that $k = 1,2..p$, a_k = is the weights for combining multiple estimators such that $\sum_{k=1}^{p} a_k = 1$.

ISSN NUMBER: 1116-249X

$t_k =$ is the estimator for total of the k^{th} variable, defined as either a weighted ratio $t_{k,R}$ or weighted a difference estimator $t_{k,D}$ such that

$$t_{k,R} = \frac{\hat{Y}_{k(3)}}{\hat{X}_{k(3)}}\hat{X}_{k(2)}$$

$$t_{k,D} = \hat{Y}_{k(3)} + \lambda_k\left(\hat{X}_{k(3)} - \hat{X}_{k(2)}\right)$$

The bias, and mean square error of the ratio estimator are given as

$$B\left(\hat{T}_{2R}\right) = \sum_{k=1}^{p}\frac{a_k}{\bar{X}_k}\left[\frac{N}{n}\sum_{i}^{N}\frac{M_i}{m_i'm_i}\left\{R\sigma^2_{jx_k} - \sigma_{jyx_k}\right\}\right]$$

$$MSE\left(\hat{T}_{2R}\right) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)S_y^2 + \frac{N}{n}\sum_{i}^{N}M_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)S_{iy}^2 + \frac{N}{n}\sum_{i}^{N}\frac{M_i}{m_i'm_i}a'B^*a$$

With $B^* = \left[\sigma^2_{jy} - R_k\sigma_{jyx_k} - R_l\sigma_{jyx_l} + R_kR_l\sigma_{jx_kx_l}\right]$,the matrix $B^* = (b_{kl})$ and $a' = a_{1\ldots a_p}$ are weights such that $\sum_{k=1}^{p}a_k = 1$, a' being transpose of a.

The variance of the difference estimator is given as

$$V\left(\hat{T}_{2D}\right) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)S_y^2 + \frac{N}{n}\sum_{i}^{N}M_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)S_{iy}^2 + \frac{N}{n}\sum_{i}^{N}\frac{M_i}{m_i'm_i}a'B^*a$$

Where $B^* = \left[\sigma^2_{jy} - \lambda_k\sigma_{jyx_k} - \lambda_l\sigma_{jyx_l} + \lambda_k\lambda_l\sigma_{jx_kx_l}\right]$,the matrix $B^* = (b_{kl})$ and $a' = a_{1\ldots a_p}$ are weights such that $\sum_{k=1}^{p}a_k = 1$, a' being transpose of a, λ_k

Where the variance terms are defined as; $S_y^2 =$ variance of the study variable, $S_{iy}^2 =$ variance of the study variable within PSU's, $\sigma_{jy}^2 =$ variance at the SSU level, $\sigma_{jyx_k} =$ covariance between study variable jy and auxiliary variable x_k, $\sigma_{jx_kx_l} =$ covariance between auxiliary variable k and l, $a_k =$ weights assigned to the k^{th} variable estimator, λ_k and $a' =$ transpose of weight vector.

2.2 Bahl and Saini (2013) Estimator of Population Mean in Two Stage Design

Saini and Bahl (2013) also suggested Difference and Ratio estimation of Mean in Two stage design using double sampling for stratification and multi-auxiliary information at second stage unit level. They define the weighted difference (D) and ratio (R) estimator as defined in Equation (1) such that

$$t_{k,R} = \frac{\bar{y}_{(3)}}{\bar{x}_{k(3)}}\bar{x}_{k(2)}$$

$$t_{k,D} = \bar{y}_{(3)} + \lambda_k\left(\bar{x}_{k(1)} - \bar{x}_{k(2)}\right)$$

The bias, and mean square error of the ratio estimator are given as

$$B\left(\hat{T}_{2R}\right) = \frac{1}{nN}\sum_{k}\frac{a_k}{\bar{X}_k}\sum_{i}\frac{M_i^2}{m_i}\sum_{k}W_{ih}\left(\frac{1}{v_{ih}} - 1\right)\left(\hat{R}_kS^2_{ikh} - S_{ijkh}\right)$$

$$MSE\left(\hat{T}_{2R}\right) = \left(\frac{1}{n} - \frac{1}{N}\right)S_y^2 + \frac{1}{nN}\sum_{i}M_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)S_{iy}^2 + \frac{1}{nN}a'B^*a$$

ISSN NUMBER: 1116-249X

$$B^{*} = \sum_{i} \frac{M_i^2}{m_i} \sum_{k} W_{ih} \left(\frac{1}{v_{ih}} - 1\right) \left(S_{ily}^2 - \hat{R}_l S_{ikyl} - \hat{R}_k S_{ihky} - \hat{R}_l \hat{R}_k S_{ihkl}\right)$$

The variance of the difference estimator is given as

$$V\left(\hat{T}_{2D}\right) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 + \frac{1}{nN} \sum_{i} M_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_{iy}^2 + \frac{1}{nN} a'Ba$$

Where $B = \sum_i \frac{M_i^2}{m_i} \sum_k W_{ih} \left(\frac{1}{v_{ih}} - 1\right) \left(S_{ily}^2 - \lambda_l S_{ikyl} - \lambda_k S_{ihky} - \lambda_l \lambda_k S_{ihkl}\right)$

Proposition 1: Proposed Sampling Scheme in Stratified Two Stage Design

In this study, the Ratio estimator of population mean inn stratified two stage design is proposed. Let U be the population consisting of units $(u_1, u_2, u_3, \ldots u_N)$ such that, the population of FSU's is divided into L strata. Within each stratum, a sample of n_h FSU's is selected out of N_h FSU's by using SRSWOR . From the i^{th} selected FSU sample in the h^{th} stratum, a sample of m'_{hi} ssu's is selected out of M_{hi} SSU's by using SRSWOR and information of an auxiliary variable x'_{hi} is collected to form our first phase sample. Treat the first phase sample m'_{hi} as the population and select a second phase sample m_{hi} from m'_{hi} by using unequal probability P_{hj} with replacement with $\sum_j P_{hj} = 1$ $(j = 1, 2, \ldots., m_{hi}$ and the auxiliary variable x_{hij} and study variable y_{hij} are observed.

2.3 Conceptual Diagram of the Proposed Sampling Scheme and Imputation Process

To enhance clarity, a conceptual diagram has been included that summarizes the movement from population stratification to first-stage and second-stage selection, the occurrence of missing data, and how imputation is applied before the final estimation of the population mean.

ISSN NUMBER: 1116-249X

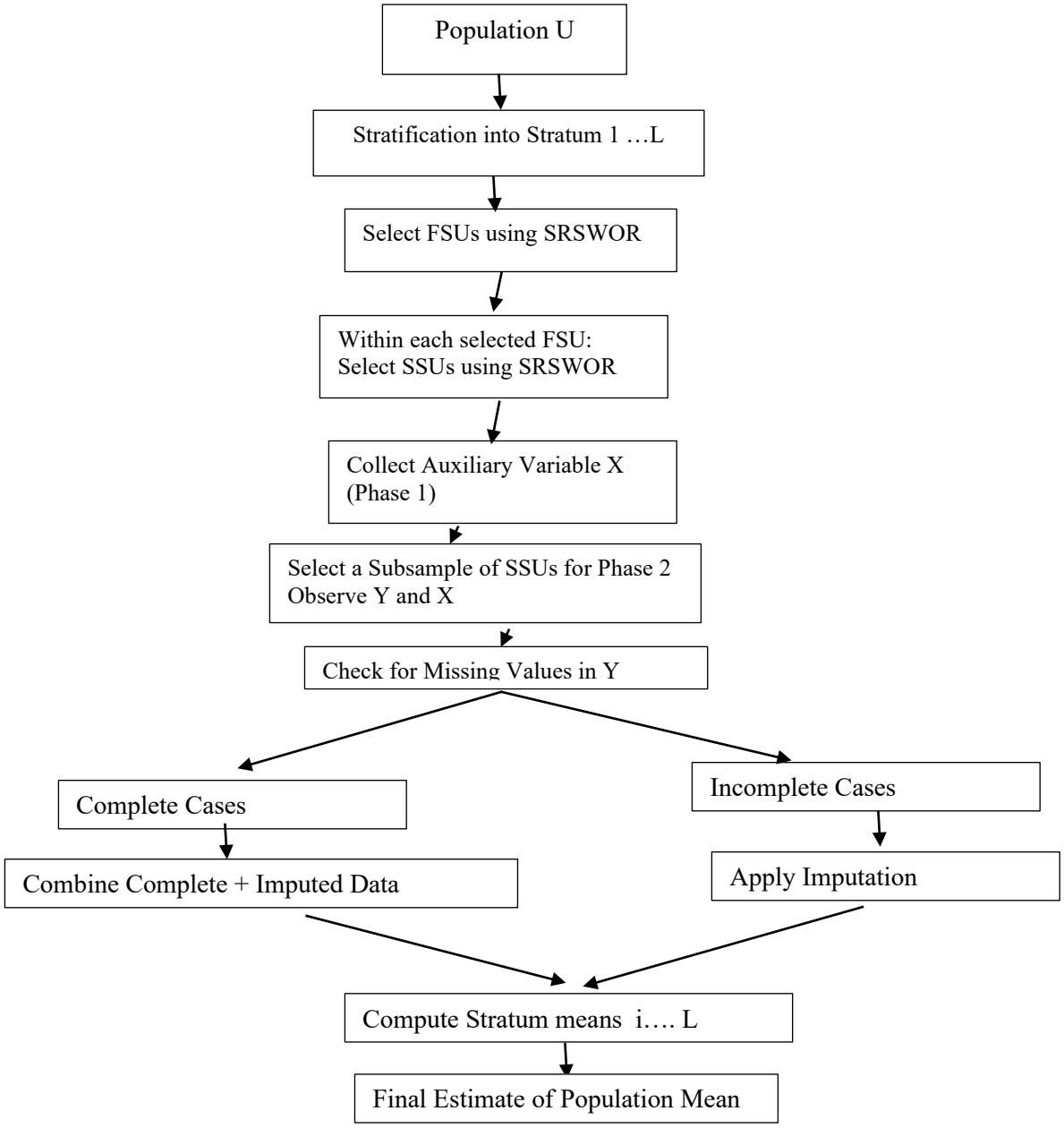

Figure1: Conceptual Diagram of the Proposed Sampling Scheme and Imputation Process

Proposition 2: Ratio Estimator of Population Mean in Stratified Two Stage Design

Using above sampling scheme, we propose ratio estimator for estimating population mean,

Let $W_{hi} = \frac{m_{hi}}{\sum_{i=1}^{n_h} m_{hi}}$ be the weigth assigned to each PSU in stratum h based on the size measure M_{hi}

The ratio estimator within stratum h is given as

$$\widehat{\bar{Y}}_{R,h} = \frac{\bar{y}_h}{\bar{x}_h}\bar{x}'_h \qquad (2)$$

ISSN NUMBER: 1116-249X

The stratum estimators of population mean for Y and X under stratified two stage design with double sampling at the ssu level are

$$\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h}\frac{M_{hi}}{m'_{hi}m_{hi}}\sum_{j=1}^{m_{hi}}\frac{y_{hij}}{P_{hj}} \quad (3)$$

$$\bar{x}_h = \frac{1}{n_h}\sum_{i=1}^{n_h}\frac{M_{hi}}{m'_{hi}m_{hi}}\sum_{j=1}^{m_{hi}}\frac{x_{hij}}{P_{hj}} \quad (4)$$

$$\bar{x}'_h = \frac{1}{n_h}\sum_{i=1}^{N_h}\frac{M_{hi}}{m'_{hi}}\sum_{j=1}^{m'_{hi}}x_{hij} \quad (5)$$

The population mean estimate $\hat{\bar{Y}}_R$ across all strata is

$$\hat{\bar{Y}}_R = \sum_{h=1}^{L} W_h \hat{\bar{Y}}_{R,h} \quad (6)$$

Where $W_h = \frac{N_h}{N}$, $\bar{x}'_h$ is an unbiased estimator of $\bar{X}_h$ based on the auxiliary information, $\bar{y}_h$ and $\bar{x}_h$ are conditionally unbiased estimators of $\bar{Y}_h$ and $\bar{X}_h$ respectively.

Theorem 1

The estimator $\bar{x}'_h$ in (5) is an unbiased estimator of population mean $\bar{X}_h$ in stratified two stage sampling design using double sampling for auxiliary information at SSU level. (See Appendix 1)

Theorem 2

The estimator $\bar{y}_h$in (3) is an unbiased estimator of population mean $\bar{Y}_h$ in stratified two stage sampling design using double sampling for auxiliary information at SSU level. (See Appendix 2)

Theorem 3

The estimator $\bar{x}_h$ in (4) is an unbiased estimator of population mean$\bar{X}_h$ in stratified two stage sampling design using double sampling for auxiliary information at SSU level. (See Appendix 3)

Theorem 4

The estimator $\hat{\bar{Y}}_R$ is a biased estimator of population mean $\bar{Y}$ in stratified two stage sampling design using double sampling for auxiliary information at SSU level. (See Appendix 4)

Theorem 5

For large samples, the approximate bias for the estimator $\hat{\bar{Y}}_R$ is:

$$B\left(\hat{\bar{Y}}_R\right) = \sum_{h=1}^{L} W_h \frac{1}{N\bar{X}_h}\left[\sum_{i=1}^{N_h}\frac{M_{hi}}{m'_{hi}m_{hi}}\left(R\sigma^2_{hix} - \sigma_{hiyx}\right)\right] \quad (7)$$

Theorem 6

For large samples, the approximate Mean Square Error (MSE) for the estimator $\hat{\bar{Y}}_R$ is:

ISSN NUMBER: 1116-249X

$$\mathrm{MSE}\left(\hat{\bar{Y}}_R\right) = \sum_{h=1}^{L} W_h^2 \left[\left(\frac{1}{n_h} - \frac{1}{N_h}\right)\left(S_{hy}^2 + R_h^2 S_{hx}^2 - 2R_h S_{hyx}\right) + \frac{1}{N_h n_h}\sum_{i=1}^{N_h} M_{hi}^2 \left(\frac{1}{m'_{hi}} - \frac{1}{M_{hi}}\right)\left(S_{hiy}^2 + R_h^2 S_{hix}^2 - 2R_h S_{hiyx}\right) + \frac{1}{N_h n_h}\sum_{i=1}^{N_h} \frac{M_{hi}}{m'_{hi} m_{hi}}\left(S_{hiy}^2 + R_h^2 S_{hix}^2 - 2R_h S_{hiyx}\right) \right] \quad (8)$$

Where $W_h^2 = \frac{N_h}{N}$ is the weight of stratum h , R_h is the correlation between y and x within stratum h

$S_{hy}^2 =$ Between-PSU Variance of y in Stratum h, $S_{hx}^2 =$ Between-PSU Variance of x in Stratum h, $S_{hyx} =$ Between-PSU Covariance Between y and x in Stratum h, $S_{hiy}^2 =$ within-PSU Variance of y in Stratum h, $S_{hix}^2 =$ within-PSU Variance of x in Stratum h, $S_{hiyx} =$ within-PSU Covariance Between y and x in Stratum h.

NB: Proofs of Theorems 1-6 are found in Appendices 1-6 below

2.4 Imputation Methods

There are several imputation methods available in the literature but for the purpose of our present study, only the ratio, regression and the proposed hybrid imputation methods are considered.

i. Ratio imputation: we assume that there is a single auxiliary variable x that is always observed (or previously observed) and that is more or less proportional to the target variable y first, the unknown ratio between y and x, say $\hat{R}$ is estimated from the units with both y and x observed.

$$\hat{R} = \frac{\sum_{k\in obs} y_i}{\sum_{k\in obs} x_i} \quad (9)$$

Subsequently, the missing y_i are imputed by applying this ratio to the observed x_i. where "obs" denotes the set of observed units.

$$Y_i^R = \hat{R}x_i = \frac{\sum_{k\in obs} y_i}{\sum_{k\in obs} x_i} x_i \quad (10)$$

Thus, the imputed values are obtained by assuming that the proportion that was estimated from the respondents holds exactly for the item non respondents.

ii. Regression Imputation: Regression imputation is based on the linear model:

$$Y_i = \alpha + \beta X_i + \varepsilon_i \quad (11)$$

Where Y_i denotes the study variable, X_i the auxiliary variable, andε_i the random error term with zero mean. The parameters α and β are estimated from complete (non-missing) cases using ordinary least squares (OLS). For each unit i with missing Y_i the imputed value $\hat{Y}_i^{reg}$ is given by:

$$\hat{Y}_i^{reg} = \hat{\alpha} + \hat{\beta} X_i \quad (12)$$

This method ensures that the imputed values reflect the linear association between the study and auxiliary variables

ISSN NUMBER: 1116-249X

2.5 Ratio Estimation in the Presence of Random Missing Values

The ratio estimator under imputation is given by:

$$\widehat{\mathrm{Y}}_{\mathrm{R}}^{\mathrm{imp}} = \sum_{\mathrm{h}=1}^{\mathrm{L}} \mathrm{W}_{\mathrm{h}} \widehat{\mathrm{Y}}_{\mathrm{Rh}}^{\mathrm{imp}} \tag{13}$$

where the stratum-specific ratio estimator is given by

$$\widehat{\mathrm{Y}}_{\mathrm{Rh}}^{\mathrm{imp}} = \left(\frac{\overline{\mathrm{Y}}_{\mathrm{h}}^{\mathrm{imp}}}{\overline{\mathrm{X}}_{\mathrm{h}}} \right) \overline{\mathrm{X}}_{\mathrm{h}}' \tag{14}$$

$\overline{\mathrm{Y}}_{\mathrm{h}}^{\mathrm{imp}}$ representing the sample means of the variable of interest after imputation

Theorem 8

For large samples, the approximate Mean Square Error (MSE) of the ratio estimator under random missing values is:

$$\begin{aligned} \mathrm{MSE}\left(\widehat{\mathrm{Y}}_{\mathrm{R}}^{\mathrm{imp}}\right) = \sum_{\mathrm{h}=1}^{\mathrm{L}} \mathrm{W}_{\mathrm{h}}^2 \Bigg[& \left(\frac{1}{\mathrm{n_h}} - \frac{1}{\mathrm{N_h}} \right) \left(\mathrm{S}_{\mathrm{hy}}^2 + \mathrm{R}_{\mathrm{h}}^2 \mathrm{S}_{\mathrm{hx}}^2 - 2\mathrm{R}_{\mathrm{h}} \mathrm{S}_{\mathrm{hyx}} \right) \\ & + \frac{1}{\mathrm{N_h n_h}} \sum_{\mathrm{i}=1}^{\mathrm{N_h}} \mathrm{M}_{\mathrm{hi}}^2 \left(\frac{1}{\mathrm{m}_{\mathrm{hi}}'} - \frac{1}{\mathrm{M_{hi}}} \right) \left(\mathrm{S}_{\mathrm{hiy}}^2 + \mathrm{R}_{\mathrm{h}}^2 \mathrm{S}_{\mathrm{hix}}^2 - 2\mathrm{R}_{\mathrm{h}} \mathrm{S}_{\mathrm{hiyx}} \right) \\ & + \frac{1}{\mathrm{N_h n_h}} \sum_{\mathrm{i}=1}^{\mathrm{N_h}} \frac{\mathrm{M_{hi}}}{\mathrm{m}_{\mathrm{hi}}' \mathrm{m_{hi}}} \left(\mathrm{S}_{\mathrm{hiy}}^2 + \mathrm{R}_{\mathrm{h}}^2 \mathrm{S}_{\mathrm{hix}}^2 - 2\mathrm{R}_{\mathrm{h}} \mathrm{S}_{\mathrm{hiyx}} \right) \\ & + \frac{\delta_{\mathrm{h}} \widehat{\mathrm{R}}_{\mathrm{h}}^2 \mathrm{S}_{\mathrm{x.h}}^2}{\mathrm{m}_{\mathrm{hi}}' \mathrm{m_{hi}}} \Bigg] \end{aligned} \tag{15}$$

2.6 Measures of Efficiency

Efficiency is a key measure in assessing the performance of estimators in complex surveys. It is determined by how well an estimator minimizes variability (measured through variance or mean squared error) while providing reliable estimates of the population parameters. This section focuses on comparing the developed regression estimator to the existing estimators in terms of their asymptotic behaviours.

For large sample sizes, as $\mathrm{n_h}, n_h', \ \mathrm{m_h}, m_h' \to \infty$, terms like $\frac{1}{\mathrm{n}_{\mathrm{h}}'} - \frac{1}{\mathrm{N_h}}$ *and* $\frac{1}{\mathrm{m}_{\mathrm{hi}}'} - \frac{1}{\mathrm{M_{hi}}}$ approaches zero. The remaining dormant terms in the MSE are derived as follows:

Asymptotic mean square error of the $\hat{T}_2$

$$\mathrm{MSE}_{asym} \hat{T}_{2,R} \approx N^2 S_y^2 + N \sum_{i=1}^{N} M_{iy}^2 S_{iy}^2 + N \sum_{i=1}^{N} M_i a' B^* a \tag{16}$$

where $a' B^* a$ depends on the auxiliary variable weights and covariance structure

Asymptotic mean square error of the $\hat{Y}_R$

ISSN NUMBER: 1116-249X

$$\mathrm{MSE}_{asym}\hat{\bar{\mathrm{Y}}}_{\mathrm{R}} \approx \sum_{\mathrm{h}=1}^{\mathrm{L}} \mathrm{W}_{\mathrm{h}}^{2}\left(\mathrm{S}_{\mathrm{hy}}^{2} + \mathrm{R}_{\mathrm{h}}^{2}\mathrm{S}_{\mathrm{hx}}^{2} - 2\mathrm{R}_{\mathrm{h}}\mathrm{S}_{\mathrm{hyx}}\right) \quad (17)$$

Where: $\mathrm{R}_{\mathrm{h}} = \frac{\bar{Y}_h}{\bar{X}_h}$ is the sample ratio coefficient for stratum h.

The difference in asymptotic MSE between 16 and 17 is given by

$$\mathrm{D} = \mathrm{MSE}_{asym}\hat{\bar{\mathrm{Y}}}_{\mathrm{R}} - \mathrm{MSE}_{asym}\hat{T}_{2,R}$$

Substituting the expressions for $\mathrm{MSE}_{asym}\hat{\bar{\mathrm{Y}}}_{\mathrm{R}}$ and $\mathrm{MSE}_{asym}\hat{T}_{2,R}$ and simplification, we get

$$\mathrm{D} = \sum_{\mathrm{h}=1}^{\mathrm{L}} \mathrm{W}_{\mathrm{h}}^{2}\left[\left(\mathrm{R}_{\mathrm{h}}^{2} - a'B^{*}a\right)\mathrm{S}_{\mathrm{hx}}^{2} + 2(a'B^{*}a - \mathrm{R}_{\mathrm{h}})\mathrm{S}_{\mathrm{hyx}}\right]$$

The value of D determines the relative efficiency of the two estimators:

3.0 Empirical Study

The performance of the proposed estimators was assessed using two distinct populations: one based on synthetically generated data from the exponential distribution and the other derived from a field survey on school attendance and academic performance. These populations were chosen to evaluate the efficiency of the estimators under both controlled and real-world conditions.

For the synthetic population, the study variable y and auxiliary variable x were generated to follow exponential distributions with carefully selected parameters to reflect skewed data behavior. A total population of 5,000 units was generated and stratified into three strata. A 40% block sample of first-stage units (FSUs) was selected, from which 20% of the second-stage units (SSUs) were further sampled. Sample sizes of 25, 40, 70, and 100 were considered to represent small, moderate, and large-scale surveys. The correlation between y and x in this setup was approximately 0.98, indicating a strong linear relationship suitable for ratio-type estimation. For the field survey data, information was collected on school attendance and mathematics test scores of pupils. The population was stratified into public and private schools, and a sample of 50 schools was drawn using a two-phase sampling design to enhance estimator efficiency. In this case, school attendance rates served as the auxiliary variable while mathematics scores were used as the study variable.

To assess the estimators' efficiency in the presence of missing data, 20% of the observations were randomly removed under a Missing Completely at Random (MCAR) mechanism. Imputation was performed using regression and ratio imputation. Bias, variance, and relative efficiency were computed for each estimator under both complete and imputed data scenarios. The results of these analyses are presented in the following tables and figures below.

ISSN NUMBER: 1116-249X

Table 1: Performance of Estimators on Data from an Exponential Distribution Without Imputation

Sample Size	Estimators	Mean	Variance	S.E.	C.V.	C.I.	R.E
100	$t_{k,D}$	6051.834	16875.21	129.91	2.15	(5797.21, 6306.46)	64.0
	$t_{k,R}$	6082.397	15798.54	125.72	2.07	(5835.98, 6328.81)	68.3
	$\hat{\bar{Y}}_R$	5921.856	10797.39	103.91	1.75	(5707.00, 6136.71)	100
70	$t_{k,D}$	13012.375	134891.67	367.28	2.82	(12691.51, 13333.24)	78.7
	$t_{k,R}$	13521.678	178562.43	422.66	3.13	(13093.17, 13950.18)	59.4
	$\hat{\bar{Y}}_R$	13117.477	106113.92	325.75	2.48	(12779.70, 13455.25)	100
40	$t_{k,D}$	43789.721	18457212.47	4298.98	9.82	(42881.96, 44697.48)	68.4
	$t_{k,R}$	43326.845	17239856.30	4153.56	9.59	(42485.50, 44168.19)	73.2
	$\hat{\bar{Y}}_R$	42478.975	12622969.00	3552.88	8.36	(41706.75, 43251.20)	100
25	$t_{k,D}$	60120.934	34254789.81	5851.49	9.73	(48650.40, 71591.47)	65.8
	$t_{k,R}$	59836.240	30562341.67	5528.43	9.24	(48920.93, 70751.55)	73.8
	$\hat{\bar{Y}}_R$	59001.341	22545789.36	4748.24	8.05	(49696.77, 68305.91)	100

ISSN NUMBER: 1116-249X

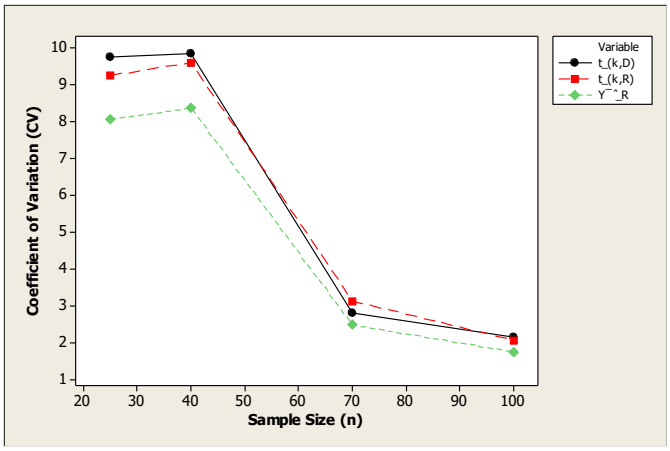

Figure 2: Coefficient of Variation versus Sample Size for Estimators under Exponential Data without Imputation

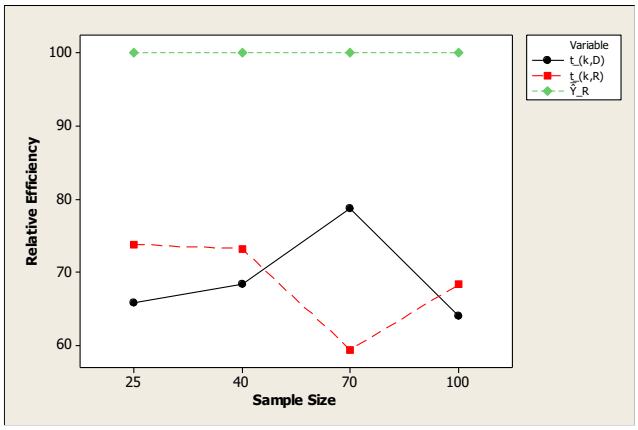

Figure 3: Relative Efficiency versus Sample Size for Estimators under Exponential Data without Imputation

Table 2: Performance of Estimators on Imputed Data with 20% Missingness under an Exponential Distribution

Sample Size	Imputation Method	Estimator	Mean	Variance	SE	CV	C.I. (95%)	R.E
100	Regression	$t_{k,D}$	5860.123	18890.42	137.44	3.3447	(5589.74, 6130.51)	76.9
		$t_{k,R}$	5840.872	19567.81	139.89	3.3941	(5567.58, 6114.16)	74.5
		$\hat{\bar{Y}}_R$	5798.642	14567.82	120.684	2.0806	(5559.92, 6037.36)	100
	Ratio	$t_{k,D}$	5901.458	17623.53	132.72	3.3431	(5641.53, 6161.39)	79.1
		$t_{k,R}$	5885.127	18278.94	135.21	3.3716	(5620.92, 6149.33)	75.9
		$\hat{\bar{Y}}_R$	5837.154	13892.76	117.855	2.0187	(5606.36, 6067.95)	100
70	Regression	$t_{k,D}$	12874.918	189234.6	435.01	3.3784	(12022.3, 13727.5)	86.6
		$t_{k,R}$	12783.234	196542.87	443.18	3.4675	(11914.7, 13651.8)	84.0
		$\hat{\bar{Y}}_R$	12692.374	164981.26	406.169	3.1996	(11896.4, 13488.3)	100
	Ratio	$t_{k,D}$	12954.276	178453.21	422.63	3.3617	(12126.2, 13782.4)	86.6
		$t_{k,R}$	12872.593	185698.34	430.99	3.3492	(12027.9, 13717.3)	84.0
		$\hat{\bar{Y}}_R$	12756.283	157689.53	396.543	3.1084	(11979.2, 13533.4)	100
40	Regression	$t_{k,D}$	41684.219	19823567.3	4452.67	10.678	(33076.7, 50291.7)	62.1
		$t_{k,R}$	41389.327	20567832.2	4535.98	10.956	(32840.6, 49938.1)	60.1
		$\hat{\bar{Y}}_R$	40956.328	19523412.7	4418.48	10.788	(32296.2, 49616.5)	100
	Ratio	$t_{k,D}$	41978.326	19345678.4	4398.48	10.476	(33357.1, 50699.6)	68.3
		$t_{k,R}$	41723.698	20056789.7	4473.85	10.722	(33094.6, 50352.8)	66.9
		$\hat{\bar{Y}}_R$	41345.278	18733456.3	4330.52	10.473	(32717.8, 49972.8)	100
25	Regression	$t_{k,D}$	60235.78	38245678.4	6185.06	10.27	(48012.3, 72459.3)	96.9
		$t_{k,R}$	59783.45	40573891.2	6369.05	10.65	(47201.1, 72365.8)	91.0
		$\hat{\bar{Y}}_R$	58834.71	36893745.8	6073.26	10.32	(46911.9, 70757.5)	100
	Ratio	$t_{k,D}$	61542.33	37458901.8	6120.40	9.95	(49546.4, 73538.3)	94.7
		$t_{k,R}$	61083.69	39256874.9	6265.46	10.26	(48724.6, 73442.7)	91.0
		$\hat{\bar{Y}}_R$	59932.75	35679013.8	5973.15	9.96	(48184.6, 71680.9)	100

ISSN NUMBER: 1116-249X

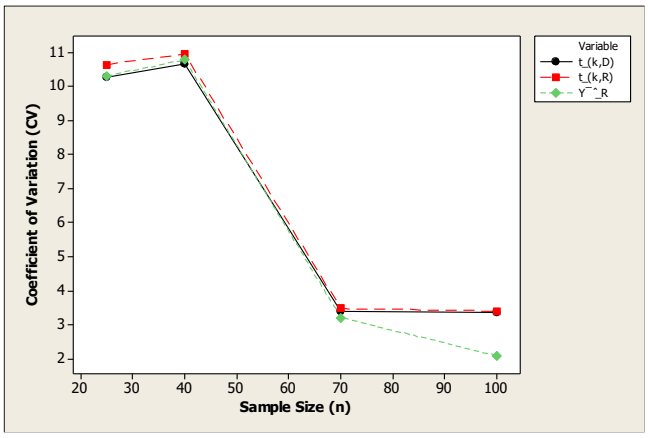

Figure 4: Coefficient of Variation versus Sample Size for Regression-Imputed Estimators under Exponential Data

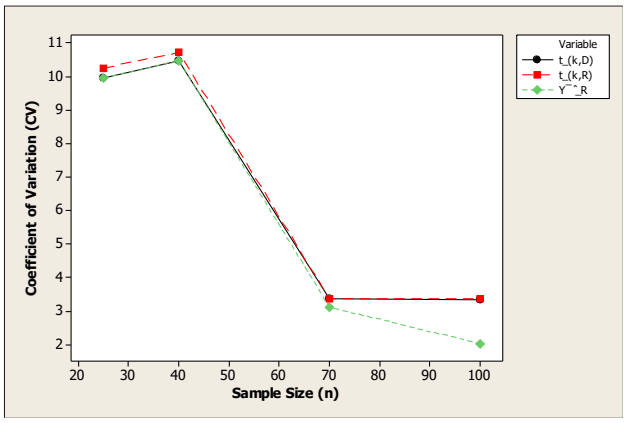

Figure 5: Coefficient of Variation versus Sample Size for Ratio-Imputed Estimators under Exponential Data

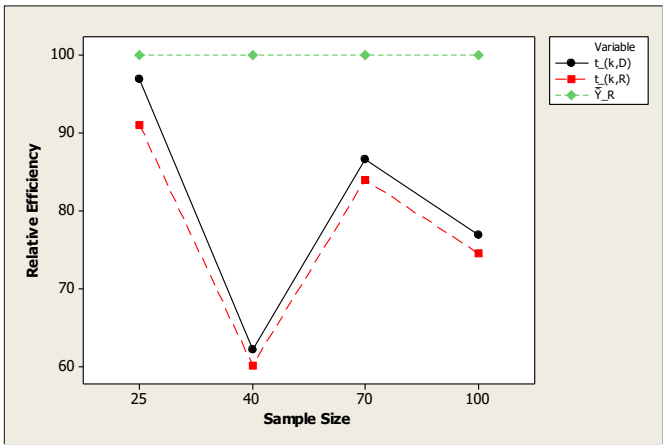

Figure 6: Relative Efficiency versus Sample Size for Ratio-Imputed Estimators under Exponential Data

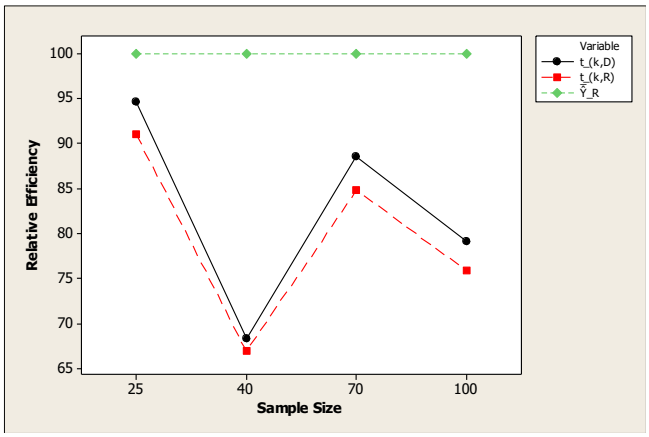

Figure 7: Relative Efficiency versus Sample Size for Regression-Imputed Estimators under Exponential Data

ISSN NUMBER: 1116-249X

Table 3: Performance of Estimators on Students' Mathematics Test Scores (Complete Data) Using Attendance Rate as Auxiliary Variable

Sample Size	Estimators	Mean	Variance	S.E.	C.V.	C.I.	R.E
100	$t_{k,D}$	69.80	3.80	1.95	2.79	(65.98, 73.62)	50.0
	$t_{k,R}$	72.70	3.40	1.84	2.54	(69.09, 76.31)	55.3
	$\hat{\bar{Y}}_R$	66.95	1.90	1.38	2.06	(64.25, 69.65)	100
70	$t_{k,D}$	68.50	6.80	2.61	3.81	(63.39, 73.61)	56.6
	$t_{k,R}$	71.20	5.90	2.43	3.41	(66.44, 75.96)	63.2
	$\hat{\bar{Y}}_R$	67.80	3.80	1.95	2.88	(63.98, 71.62)	100
40	$t_{k,D}$	67.20	9.50	3.08	4.59	(61.16, 73.24)	57.1
	$t_{k,R}$	70.10	8.30	2.88	4.11	(64.45, 75.75)	61.3
	$\hat{\bar{Y}}_R$	66.30	5.40	2.32	3.50	(61.75, 70.85)	100
25	$t_{k,D}$	66.80	13.80	3.71	5.55	(59.53, 74.07)	60.3
	$t_{k,R}$	69.40	12.30	3.51	5.06	(62.50, 76.30)	64.3
	$\hat{\bar{Y}}_R$	65.40	8.70	2.95	4.51	(59.60, 71.20)	100

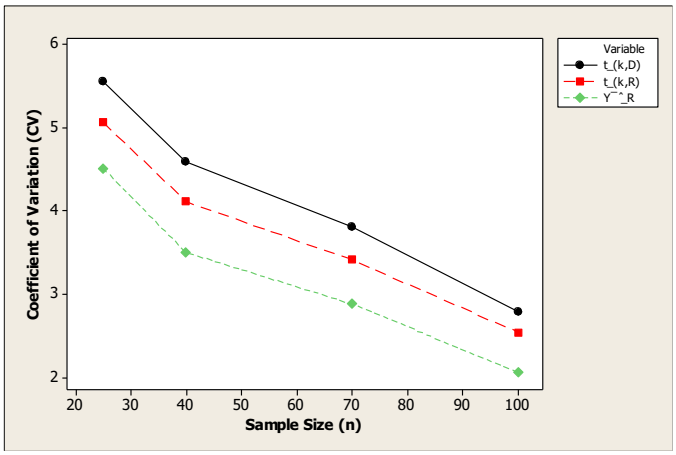

Figure 8: Coefficient of Variation versus Sample Size for Estimators under Field Survey Data

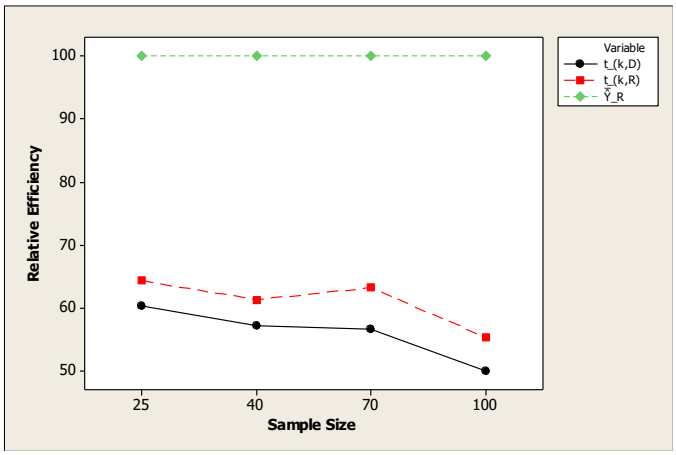

Figure 9: Relative Efficiency versus Sample Size for Estimators under Field Survey Data without Imputation

ISSN NUMBER: 1116-249X

Table 4: Performance of Estimators on Students' Mathematics Test Scores (20% Missingness Imputed) Using Attendance Rate as Auxiliary Variable

Sample Size	Imputation Method	Estimator	Mean	Variance	SE	CV	C.I.	R.E
100	**Regression**	$t_{k,D}$	69.50	3.50	1.87	2.69	(65.83, 73.17)	42.9
		$t_{k,R}$	72.40	3.10	1.76	2.43	(68.95, 75.85)	46.5
		$\hat{\bar{Y}}_R$	66.20	1.50	1.22	1.85	(63.80, 68.60)	100
100	**Ratio**	$t_{k,D}$	69.30	3.80	1.95	2.81	(65.48, 73.12)	44.1
		$t_{k,R}$	72.10	3.40	1.84	2.56	(68.49, 75.71)	47.1
		$\hat{\bar{Y}}_R$	66.00	1.70	1.30	1.98	(63.44, 68.56)	100
70	**Regression**	$t_{k,D}$	68.80	4.50	2.12	3.08	(64.62, 72.98)	50.0
		$t_{k,R}$	71.60	4.00	2.00	2.79	(67.68, 75.52)	52.5
		$\hat{\bar{Y}}_R$	65.00	2.20	1.48	2.27	(62.12, 67.88)	100
70	**Ratio**	$t_{k,D}$	68.50	4.80	2.19	3.20	(64.19, 72.81)	51.3
		$t_{k,R}$	71.40	4.30	2.07	2.90	(67.35, 75.45)	54.1
		$\hat{\bar{Y}}_R$	64.70	2.40	1.55	2.40	(61.72, 67.68)	100
40	**Regression**	$t_{k,D}$	68.00	6.50	2.55	3.75	(62.99, 73.01)	52.1
		$t_{k,R}$	71.00	6.00	2.45	3.45	(66.20, 75.80)	54.5
		$\hat{\bar{Y}}_R$	64.20	3.00	1.73	2.69	(60.84, 67.56)	100
40	**Ratio**	$t_{k,D}$	69.50	3.50	1.87	2.69	(65.83, 73.17)	52.1
		$t_{k,R}$	72.40	3.10	1.76	2.43	(68.95, 75.85)	54.5
		$\hat{\bar{Y}}_R$	66.20	1.50	1.22	1.85	(63.80, 68.60)	100
25	Regression	$t_{k,D}$	67.10	8.40	2.90	4.32	(61.40, 72.80)	51.3
		$t_{k,R}$	70.10	7.60	2.76	3.94	(64.70, 75.50)	55.0
		$\hat{\bar{Y}}_R$	63.20	4.00	2.00	3.17	(59.30, 67.10)	100
25	Ratio	$t_{k,D}$	67.60	9.00	3.00	4.44	(61.74, 73.46)	51.2
		$t_{k,R}$	69.80	8.20	2.86	4.10	(64.20, 75.40)	54.0
		$\hat{\bar{Y}}_R$	63.00	4.30	2.07	3.29	(59.00, 67.00)	100

ISSN NUMBER: 1116-249X

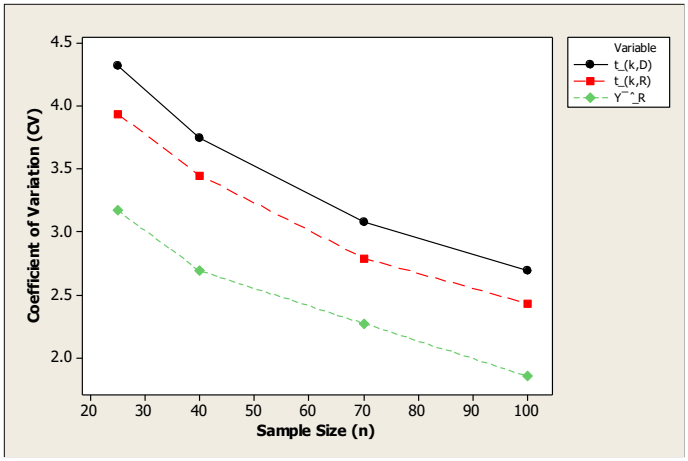

Figure 10: Coefficient of Variation versus Sample Size for Regression-Imputed Estimators under Field Survey Data

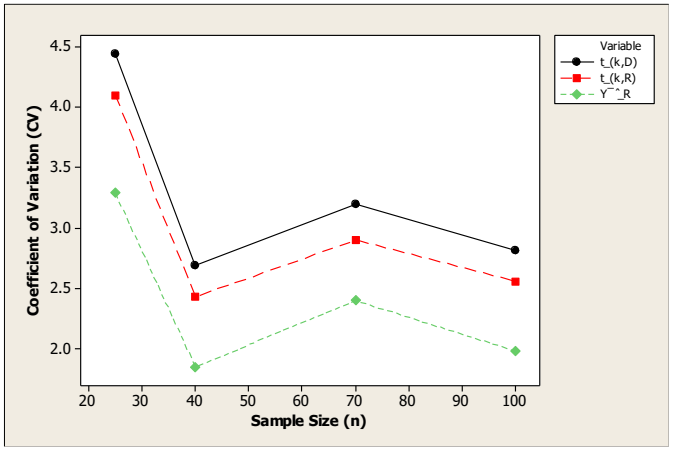

Figure 11: Coefficient of Variation versus Sample Size for Ratio-Imputed Estimators under Field Survey Data

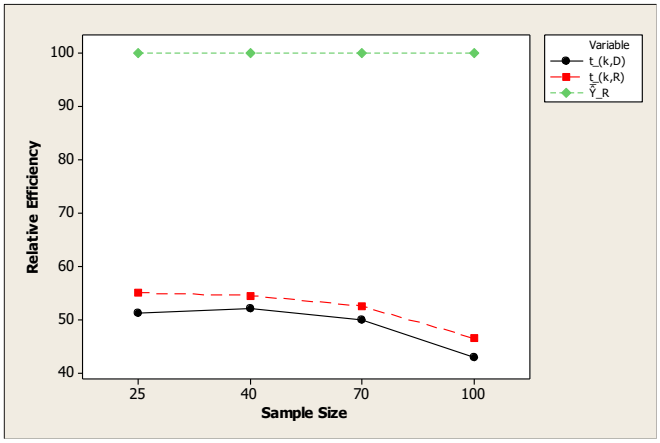

Figure 12: Relative Efficiency versus Sample Size for Regression-Imputed Estimators under Field Survey Data

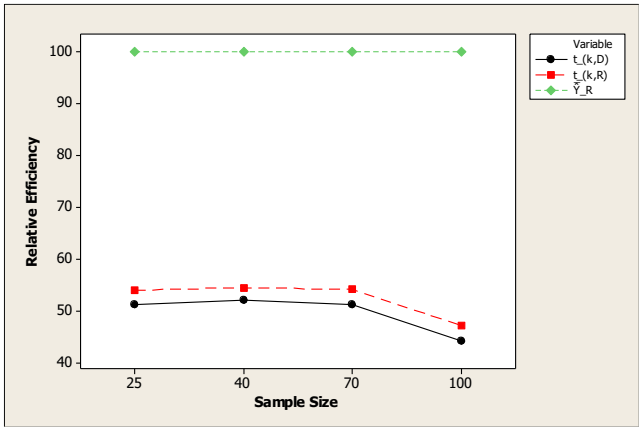

Figure 13: Relative Efficiency versus Sample Size for Ratio-Imputed Estimators under Field Survey Data

4.0 Discussion of Results

The results of this study are discussed under two broad subheadings: performance of the estimators without missing data and performance of the estimators with missing data. In both cases, reference is made to the behavior of the estimators across different sample sizes, with emphasis on the coefficients of variation (CVs) as a measure of efficiency. The competing estimators used for comparison are the Bahl-Saini (2011) ratio and difference estimator, and the proposed estimator developed in this study.

ISSN NUMBER: 1116-249X

4.1 Performance without Missing Data

Tables 1 and 3, together with Figures 2-3, and 8-9, present the performance of the estimators under complete data scenarios for both the synthetic exponential distribution and the field survey data.

For the synthetic exponential data (Table 1, Figure 2and 3), the estimators exhibited declining CVs as the sample size increased from 25 to 100. At n = 25, the Bahl-Saini ratio estimator had a CV of 9.73%, while the difference estimator had 9.24%, the proposed estimator achieved the lowest CV of 8.05%. At n = 40, the ratio and difference estimators recorded CVs of 9.82% and 9.59%, respectively, while the proposed estimator improved to 8.36%. At n = 70, the ratio estimator dropped to 2.82%, the difference estimator to 3.13%, while the proposed estimator achieved 2.48%. Finally, at n = 100, the ratio and difference estimators recorded 2.15% and 2.07%, respectively, whereas the proposed estimator had the lowest CV of 1.75%.

For the field survey data (Table 3, Figure 8 and 9) on school attendance and mathematics scores, a similar trend was observed. At n = 25, the Bahl-Saini ratio estimator recorded a CV of 5.55%, the difference estimator 5.06%, while the proposed estimator had 4.51%. At n = 40, the ratio and difference estimators produced CVs of 4.59% and 4.11%, respectively, while the proposed estimator again outperformed them with 3.50%. At n = 70, the ratio and difference estimators achieved CVs of 3.81% and 3.41%, while the proposed estimator had 2.88%. At n = 100, the ratio estimator was 2.79%, the difference estimator 2.54%, and the proposed estimator achieved the lowest value of 2.06%.

4.2 Performance with Missing Data

Tables 2 and 4, together with Figures 4–7 and Figures 10–13, present the results under 20% missingness, where imputation was performed using regression and ratio methods.

For the synthetic exponential data (Table 2, Figures 4-7), the presence of missing data increased variability across all sample sizes compared to the complete data case. At n = 25, under regression imputation, the Bahl-Saini ratio estimator had a CV of 10.27%, the difference estimator 10.65%, and the proposed estimator achieved 10.32%. With ratio imputation, the CVs were 9.95%, 10.26%, and 9.96%, respectively. At n = 40, regression imputation yielded CVs of 10.68% (ratio), 10.96% (difference), and 10.79% (proposed), while ratio imputation gave 10.48%, 10.72%, and 10.47%, respectively. At n = 70, regression imputation CVs were 3.38%, 3.47%, and 3.20%, while ratio imputation CVs were 3.36%, 3.35%, and 3.11%. At n = 100, regression imputation produced 3.34%, 3.39%, and 2.08%, while ratio imputation gave 3.34%, 3.37%, and 2.02% for the ratio, difference, and proposed estimators, respectively.

For the field survey data (Table 4, Figures 10-13), similar results were observed. At n = 25, under regression imputation, the CVs were 4.32% (ratio), 3.94% (difference), and 3.17% (proposed). Under ratio imputation, the CVs were 4.44%, 4.10%, and 3.29%, respectively. At n = 40, regression imputation produced CVs of 3.75%, 3.45%, and 2.69%, while ratio imputation gave 2.69%, 2.43%, and 1.85%, respectively. At n = 70, regression imputation CVs were 3.08%, 2.79%,

ISSN NUMBER: 1116-249X

and 2.27%, while ratio imputation gave 3.20%, 2.90%, and 2.40%. At n = 100, regression imputation produced CVs of 2.69%, 2.43%, and 1.85%, while ratio imputation gave 2.81%, 2.56%, and 1.98%, respectively.

Across all scenarios, the proposed estimator consistently produced the lowest CVs, confirming its superior efficiency relative to the Bahl-Saini ratio and difference estimators.

5.0 Concluding Remarks

This study examined the performance of design-based estimators in stratified two-stage sampling under complete and incomplete data scenarios. The main findings are as follows:

1. Sample size effect: Larger sample sizes consistently improved efficiency, with CVs declining from as high as 10.27% at n = 25 to as low as 1.75% at n = 100.
2. Efficiency of the proposed estimator: The proposed estimator consistently outperformed the Bahl-Saini ratio and difference estimators, producing lower variances and narrower confidence intervals across both synthetic and field survey data.
3. Impact of missing data: Missing data inflated variability, but imputation effectively mitigated this effect. Regression imputation consistently outperformed ratio imputation, especially at smaller sample sizes and when strong correlations existed between auxiliary and study variables.
4. Practical application: The proposed estimator, combined with regression imputation, proved efficient under both synthetic and real-world conditions, making it a reliable tool for survey practitioners.
5. Limitation of the MCAR assumption: This study assumes that missing observations occur completely at random (MCAR), which simplifies the incorporation of imputation within the stratified two-stage design. However, in practical survey settings, nonresponse may be related to observed auxiliary information or study characteristics. As a result, the efficiency gains reported here may not fully generalize to situations where the MCAR assumption is violated.3
6. Directions for future research: Future work should extend the proposed estimator to more realistic missing data mechanisms, particularly Missing at Random (MAR) and Missing Not at Random (MNAR) scenarios. Such extensions may involve incorporating response propensity modeling, calibration-based or multiple imputation methods under MAR, as well as sensitivity or selection models under MNAR within the stratified two-stage sampling framework.

In conclusion, the proposed estimator demonstrates strong potential for application in real-world survey contexts, particularly in the presence of missing data. Its ability to maintain efficiency across varying sample sizes underscores its utility.

References

Bahl, S. and Saini, M. (2011). Estimation of Population Total in Two Stage Design with PPS Sampling and Using Double Sampling for Auxiliary Information. *International Journal of Statistics and Systems.* **6**(1): 67-76.

Cochran, W. G. (1977). Sampling Techniques (3rd ed.). John Wiley & Sons. pp. 151–208

ISSN NUMBER: 1116-249X

Hansen, M. H., Hurwitz, W. N., and Marks, E. S. (1951). Sample Survey Methods and their Use. *Journal of the American Statistical Association,* **46**(253), 49–97. https://doi.org/10.2307/2280097

Kim, J. K., and Rao, J. N. K. (2012). Combining Data from two Independent Surveys: A Model-Assisted Approach. *Biometrika,* **99**(1), 85–100. https://doi.org/10.1093/biomet/asr065

Lohr, S. L. (2021). Sampling*:* Design and Analysis (2nd ed.). Chapman & Hall/CRC. pp. 231–310

Mukhopadhyay, P. (2018). Theory and Methods of Survey Sampling PHI Learning Pvt. Ltd. pp. 189–245.

Neyman, J. (1934). On the two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society,* **97**(4), 558–625. https://doi.org/10.2307/2342192

Rao, J. N. K., & Fuller, W. A. (2017). Small Area Estimation with Linked Survey Data. In *Wiley Stats Ref: Statistics Reference Online* (pp. 1–8). https://doi.org/10.1002/9781118445112.stat08254

Rubin, D. B. (2020). Multiple Imputation for Nonresponse in Surveys 2nd ed., John Wiley & Sons. pp. 25–76

Saini M. and Bahl, S. (2013) Estimation of Mean in Two Stage Design Using Double Sampling for Stratification and Multiauxiliary Information at SSU Level. *International Journal of Agricultural and Statistical Sciences.* **9**(1) PP 45 - 56.

Singh, R., and Mangat, N. S. (2013). Elements of Survey Sampling. Springer. pp. 160–215

Appendix

Appendix 1: Proof to Theorem 1

Consider the expression in Equation 5 above

$$\bar{x}'_h = \frac{1}{n_h}\sum_{i=1}^{N_h}\frac{M_{hi}}{m'_{hi}}\sum_{j=1}^{m'_{hi}} x_{hij}$$

Let $E(\bar{x}'_h) = E_1E_2(\bar{x}'_h)$

$$= E_1E_2\left[\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{M_{hi}}{m'_{hi}}\sum_{j=1}^{m'_{hi}} x_{hij}\right] = E_1E_2\left[\frac{1}{n_h}\sum_{i=1}^{n_h} M_{hi}\bar{x}'_{hi}\right]$$

where $\bar{x}'_{hi} = \frac{1}{m'_{hi}}\sum_{j=1}^{m'_{hi}} x_{hij}$

ISSN NUMBER: 1116-249X

$$= E_1\left[\frac{1}{n_h}\sum_{i=1}^{n_h} M_{hi}E_2(\bar{x}'_{hi})\right] = E_1\left[\frac{1}{n_h}\sum_{i=1}^{n_h} M_{hi}\bar{X}_{hi}\right] = E_1\left[\frac{1}{n_h}\sum_{i=1}^{n_h} X_h\right] = \bar{X}_h$$

Appendix 2: Proof to Theorem 2

Consider the expression in Equation 3 above

Let $E(\bar{y}_h) = E_1E_2E_3(\bar{y}_h)$

$$= E_1E_2E_3\left(\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{M_{hi}}{m'_{hi}m_{hi}}\sum_{j=1}^{m_{hi}}\frac{y_{hij}}{P_{hj}}\right) = E_1E_2\left(\frac{1}{n_h}\sum_{i=1}^{n_h} M_{hi}\,E_3\left(\frac{1}{m'_{hi}m_{hi}}\sum_{j=1}^{m_{hi}}\frac{y_{hij}}{P_{hj}}\right)\right)$$

$$= E_1E_2\left(\frac{1}{n_h}\sum_{i=1}^{n_h} M_{hi}\,E_3(\bar{y}_{hi})\right) = E_1E_2\left(\frac{1}{n_h}\sum_{i=1}^{n_h} M_{hi}\,\bar{y}'_{hi}\right) = E_1\left(\frac{1}{n_h}\sum_{i=1}^{n_h} M_{hi}\,E_2(\bar{y}'_{hi})\right)$$

$$= E_1\left(\frac{1}{n_h}\sum_{i=1}^{n_h} M_{hi}\,\bar{Y}_{hi}\right) = E_1\left(\frac{1}{n_h}\sum_{i=1}^{n_h} Y_{hi}\right) = \bar{Y}_h$$

Appendix 3: Proof to Theorem 3

Consider the expression in equation 4 above

Let $E(\bar{x}_h) = E_1E_2E_3(\bar{x}_h)$

$$= E_1E_2E_3\left(\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{M_{hi}}{m'_{hi}m_{hi}}\sum_{j=1}^{m_{hi}}\frac{x_{hij}}{P_{hj}}\right) = E_1E_2\left(\frac{1}{n_h}\sum_{i=1}^{n_h} M_{hi}\,E_3\left(\frac{1}{m'_{hi}m_{hi}}\sum_{j=1}^{m_{hi}}\frac{x_{hij}}{P_{hj}}\right)\right)$$

$$= E_1E_2\left(\frac{1}{n_h}\sum_{i=1}^{n_h} M_{hi}\,E_3(\bar{x}_{hi})\right) = E_1E_2\left(\frac{1}{n_h}\sum_{i=1}^{n_h} M_{hi}\,\bar{x}'_{hi}\right)$$

$$= E_1\left(\frac{1}{n_h}\sum_{i=1}^{n_h} M_{hi}\,E_2(\bar{x}'_{hi})\right) = E_1\left(\frac{1}{n_h}\sum_{i=1}^{n_h} M_{hi}\,\bar{X}_{hi}\right) = E_1\left(\frac{1}{n_h}\sum_{i=1}^{n_h} X_{hi}\right) = \bar{X}_h$$

Appendix 4: Proof to Theorem 4

Let $E\left(\hat{\bar{Y}}_R\right) = E_1E_2E_3\left(\sum_{h=1}^{L}\frac{N_h}{N}\hat{\bar{Y}}_{R,h}\right)$

$$E_1E_2E_3\left(\sum_{h=1}^{L}\left(\frac{N_h}{N}\right)\frac{\bar{y}_h}{\bar{x}_h}\bar{x}'_h\right)$$

$$= E_1E_2\left[\sum_{h=1}^{L}\frac{N_h}{N}\bar{x}'_h\,E_3\left(\frac{\bar{y}_h}{\bar{x}_h}\right)\right] = E_1E_2\left[\sum_{h=1}^{L}\frac{N_h}{N}\bar{x}'_h\frac{\bar{y}'_h}{\bar{x}'_h}\right]$$

ISSN NUMBER: 1116-249X

$$= E_1\left[\sum_{h=1}^{L}\frac{N_h}{N}E_2(\bar{y}_h')\right] = \left[\sum_{h=1}^{L}\frac{N_h}{N}E_1(\bar{Y}_h)\right] = \frac{1}{N}\sum_{h=1}^{L}N_h\bar{Y} \neq \bar{Y}$$

Thus, $\hat{\bar{Y}}_R$ is a biased estimator of population mean $\bar{Y}$ in stratified two stage sampling design using double sampling for auxiliary information at SSU level.

Appendix 5: Proof to Theorem 5

Let

$$e_0 = \frac{\bar{y}_h - \bar{Y}_h}{\bar{Y}_h} \Rightarrow \bar{y}_h = \bar{Y}_h(1+e_0), e_1 = \frac{\bar{x}_h - \bar{X}_h}{\bar{X}_h} \Rightarrow \bar{x}_h = \bar{X}_h(1+e_1), e_2 = \frac{\bar{x}_h' - \bar{X}_h}{\bar{X}_h} \Rightarrow \bar{x}_h' = \bar{X}_h(1+e_2)$$

$$E(e_0) = E\left(\frac{\bar{y}_h - \bar{Y}_h}{\bar{Y}_h}\right) = 0, E(e_1) = E\left(\frac{\bar{x}_h - \bar{X}_h}{\bar{X}_h}\right) = 0, E(e_2) = E\left(\frac{\bar{x}_h' - \bar{X}_h}{\bar{X}_h}\right) = 0$$

$$E(e_1^2) = E\left(\frac{\bar{x}_h - \bar{X}_h}{\bar{X}_h}\right)^2 = \frac{V(\bar{X}_h)}{\bar{X}_h^2}, \qquad E(e_0e_1) = E\left(\frac{(\bar{y}_h - \bar{Y}_h)(\bar{x}_h - \bar{X}_h)}{\bar{X}_h\bar{Y}_h}\right) = \frac{COV(\bar{X}_h, \bar{Y}_h)}{\bar{X}_h\bar{Y}_h}$$

$$E(e_0e_2) = E\left(\frac{(\bar{y}_h - \bar{Y}_h)(\bar{x}_h' - \bar{X}_h)}{\bar{X}_h\bar{Y}_h}\right) = \frac{COV(\bar{X}_h', \bar{Y}_h)}{\bar{X}_h\bar{Y}_h} = 0,$$

$$E(e_1e_2) = E\left(\frac{(\bar{x}_h - \bar{X}_h)(\bar{x}_h' - \bar{X}_h)}{\bar{X}_h\bar{Y}_h}\right) = \frac{cov(\bar{X}_h', \bar{X}_h)}{\bar{X}_h\bar{Y}_h} = 0$$

Equation (6) becomes

$$\hat{\bar{Y}}_R = \sum_{h=1}^{L} W_h\left[\bar{Y}_h(1+e_0)(1+e_1)^{-1}(1+e_2)\right], = \sum_{h=1}^{L} W_h\left[\bar{Y}_h(1+e_0)(1+e_2)(1-e_1+e_1^2)\right]$$

Expanding and ignoring terms of degree greater than two, we have

$$= \sum_{h=1}^{L} W_h\left[\bar{Y}_h(1+e_0+e_2+e_0e_2)(1-e_1+e_1^2)\right]$$

$$= \sum_{h=1}^{L} W_h\left[\bar{Y}_h(1-e_1+e_1^2+e_0-e_0e_1+e_0e_1^2+e_2-e_1e_2+e_1^2e_2+e_0e_2-e_0e_1e_2+e_0e_2e_1^2)\right]$$

$$= \sum_{h=1}^{L} W_h\left[\bar{Y}_h(1-e_1+e_1^2+e_0-e_0e_1+e_2-e_1e_2+e_0e_2)\right]$$

$$B\left(\hat{\bar{Y}}_R\right) = E\left(\hat{\bar{Y}}_R\right) - \bar{Y}_h$$

$$B\left(\hat{\bar{Y}}_R\right) = E\left(\sum_{h=1}^{L} W_h\bar{Y}_h\left[(1-e_1+e_1^2+e_0-e_0e_1+e_2-e_1e_2+e_0e_2)\right]\right) - \bar{Y}_h$$

$$= \sum_{h=1}^{L} W_h\bar{Y}_h\left[(e_1^2-e_0e_1)\right]$$

ISSN NUMBER: 1116-249X

$$B\left(\hat{\bar{Y}}_R\right) = \sum_{h=1}^{L} W_h \bar{Y}_h \left[\frac{V(\bar{X}_h)}{\bar{X}_h^2} - \frac{COV(\bar{X}_h, \bar{Y}_h)}{\bar{X}_h \bar{Y}_h}\right]$$

$$= \sum_{h=1}^{L} W_h \bar{Y}_h \left[\frac{V(\bar{X}_h)}{\bar{X}_h^2} - \frac{COV(\bar{X}_h, \bar{Y}_h)}{\bar{X}_h \bar{Y}_h}\right]$$

$$= \sum_{h=1}^{L} W_h \frac{1}{\bar{X}_h} \left[\frac{V(\bar{X}_h)}{\bar{X}_h} - COV(\bar{X}_h, \bar{Y}_h)\right]$$

$$B\left(\hat{\bar{Y}}_R\right) = \sum_{h=1}^{L} W_h \frac{1}{n_h \bar{X}_h} \left[\sum_{i=1}^{N_h} \frac{M_{hi}}{m'_{hi} m_{hi}} \left(R\sigma_{hix}^2 - \sigma_{hiyx}\right)\right]$$

Appendix 6: Proof to Theorem 6

The Mean Square Error (MSE) of the estimator $\hat{\bar{Y}}_R$ is defined as:

$$MSE\left(\hat{\bar{Y}}_R\right) = \sum_{h=1}^{L} W_h^2 \left[V\left(\bar{y}_{h,}\right) + R_h^2 V(\bar{x}_h) - 2RCov(\bar{y}_{h,}, \bar{x}_h)\right] \quad (18)$$

Where

$$V(\bar{y}_h) = V_1 E_2 E_3(\bar{y}_h) + E_1 V_2 E_3(\bar{y}_h) + E_1 E_2 V_3(\bar{y}_h) \quad (19)$$
$$= V_1 + V_2 + V_3$$

Now

$$V_1 = V_1 E_2 E_3(\bar{y}_h) = V_1 E_2 E_3 \left(\frac{1}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}}{m'_{hi} m_{hi}} \sum_{j=1}^{m_{hi}} \frac{y_{hij}}{P_{hj}}\right)$$

$$= V_1 E_2 \left(\frac{1}{n_h} \sum_{i=1}^{n_h} M_{hi} E_3(\bar{y}_{hi})\right) = V_1 \left(\frac{1}{n_h} \sum_{i=1}^{n_h} M_{hi} E_2(\bar{y}'_{hi})\right)$$

$$= V_1 \left(\frac{1}{n_h} \sum_{i=1}^{n_h} M_{hi} \bar{Y}_{hi}\right) = V_1 \left(\frac{1}{n_h} \sum_{i=1}^{n_h} Y_{hi}\right)$$

$$= \left(\frac{1}{n'_h} - \frac{1}{N_h}\right) \left(\frac{1}{n_h - 1} \sum_{j=1}^{m_{hi}} (Y_{hi} - \bar{Y}_h)^2\right)$$

$$= \left(\frac{1}{n'_h} - \frac{1}{N_h}\right) S_{hy}^2 \quad (20)$$

Where S_{hy}^2 can be estimated by

$$S_{hy}^2 = \frac{1}{n'_h - 1} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2$$

$$Y_{hi} = M_{hi} \bar{y}_{hi}$$

ISSN NUMBER: 1116-249X

$$V_2 = E_1V_2E_3(\bar{y}_h) = E_1V_2E_3\left(\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{M_{hi}}{m'_{hi}m_{hi}}\sum_{j=1}^{m_{hi}}\frac{y_{hij}}{P_{hj}}\right)$$

$$= E_1V_2\left(\frac{1}{n_h}\sum_{i=1}^{n_h}M_{hi}\,\bar{y}'_{hi}\right) = E_1\left(\frac{1}{n_h}\sum_{i=1}^{n_h}M_{hi}^2V_2(\bar{y}'_{hi})\right)$$

$$= E_1\left(\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{M_{hi}^2}{m_{hi}}\left(\frac{1}{m_{hi}-1}\right)S_{hiy}^{2\prime}\right)$$

$$= \frac{1}{N_h n_h}\sum_{i=1}^{N_h}M_{hi}^2\left(\frac{1}{m'_{hi}}-\frac{1}{M_{hi}}\right)S_{hiy}^2 \qquad (21)$$

where S_{hiy}^2 can be estimated by

$$S_{hiy}^2 = \frac{1}{M_{hi}-1}\sum_{j=1}^{M_{hi}}\left(y_{hij}-\bar{Y}_h\right)^2$$

$$V_3 = E_1E_2V_3(\bar{y}_h)$$

$$= E_1E_2V_3\left(\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{M_{hi}}{m'_{hi}m_{hi}}\sum_{j=1}^{m_{hi}}\frac{y_{hij}}{P_{hj}}\right) = E_1E_2\left(\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{M_{hi}^2}{m_{hi}^{2\prime}m_{hi}^2}\sum_{j=1}^{m_{hi}}V_3\left(\frac{y_{hij}}{P_{hj}}\right)\right)$$

$$= E_1E_2\left(\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{M_{hi}^2}{m_{hi}^{2\prime}m_{hi}^2}\sum_{j=1}^{m'_{hi}}\left(\frac{y_{hij}}{P_{hj}}-\bar{y}'_{hi}\right)^2\right) \quad = \frac{1}{N_h}\sum_{i=1}^{N_h}\frac{M_{hi}}{m'_{hi}m_{hi}}\sum_{j=1}^{m_{hi}}\left(\frac{y_{hij}}{P_{hj}}-\bar{Y}_h\right)^2$$

$$= \frac{1}{N_h n_h}\sum_{i=1}^{N_h}\frac{M_{hi}}{m'_{hi}m_{hi}}S_{hiy}^2 \qquad (22)$$

where

$$S_{hiy}^2 = \sum_{j=1}^{m_{hi}}\left(\frac{y_{hij}}{P_{hj}}-\bar{Y}_h\right)^2$$

Substituting (20), (21) and (22) in (19) we realize

$$V(\bar{y}_h) = \left(\frac{1}{n'_h}-\frac{1}{N_h}\right)S_{hy}^2 + \frac{1}{N_h n_h}\sum_{i=1}^{N_h}M_{hi}^2\left(\frac{1}{m'_{hi}}-\frac{1}{M_{hi}}\right)S_{hiy}^2 + \frac{1}{N_h n_h}\sum_{i=1}^{N_h}\frac{M_{hi}}{m'_{hi}m_{hi}}S_{hiy}^2$$

Similarly,

$$V(\bar{x}_h) = \left(\frac{1}{n'_h}-\frac{1}{N_h}\right)S_{hx}^2 + \frac{1}{N_h n_h}\sum_{i=1}^{N_h}M_{hi}^2\left(\frac{1}{m'_{hi}}-\frac{1}{M_{hi}}\right)S_{hix}^2 + \frac{1}{N_h n_h}\sum_{i=1}^{m_{hi}}\frac{M_{hi}}{m'_{hi}m_{hi}}S_{hix}^2$$

ISSN NUMBER: 1116-249X

Corollary: The covariance factor in Equation (18) is given by

$$\mathrm{Cov}\left(\bar{y}_{h,},\bar{x}_h\right)=\left(\frac{1}{n'_h}-\frac{1}{N_h}\right)S^2_{hyx}\ +\frac{1}{N_h n_h}\sum_{i=1}^{N_h}M^2_{hi}\left(\frac{1}{m'_{hi}}-\frac{1}{M_{hi}}\right)S^2_{hiyx}+\frac{1}{N_h n_h}\sum_{i=1}^{N_h}\frac{M_{hi}}{m'_{hi}m_{hi}}S^2_{hiyx}\quad(23)$$

Proof

Let $\mathrm{Cov}\left(\bar{y}_{h,},\bar{x}_h\right)=C_1E_2E_3(\bar{x}_h,\bar{y}_h)+E_1C_2E_3(\bar{x}_h,\bar{y}_h)+E_1E_2C_3(\bar{x}_h,\bar{y}_h)$ (24)

$$C_1E_2E_3\left(\bar{y}_{h,},\bar{x}_h\right)=\ C_1E_2E_3\left(\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{M_{hi}}{m'_{hi}m_{hi}}\sum_{j=1}^{m_{hi}}\frac{x_{hij}}{P_{hj}},\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{M_{hi}}{m'_{hi}m_{hi}}\sum_{j=1}^{m_{hi}}\frac{y_{hij}}{P_{hj}}\right)$$

$$=C_1\left(\frac{1}{n_h}\sum_{i=1}^{n_h}M_{hi}\,E_3(\bar{x}_{hi}),\frac{1}{n_h}\sum_{i=1}^{n_h}M_{hi}\,E_3(\bar{y}_{hi})\right)=\left(\frac{1}{n_h}\sum_{i=1}^{N_h}M^2_{hi}S^2_{hiy}\ \ ,\frac{1}{n_h}\sum_{i=1}^{N_h}\frac{M^2_{hi}}{m_{hi}}S^2_{hiy}\ \right)$$

$$=\left(\frac{1}{n'_h}-\frac{1}{N_h}\right)S^2_{hy}\,,\left(\frac{1}{n'_h}-\frac{1}{N_h}\right)S^2_{hx}$$

$$=\left(\frac{1}{n'_h}-\frac{1}{N_h}\right)S^2_{hyx}\quad(25)$$

$$E_1C_2E_3\left(\bar{y}_{h,},\bar{x}_h\right)=E_1C_2E_3\left(\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{M_{hi}}{m'_{hi}m_{hi}}\sum_{j=1}^{m_{hi}}\frac{x_{hij}}{P_{hj}},\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{M_{hi}}{m'_{hi}m_{hi}}\sum_{j=1}^{m_{hi}}\frac{y_{hij}}{P_{hj}}\right)$$

$$=E_1C_2\left(\frac{1}{n_h}\sum_{i=1}^{n_h}M^2_{hi}V_2(\bar{x}'_{hi}),\ \ \frac{1}{n_h}\sum_{i=1}^{n_h}M^2_{hi}V_2(\bar{y}'_{hi})\right)$$

$$=E_1\left(\frac{1}{n_h}\sum_{i=1}^{n_h}M^2_{hi}\left(\frac{1}{m_{hi}-1}\right)S^{2\prime}_{hiy}\,,\frac{1}{n_h}\sum_{i=1}^{n_h}M^2_{hi}\left(\frac{1}{m_{hi}-1}\right)S^{2\prime}_{hix}\ \right)$$

$$=\ \ \frac{1}{N_h n_h}\sum_{i=1}^{N_h}M^2_{hi}\left(\frac{1}{m'_{hi}}-\frac{1}{M_{hi}}\right)S^2_{hiyx}\quad(26)$$

where

$$S^2_{hiyx}=\frac{1}{M_{hi}-1}\sum_{j=1}^{m_{hi}}\left(\frac{y_{hij}}{P_{hj}}-\bar{Y}_h\right)^2\left(\frac{x_{hij}}{P_{hj}}-\bar{X}_h\right)^2$$

$$E_1E_2C_3\left(\bar{y}_{h,},\bar{x}_h\right)\ \ =E_1E_2C_3\left(\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{M_{hi}}{m'_{hi}m_{hi}}\sum_{j=1}^{m_{hi}}\frac{x_{hij}}{P_{hj}},\frac{1}{n_h}\sum_{i=1}^{n_h}\frac{M_{hi}}{m'_{hi}m_{hi}}\sum_{j=1}^{m_{hi}}\frac{y_{hij}}{P_{hj}}\right)$$

ISSN NUMBER: 1116-249X

$$= \left(\frac{1}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}^2}{m_{hi}^{2\prime} m_{hi}^2} \sum_{j=1}^{m_{hi}} V_3 \left(\frac{x_{hij}}{P_{hj}} \right), \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}^2}{m_{hi}^{2\prime} m_{hi}^2} \sum_{j=1}^{m_{hi}} V_3 \left(\frac{y_{hij}}{P_{hj}} \right) \right)$$

$$= E_1 E_2 \left(\frac{1}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}^2}{m_{hi}^{2\prime} m_{hi}^2} \sum_{j=1}^{m'_{hi}} \left(\frac{x_{hij}}{P_{hj}} - \bar{x}'_{hi} \right)^2, \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{M_{hi}^2}{m_{hi}^{2\prime} m_{hi}^2} \sum_{j=1}^{m'_{hi}} \left(\frac{y_{hij}}{P_{hj}} - \bar{y}'_{hi} \right)^2 \right)$$

$$= \frac{1}{N_h n_h} \sum_{i=1}^{N_h} \frac{M_{hi}}{m'_{hi} m_{hi}} S_{hiyx}^2 \qquad (27)$$

Putting the values of (25) , (26) and (27) in (24), we get

$$\mathrm{Cov}\left(\bar{y}_{h,}, \bar{x}_h\right) = \left(\frac{1}{n'_h} - \frac{1}{N_h} \right) S_{hyx}^2 + \frac{1}{N_h n_h} \sum_{i=1}^{N_h} M_{hi}^2 \left(\frac{1}{m'_{hi}} - \frac{1}{M_{hi}} \right) S_{hiyx}^2 + \frac{1}{N_h n_h} \sum_{i=1}^{N_h} \frac{M_{hi}}{m'_{hi} m_{hi}} S_{hiyx}^2$$

Putting the values of $V\left(\bar{y}_{h,}\right)$, $V\left(\bar{x}_{h,}\right)$ and $\mathrm{Cov}\left(\bar{y}_{h,}, \bar{x}_h\right)$ in (18) and factoring common terms, we have

$$\mathrm{MSE}\left(\hat{\bar{Y}}_R\right) = \sum_{h=1}^{L} W_h^2 \left[\left(\frac{1}{n'_h} - \frac{1}{N_h} \right) \left(S_{hy}^2 + R_h^2 S_{hx}^2 - 2R_h S_{hyx} \right) + \frac{1}{N_h n_h} \sum_{i=1}^{N_h} M_{hi}^2 \left(\frac{1}{m'_{hi}} - \frac{1}{M_{hi}} \right) \left(S_{hiy}^2 + R_h^2 S_{hix}^2 - 2R_h S_{hiyx} \right) + \frac{1}{N_h n_h} \sum_{i=1}^{N_h} \frac{M_{hi}}{m'_{hi} m_{hi}} \left(S_{hiy}^2 + R_h^2 S_{hix}^2 - 2R_h S_{hiyx} \right) \right]$$

Where

W_h^2 Represents the weight for stratum h

$\left(\frac{1}{n_h} - \frac{1}{N_h} \right)$ is a sampling factor adjusting the variance based on the sample and population sizes.

S_{hy}^2, S_{hx}^2 and S_{hyx} are Variances and covariance's at the stratum level.

R_h^2 is the Squared ratio of means used in ratio estimation.

Appendix 8: Proof of Theorem 8

Given that the estimator retains the same functional form as in Equation (6), but now relies on imputed values, its Mean Square Error (MSE) must account not only for the sampling error but also for the uncertainty introduced by the imputation process.

In the complete data scenario (Equation 6), the approximate MSE is obtained as given in Equation 8. However, under missing data conditions, a proportion of δ_h of the second-stage units within each stratum h are missing and are imputed. Imputation itself introduces an additional source of uncertainty, characterized by its variance σ_{Ih}^2 . Since imputation is applied at the second-stage unit

ISSN NUMBER: 1116-249X

level, the contribution of this additional variance is scaled by the second-stage sampling factor $\frac{1}{m'_{hi}m_{hi}}$

The imputation variance for a missing unit under ratio imputation is

$\sigma^2_{Ih,R} = \widehat{R}^2_h S^2_{x.h}$

Hence, the contribution of imputation variance to MSE is

$$\frac{\delta_h \sigma^2_{Ih,R}}{m'_{hi}m_{hi}} = \frac{\delta_h \widehat{R}^2_h S^2_{x.h}}{m'_{hi}m_{hi}}$$

Adding this term to the sampling variance component from the complete case scenario in theorem 8 gives

$$\begin{aligned} MSE\left(\hat{Y}_R^{imp}\right) = \sum_{h=1}^{L} W_h^2 \Bigg[& \left(\frac{1}{n_h} - \frac{1}{N_h}\right)\left(S^2_{hy} + R^2_h S^2_{hx} - 2R_h S_{hyx}\right) \\ & + \frac{1}{N_h n_h}\sum_{i=1}^{N_h} M^2_{hi}\left(\frac{1}{m'_{hi}} - \frac{1}{M_{hi}}\right)\left(S^2_{hiy} + R^2_h S^2_{hix} - 2R_h S_{hiyx}\right) \\ & + \frac{1}{N_h n_h}\sum_{i=1}^{N_h} \frac{M_{hi}}{m'_{hi}m_{hi}}\left(S^2_{hiy} + R^2_h S^2_{hix} - 2R_h S_{hiyx}\right) + \frac{\delta_h \widehat{R}^2_h S^2_{x.h}}{m'_{hi}m_{hi}} \Bigg] \end{aligned}$$

As required.

ISSN NUMBER: 1116-249X