Shalabh; Toutenburg, Helge

Working Paper

On the regression method of estimation of population mean from incomplete survey data through imputation

Discussion Paper, No. 442

Provided in Cooperation with:
Collaborative Research Center (SFB) 386: Statistical Analysis of discrete structures - Applications in Biometrics and Econometrics, University of Munich (LMU)

Suggested Citation: Shalabh; Toutenburg, Helge (2005) : On the regression method of estimation of population mean from incomplete survey data through imputation, Discussion Paper, No. 442, Ludwig-Maximilians-Universität München, Sonderforschungsbereich 386 - Statistische Analyse diskreter Strukturen, München, https://doi.org/10.5282/ubm/epub.1811

This Version is available at:
https://hdl.handle.net/10419/31123

On the regression method of estimation of population mean from incomplete survey data through imputation

Shalabh

Department of Mathematics and Statistics
Indian Institute of Technology
Kanpur 208016, India.
E-mail: shalab@iitk.ac.in ; shalabh1@yahoo.com

H. Toutenburg

Department für Statistik
Ludwig–Maximilians–Universität München
Ludwigstr. 33, 80539 Munich, Germany
E-mail : helge.toutenburg@stat.uni-muenchen.de

May 23, 2005

Abstract

When some observations in the sample data are missing, the application of the regression method is considered for the estimation of population mean with and without the use of imputation. The performance properties of the estimators based on the methods of mean imputation, regression imputation and no imputation are analyzed and the superiority of one method over the other is examined.
KEY WORDS: missing data mechanism, regression analysis, generalized additive models, imputation, MSE–superiority;

1 Introduction

Despite a careful collection of information from the selected sampling units, the sample data set is often found to contain some missing values in many surveys. It is then desirable to employ an imputation procedure for filling in the values of missing observations in order to complete the data set. The thus repaired data mimics as if there was no non–response, and permits the application of standard familiar techniques for the purpose of statistical analysis.

There are several ways to find the imputed values for the missing observations in the sample data; see, e.g., **?**, **?**, **?** and **?** for an interesting exposition. Among them, a popular technique is the method of mean imputation in which the mean of observations is employed to fill in the missing observations. As this technique does not utilize the available information on the auxiliary characteristics, one may employ the method of regression imputation in which the regression method of estimation is utilized to find the imputed values. The implications of these two techniques of imputation on the estimation of the population mean of the study characteristic are investigated in this paper.

The plan of our presentation is as follows. In Section 2, we describe the framework in which some observations in the sample data are missing randomly. The missingness of observations relates to one of the two characteristics at a time but not simultaneously. Employing the techniques of mean imputation and regression imputation, four estimators for the population mean of study characteristic arising from the regression method of estimation in survey sampling are formulated. An estimator that does not use any imputation technique is also presented. In Section 3, a comparison of the performance properties of the estimators based on imputation is reported. In Section 4, we examine the role of imputation in the formulation of effective estimators for the population mean. Section 6 offers some concluding remarks. In the last, the Appendix presents an outline for the derivation of main expressions.

2 Estimation of Population Mean

Let the means of study and auxiliary characteristics in a finite population of size N be $\bar{Y}$ and $\bar{X}$ respectively such that $\bar{X}$ is known while $\bar{Y}$ is to be estimated with the help of n pairs

$$\begin{aligned}(x_1, y_1), (x_2, y_2), (x_2, y_2), \ldots, (x_{n-p-q}, y_{n-p-q}),\\ (x_1^*, y_1^*), (x_2^*, y_2^*), \ldots, (x_p^*, y_p^*),\\ (x_1^{**}, y_1^{**}), (x_2^{**}, y_2^{**}, \ldots, (x_q^{**}, y_q^{**})\end{aligned}$$

of observations drawn from the given population according to the procedure of simple random sampling without replacement. It is specified that some observations in the sample data are missing. Following **?**, we assume that the p values $y_1^*, y_2^*, \ldots, y_p^*$ of the study characteristic and q values $x_1^{**}, x_2^{**}, \ldots, x_q^{**}$ of the auxiliary characteristic are missing randomly.

Now let us define the following quantities in the sample data:

$$
\begin{aligned}
\bar{x} &= \frac{1}{(n-p-q)} \sum^{(n-p-q)} x_i\,, \quad \bar{x}^* = \frac{1}{p}\sum^{p} x_i^*\,, \quad \bar{x}^{**} = \frac{1}{q}\sum^{q} x_i^{**}\,,\\
\bar{y} &= \frac{1}{(n-p-q)} \sum^{(n-p-q)} y_i\,, \quad \bar{y}^* = \frac{1}{p}\sum^{p} y_i^*\,, \quad \hat{y}^{**} = \frac{1}{q}\sum^{q} y_i^{**}\,,\\
s_{xx} &= \frac{1}{(n-p-q-1)} \sum^{(n-p-q)} (x_i-\bar{x})^2\,,\\
s_{xy} &= \frac{1}{(n-p-q-1)} \sum^{(n-p-q)} (x_i-\bar{x})(y_i-\bar{y})\,.
\end{aligned}
$$

The optimal difference estimator for the population mean $\bar{Y}$ is given by

$$
\frac{(n-p-q)\bar{y}+p\bar{y}^*+q\bar{y}^{**}}{n} + \beta\left[\bar{X} - \frac{(n-p-q)\bar{x}+p\bar{x}^*+q\bar{x}^{**}}{n}\right] \tag{2.1}
$$

provided that the quantity $\beta = \frac{\left[\sum^N (x_i-\bar{X})(y_i-\bar{Y})\right]}{\left[\sum^N (x_i-\bar{X})^2\right]}$ is known and no observation is missing; see, e.g., **?** (Chapter 7), **?** (Chapter 6), **?** (Chapter 10).

As β is generally unknown, a feasible version of the optimal difference estimator is chosen as follows:

$$
\bar{y}_{\text{reg}} = \frac{(n-p-q)\bar{y}+p\bar{y}^*+q\bar{y}^{**}}{n} + \frac{s_{xy}}{s_{xx}}\left[\bar{X} - \frac{(n-p-q)\bar{x}+p\bar{x}^*+q\bar{x}^{**}}{n}\right] \tag{2.2}
$$

Due to the missingness of some observation in the sample data, the means $\bar{y}^*$ and $\bar{x}^{**}$ cannot be found and consequently the estimator (2.2) cannot be used in practice.

If we follow the rule of mean imputation, we may replace $\bar{y}^*$ and $\bar{x}^{**}$ in (2.2) by $\bar{y}$ and $\bar{x}$ respectively. This proposition provides the following estimator of $\bar{Y}$:

$$
\hat{\bar{Y}}_1 = \frac{(n-q)\bar{y}+q\bar{y}^{**}}{n} + \frac{s_{xy}}{s_{xx}}\left[\bar{X} - \frac{(n-p)\bar{x}+p\bar{x}^*}{n}\right]. \tag{2.3}
$$

As $\bar{X}$ is known, it may be tempting to use $\bar{X}$ rather than $\bar{x}$ for the imputation of $\bar{x}^{**}$. This leads to the following estimator of $\bar{Y}$:

$$
\hat{\bar{Y}}_2 = \frac{(n-q)\bar{y}+q\bar{y}^{**}}{n} + \frac{s_{xy}}{s_{xx}}\left[\frac{n-q}{n}(\bar{X}-\bar{x}) + \frac{p}{n}(\bar{x}-\bar{x}^*)\right]. \tag{2.4}
$$

Instead of the rule of mean imputation, if we employ the regression method of imputation for the missing observations on the study characteristic, the imputation estimator of $\bar{y}^*$ is given by:

$$
\hat{\bar{y}}^* = \bar{y} + \frac{s_{xy}}{s_{xx}}(\bar{x}^*-\bar{x}). \tag{2.5}
$$

Using it in place of $\bar{y}^*$ and $\bar{X}$ in place of $\bar{x}^{**}$ in (2.2), we get the following estimator for $\bar{Y}$:

$$\hat{\bar{Y}}_3 = \frac{(n-q)\bar{y} + q\bar{y}^{**}}{n} + \frac{s_{xy}}{s_{xx}} \left(\frac{n-q}{n}\right) (\bar{X} - \bar{x}). \tag{2.6}$$

If we employ the regression method of imputation for the p missing values of the study characteristic and q missing values of the auxiliary characteristic so that $\bar{y}^*$ and $\bar{x}^{**}$ in (2.2) are replaced by $\hat{\bar{y}}^*$ and

$$\hat{\bar{x}}^{**} = \bar{x} + \frac{s_{yy}}{s_{xy}} (\bar{y}^{**} - \bar{y}), \tag{2.7}$$

we find the following estimator of $\bar{Y}$:

$$\hat{\bar{Y}}_4 = \frac{(n-q)\bar{y} + q\bar{y}^{**}}{n} + \frac{s_{xy}}{s_{xx}} (\bar{X} - \bar{x}) + \frac{q s_{xy}^2}{n s_{xx} s_{yy}} (\bar{y} - \bar{y}^{**}). \tag{2.8}$$

We have thus formulated four estimators of the population mean $\bar{Y}$ employing the methodology of imputation for the missing values in the sample data.

If we do not use any imputation procedure and simply employ the available observations in the sample data, we may approximate (2.2) by the following

$$\tilde{\bar{Y}} = \frac{(n-p-q)\bar{y} + q\bar{y}^{**}}{(n-p)} + \frac{s_{xy}}{s_{xx}} \left[\bar{X} - \frac{(n-p-q)\bar{x} + p\bar{x}^*}{(n-q)}\right] \tag{2.9}$$

which may serve as an estimator of $\bar{Y}$.

3 Comparison of Estimators

In order to compare the performance properties of the estimators of $\bar{Y}$ under the criteria of the relative bias and relative mean squared error to the first order of approximation, we introduce the following quantities in the population:

$$\begin{aligned}
C_Y^2 &= \frac{1}{N} \sum^N \left(\frac{Y_i - \bar{Y}}{\bar{Y}}\right)^2 \\
\rho &= \frac{\sum^N (X_i - \bar{X})(Y_i - \bar{Y})}{\left[\sum^N (X_i - \bar{X})^2 * \sum^N (Y_i - \bar{Y})^2\right]^{\frac{1}{2}}} \\
\theta &= \left(\frac{\bar{Y}}{\bar{X}}\right) \left[\frac{\sum (X_i - \bar{X})^2}{\sum (Y_i - \bar{Y})^2}\right]^{\frac{1}{2}} \\
K &= \frac{1}{\theta^3 C_Y^2 N} \sum^N \left(\frac{X_i - \bar{X}}{\bar{X}}\right)^2 \left[\rho \left(\frac{X_i - \bar{X}}{\bar{X}}\right) - \theta \left(\frac{Y_i - \bar{Y}}{\bar{Y}}\right)\right] \\
G &= \frac{\rho}{\theta^3 C_Y^2 N} \sum^N \left[\rho \left(\frac{X_i - \bar{X}}{\bar{X}}\right)^2 + \rho\theta^2 \left(\frac{Y_i - \bar{Y}}{\bar{Y}}\right)^2 \right. \\
&\quad \left. -\theta \left(\frac{X_i - \bar{X}}{\bar{X}}\right) \left(\frac{Y_i - \bar{Y}}{\bar{Y}}\right)\right] * \left(\frac{Y_i - \bar{Y}}{\bar{Y}}\right).
\end{aligned} \tag{3.1}$$

Theorem I: To the first order of approximation, the relative biases of the estimators $\hat{\bar{Y}}_1, \hat{\bar{Y}}_2, \hat{\bar{Y}}_3$ and $\hat{\bar{Y}}_4$ are given by

$$\begin{aligned} RB(\hat{\bar{Y}}_1) &= E\left(\frac{\hat{\bar{Y}}_1 - \bar{Y}}{\bar{Y}}\right) \\ &= \frac{K}{n} E_{p,q}\left(\frac{n-p}{n-p-q}\right) \qquad (3.2) \\ RB(\hat{\bar{Y}}_2) &= E\left(\frac{\hat{\bar{Y}}_2 - \bar{Y}}{\bar{Y}}\right) \\ &= \frac{K}{n} \qquad (3.3) \\ RB(\hat{\bar{Y}}_3) &= E\left(\frac{\hat{\bar{Y}}_3 - \bar{Y}}{\bar{Y}}\right) \\ &= \frac{K}{n} E_{p,q}\left(\frac{n-q}{n-p-q}\right) \qquad (3.4) \\ RB(\hat{\bar{Y}}_4) &= E\left(\frac{\hat{\bar{Y}}_4 - \bar{Y}}{\bar{Y}}\right) \\ &= K E_{p,q}\left(\frac{1}{n-p-q}\right) - \frac{G}{n} E_{p,q}\left(\frac{q}{n-p-q}\right) \qquad (3.5) \end{aligned}$$

where the operators $E_{p,q}$ refers to the expectation taken over all possible values of the non–negative integer valued random variables p and q in the sample of fixed size n.
Proof: See Appendix.

Looking at the expressions for the relative biases, it is obvious that all the four estimators are generally biased.

When $\bar{X}$ is used for the imputation of missing values of the auxiliary characteristics, it is seen from (3.3) and (3.4) that the estimator $\hat{\bar{Y}}_2$ has smaller magnitudes of bias in comparison to $\hat{\bar{Y}}_3$. It means that the method of mean imputation for the missing values of the study characteristics has better performance than the method of regression imputation with respect to the criterion of the magnitudes of bias.

On the other hand, when $\bar{X}$ is not used for the purpose of imputation of missing values of the auxiliary characteristic and instead the same method of imputation is applied to find the substitutes for the missing values of both the study and auxiliary characteristics in the sample data, we observe from (3.2) and (3.4) that the magnitude of bias of the estimator $\hat{\bar{Y}}_1$ is smaller than that of $\hat{\bar{Y}}_4$ when

$$(h_1 K - G)(h_2 K - G) > 0 \qquad (3.6)$$

where

$$h_1 = E_{p,q}\left(\frac{p}{n-p-q}\right)\left[E_{p,q}\left(\frac{q}{n-p-q}\right)\right]^{-1} \tag{3.7}$$

$$h_2 = E_{p,q}\left(\frac{2n-p}{n-p-q}\right)\left[E_{p,q}\left(\frac{q}{n-p-q}\right)\right]^{-1}. \tag{3.8}$$

The condition (3.6) holds true when the quantities G and K have opposite signs. If G and K have same signs, the condition (3.6) is satisfied so long as any one of the following inequalities holds:

$$0 < G < h_1 K \tag{3.9}$$

$$G > h_2 K > 0 \tag{3.10}$$

$$h_1 K < G < 0 \tag{3.11}$$

$$G < h_2 K < 0. \tag{3.12}$$

On the other hand, the condition (3.6) with a reversed inequality sign holds when one of the following is true:

$$0 > h_1 K < G < h_2 K \tag{3.13}$$

$$h_2 < G < h_1 K < 0 \tag{3.14}$$

which specifies the situations where the method of regression imputation is preferable to the method of mean imputation under the criterion of the magnitude of bias.

Next, let us examine the mean squared errors to the first order of approximation.

Theorem II: To the first order of approximation, the relative mean squared

errors of the estimators $\hat{\bar{Y}}_1$, $\hat{\bar{Y}}_2$, $\hat{\bar{Y}}_3$ and $\hat{\bar{Y}}_4$ are given by

$$\begin{aligned}
RMSE(\hat{\bar{Y}}_1) &= E\left(\frac{\hat{\bar{Y}}_1 - \bar{Y}}{\bar{Y}}\right)^2 \\
&= D - \frac{C_Y^2\rho^2}{n^2}\left[n - E_{p,q}\left(\frac{q(n-3p)}{n-p-q}\right)\right] \\
RMSE(\hat{\bar{Y}}_2) &= E\left(\frac{\hat{\bar{Y}}_2 - \bar{Y}}{\bar{Y}}\right)^2 \qquad (3.15) \\
&= D - \frac{C_Y^2\rho^2}{n^2}E_{p,q}(n-q) \qquad (3.16) \\
RMSE(\hat{\bar{Y}}_3) &= E\left(\frac{\hat{\bar{Y}}_3 - \bar{Y}}{\bar{Y}}\right)^2 \\
&= D - \frac{C_Y^2\rho^2}{n^2}E_{p,q}\left[\frac{(n-q)^2}{n-p-q}\right] \qquad (3.17) \\
RMSE(\hat{\bar{Y}}_4) &= E\left(\frac{\hat{\bar{Y}}_4 - \bar{Y}}{\bar{Y}}\right)^2 \\
&= D - \frac{C_Y^2\rho^2}{n^2}\left[n - E_{p,q}\left(\frac{q(n+2p)-np}{n-p-q}\right)\right] \\
&\quad -\frac{C_Y^2\rho^4}{n^2}E_{p,q}\left[\frac{q(n+p)}{n-p-q}\right] \qquad (3.18)
\end{aligned}$$

where

$$D = \frac{C_Y^2}{n}\left[1 + E_{p,q}\left(\frac{p(n-q)}{n(n-p-q)}\right)\right]. \qquad (3.19)$$

Proof: See Appendix.

Looking at the expressions for the relative mean squared errors of the estimators, we observe that the sign (positive or negative) of the correlation coefficient between the study and auxiliary characteristics has no influence on the performance of estimators.

When we employ $\bar{X}$ to impute the average of missing values on the auxiliary characteristic, a comparison of (3.15) and (3.17) reveals that the estimator based on the method of regression imputation is more efficient than the one based on the method of mean imputation for the missing values of the study characteristic.

Comparing $\hat{\bar{Y}}_1$ and $\hat{\bar{Y}}_4$, it is observed from (3.15) and (3.15) that $\hat{\bar{Y}}_4$ has smaller mean squared mean squared error than $\hat{\bar{Y}}_1$ when

$$\rho^2 > E_{p,q}\left(\frac{p(5q-n)}{n-p-q}\right)\left[E_{p,q}\left(\frac{q(n+p)}{n-p-q}\right)\right]^{-1} \qquad (3.20)$$

which is a condition for the superiority of the method of regression imputation over the method of mean imputation.

The opposite is true, i. e., the mean imputation provides more efficient estimator of $\bar{Y}$ than the regression imputation when the inequality (3.20) holds with a reversed sign.

An interesting particular case arises when it is known that the sample contains no missing values of the auxiliary characteristic, i. e., the random variable q takes the value zero with probability one. In this case, the estimators $\hat{\bar{Y}}_1$ and $\hat{\bar{Y}}_2$ are identical. Similarly, $\hat{\bar{Y}}_3$ and $\hat{\bar{Y}}_4$ become equal. Further, it follows from the results of Theorem I and Theorem II that the regression imputation technique leads to a decrease in the mean squared error accompanied with an increase in the magnitude of bias when compared with the technique of mean imputation.

Similarly, when the sample is known to contain only the missing values of the auxiliary characteristic, i. e., the random variable p takes the value zero with probability one, we observe that $\hat{\bar{Y}}_2$ and $\hat{\bar{Y}}_3$ become identical. Now, from the results in Theorem I and Theorem II, it is seen that the estimators $\hat{\bar{Y}}_2$ and $\hat{\bar{Y}}_3$ posses not only smaller magnitude of bias but lower mean squared error too in comparison to the estimator $\hat{\bar{Y}}_1$.

Comparing $\hat{\bar{Y}}_4$ with $\hat{\bar{Y}}_2$ and $\hat{\bar{Y}}_3$, it is observed from Theorem I that the estimator $\hat{\bar{Y}}_2$ and $\hat{\bar{Y}}_3$ will have smaller magnitude of bias in comparison to $\hat{\bar{Y}}_4$ when

$$(K - G)(hK - G) > 0 \tag{3.21}$$

where

$$h = \left[1 + E_q\left(\frac{n}{n-q}\right)\right]\left[E_q\left(\frac{q}{n-q}\right)\right]^{-1} > 1. \tag{3.22}$$

The condition (3.21) holds true when the quantities G and K have opposite signs. When they have same signs, the condition (3.21) is satisfied as long as any one of following inequalities is true:

$$0 < G < K \tag{3.23}$$

$$0 < hK < G \tag{3.24}$$

$$G < hK < 0 \tag{3.25}$$

$$K < G < 0. \tag{3.26}$$

Similarly, when

$$\rho^2 < \frac{1}{n} E_q\left(\frac{q^2}{n-q}\right)\left[E_q\left(\frac{nq}{n-q}\right)\right]^{-1} \tag{3.27}$$

the estimators $\hat{\bar{Y}}_2$ and $\hat{\bar{Y}}_3$ are found to have lower mean squared error than $\hat{\bar{Y}}_4$.

4 Usefulness of Imputation

It may be observed that all the five formulated estimators of $\bar{Y}$ make full utilization of the available values in the sample data. Out of these, the estimator

$\tilde{\bar{Y}}$ specified by (2.9) does not utilize the technique of imputation while the remaining four estimators $\hat{\bar{Y}}_1$, $\hat{\bar{Y}}_2$, $\hat{\bar{Y}}_3$ and $\hat{\bar{Y}}_4$ do. Thus a comparison of $\tilde{\bar{Y}}$ with $\hat{\bar{Y}}_1$, $\hat{\bar{Y}}_2$, $\hat{\bar{Y}}_3$ and $\hat{\bar{Y}}_4$ may shed light on the role of imputation on the construction of estimators for $\bar{Y}$.

Theorem III: The first order approximations for the relative bias and relative mean squared error of the estimator $\tilde{\bar{Y}}$ are given by

$$\begin{aligned} RB(\tilde{\bar{Y}}) &= E\left(\frac{\tilde{\bar{Y}}-\bar{Y}}{\bar{Y}}\right) \\ &= KE_{p,q}\left(\frac{1}{n-q}\right) \qquad (4.1) \\ RMSE(\tilde{\bar{Y}}) &= E\left(\frac{\tilde{\bar{Y}}-\bar{Y}}{\bar{Y}}\right)^2 \\ &= D-\frac{C_Y^2}{n^2}E_{p,q}\left[\frac{qp^2}{(n-p)(n-p-q)}\right] \\ &\quad -\rho^2 C_Y^2 E_{p,q}\left[\frac{n-p-2q}{(n-p)(n-q)}\right] \qquad (4.2) \end{aligned}$$

Proof: See Appendix.

Like the estimators based on an imputation procedure, the estimator (4.1) is also generally biased.

Comparing (4.1) with the results mentioned in Theorem I, we observe that the estimator $\tilde{\bar{Y}}$ has always smaller magnitude of bias than $\hat{\bar{Y}}_1$ and $\hat{\bar{Y}}_2$. The estimator $\tilde{\bar{Y}}$ continues to have smaller magnitude of bias than $\hat{\bar{Y}}_3$ provided that

$$E_{p,q}\left[\frac{n(p-q)+pq}{n(n-q)(n-p-q)}\right] > 0. \qquad (4.3)$$

Similarly, the estimator $\tilde{\bar{Y}}$ remains superior to $\hat{\bar{Y}}_4$ under the criterion of the magnitude of bias when

$$(f_1K-G)(f_2K-G) > 0 \qquad (4.4)$$

with

$$\begin{aligned} f_1 &= nE_{p,q}\left(\frac{p}{(n-q)(n-p-q)}\right)\left[E_{p,q}\left(\frac{q}{n-p-q}\right)\right]^{-1} \qquad (4.5) \\ f_2 &= nE_{p,q}\left(\frac{2(n-q)-p}{(n-q)(n-p-q)}\right)\left[E_{p,q}\left(\frac{q}{n-p-q}\right)\right]^{-1}. \qquad (4.6) \end{aligned}$$

As f_1 is less than f_2 , the condition (4.4) is satisfied when G and K have opposite signs. When both are either positive or negative, the condition (4.4) holds so

long as any one of the following is true:

$$0 < G < f_1 K \tag{4.7}$$
$$0 < f_2 K < G \tag{4.8}$$
$$f_1 K < G < 0 \tag{4.9}$$
$$G < f_2 K < 0. \tag{4.10}$$

Comparing (4.2) with the results stated in Theorem II, we observe that

$$\begin{aligned} RMSE(\hat{\bar{Y}}_1) - RMSE(\tilde{\bar{Y}}) &= \frac{C_Y^2}{n^2} E_{p,q}\left[\frac{qp^2}{(n-p)(n-p-q)}\right] \\ &\quad + \frac{\rho^2 C_y^2}{n^2} E_{p,q}(Z) \end{aligned} \tag{4.11}$$

where

$$Z = \frac{pq[2n(n-p-q) + (n-p)(2n-q)]}{(n-p)(n-q)(n-p-q)}. \tag{4.12}$$

It is thus seen that the estimator $\tilde{\bar{Y}}$ is always superior to $\hat{\bar{Y}}_1$. This implies that the imputation by mean is not necessarily a good strategy so far as the estimation of $\bar{Y}$ is concerned.

Similarly, from (3.15), (3.17) and (4.2), we have

$$RMSE(\hat{\bar{Y}}_2) - RMSE(\tilde{\bar{Y}}) = \frac{C_Y^2}{n^2}(1-\rho^2 d_2) E_{p,q}\left[\frac{qp^2}{(n-p)(n-p-q)}\right] \tag{4.13}$$

$$RMSE(\hat{\bar{Y}}_3) - RMSE(\tilde{\bar{Y}}) = \frac{C_Y^2}{n^2}[(1-\rho^2(d_2+d_3)] E_{p,q}\left[\frac{qp^2}{(n-p)(n-p-q)}\right] \tag{4.14}$$

where

$$d_2 = \left[E_{p,q}\left(\frac{qp^2}{(n-p)(n-p-q)}\right)\right]^{-1} E_{p,q}\left[\frac{q(2np+nq-pq)}{(n-p)(n-q)}\right] > 1 \tag{4.15}$$

$$d_3 = \left[E_{p,q}\left(\frac{qp^2}{(n-p)(n-p-q)}\right)\right]^{-1} E_{p,q}\left[\frac{p(n-q)}{(n-p-q)}\right]. \tag{4.16}$$

When $\bar{X}$ is used for the imputation of missing values of the auxiliary characteristic, the imputation of missing values of the study characteristic by the sample mean $\bar{y}$ leads to the estimator $\hat{\bar{Y}}_2$ which is more efficient in comparison to the estimator $\tilde{\bar{Y}}$ provided that

$$\rho^2 > \left(\frac{1}{d_2}\right). \tag{4.17}$$

Instead of the method of mean imputation if we follow the method of regression imputation and continue to substitute $\bar{X}$ as before, the estimator $\hat{\bar{Y}}_3$ is more efficient than $\tilde{\bar{Y}}$ when

$$\rho^2 > \left(\frac{1}{d_2 + d_3}\right). \tag{4.18}$$

Looking at (4.17) and (4.18), it may be noticed that the superiority of regression imputation over no imputation holds for a relatively wider range of situations when compared with the superiority of mean imputation over no imputation.

It can be well appreciated from (3.18) and (4.2) that it is hard to find any clear condition for the superiority of $\tilde{\bar{Y}}$ over $\hat{\bar{Y}}_4$ or vice–versa. However, if the missingness pertains to only one of the two characteristics, i. e., either of the random variables takes the value zero with probability one, it is interesting to find that the regression imputation is definitely a better strategy than no imputation.

5 Monte-Carlo Simulation Study

We conducted a Monte-Carlo Simulation experiment to study the behavior of the estimators arising after five proposed ways to impute the missing values. In fact, the large sample asymptotic approximation theory gives an idea of the behavior of the distribution of the estimator in the central part of distribution only. The Monte-Carlo study may shed some light on the finite (analytic) sample properties and overall performance of these estimators.

We considered the set up of a linear regression model $y_i = \alpha + \beta X_i + \epsilon_i (i = 1, 2, \ldots, n)$ with $\epsilon_i \sim N(0, \sigma_\epsilon^2)$. The $(n - p - q)$ complete observations on X_i, say X_{comp}, are generated following the normal distribution $N(\mu_x, \sigma_x^2)$, where μ_X and σ_X^2 are pre-specified. The corresponding $(n - p - q)$ complete observations on y_i, say y_{comp}, are then generated following $y_{comp} = \alpha + \beta X_{comp} + \epsilon$ with preassigned α and β. The p corresponding observations on X_i when y_i are missing are obtained following $N(\mu_X, \sigma_\epsilon^2)$ whereas q observations on y_i when $X_i's$ are missing are obtained from $N(\alpha + \beta\mu_X, \sigma_X^2 + \sigma_\epsilon^2)$. We considered the following setup of values:
$\alpha = 1$, $\beta = 1$, $\mu_X = 4$, $\sigma_X^2 = 0.4, 0.1$, $\sigma_\epsilon^2 = 0.6, 1$ and $n = 20, 40, 100, 200$.

The values of p and q are varied from 5% to 50% and then different combinations of them are considered. The expected value of estimators and their expected mean squared errors are calculated on the basis of 15000 replications. These values are reported in Tables 1 - 4. The expected values of estimators with $\sigma_X^2 = 0.6$, $\sigma_\epsilon^2 = 0.4$ and $\sigma_X^2 = 1$, $\sigma_\epsilon^2 = 1$ are given in Tables 1 and 2 respectively. The expected value of mean squared errors of the estimators with $\sigma_X^2 = 0.6$, $\sigma_\epsilon^2 = 0.4$ and $\sigma_X^2 = 1$, $\sigma_\epsilon^2 = 1$ are reported in Table 3 and 4 respectively.

First of all, we study the behavior of estimators under the criterion of magnitude of bias. We observe from Tables 1 and 2 that when p and q are small, say 5% each, the magnitude of bias is almost same for all the five estimators, viz., $\hat{\bar{Y}}_1$, $\hat{\bar{Y}}_2$, $\hat{\bar{Y}}_3$, $\hat{\bar{Y}}_4$ and $\hat{\bar{Y}}_5$. This simply indicates that if the percentage of missing observation is very low then all the estimators are almost equally bias efficient. When σ_X^2 and σ_ϵ^2 are high, say $\sigma_X^2 = \sigma_\epsilon^2 = 1$, then $\hat{\bar{Y}}_3$ and $\hat{\bar{Y}}_4$ turns out to be most preferred estimators whereas $\hat{\bar{Y}}_1$ and $\hat{\bar{Y}}_5$ emerges out to be the least preferred estimators. This observation holds true even when the number of missing observations in X and y are high. The dominance of $\hat{\bar{Y}}_4$ over $\hat{\bar{Y}}_3$ increases as in small samples as σ_X^2 and σ_ϵ^2 increases. There is not much variation in the magnitude of bias when σ_X^2 and σ_ϵ^2 increases. The estimator $\hat{\bar{Y}}_5$ seems to have its utility only when the percentage of missing observations is very high, say more than $60 - 70\%$. In the small to moderately large sample sizes, say up to 40, the magnitude of bias is relatively high when σ_X^2 and σ_ϵ^2 along with percentage of missing observations are high, say $\sigma_X^2 = \sigma_\epsilon^2 = 1$ and p and q are more than 50%. On the other hand, the magnitude of bias is not much high, when σ_X^2 and σ_ϵ^2 are small. So, overall, $\hat{\bar{Y}}_3$ and $\hat{\bar{Y}}_4$ emerges out to be better than others and which indicates that the regression method of imputation gives more efficient results under the criterion of magnitude of bias.

Next, we consider the criterion of mean squared error to study the behavior of these estimators. The overall picture of the mean squared errors indicate that $\hat{\bar{Y}}_1$ and $\hat{\bar{Y}}_5$ are the least preferred estimators whereas $\hat{\bar{Y}}_3$ and $\hat{\bar{Y}}_4$ emerge out to be more favored than other estimators. This clearly indicates that the use of the regression method of estimation to impute missing observation yields more efficient estimator than based on mean imputation. The dominance of $\hat{\bar{Y}}_4$ over $\hat{\bar{Y}}_3$ increases as σ_X^2 and σ_ϵ^2 increases.. Even in those cases, when percentage of missing observations is small, the difference in the mean squared error of estimators is clearly visible from Tables 3 and 4. As the sample size increases, the dominance of $\hat{\bar{Y}}_4$ over $\hat{\bar{Y}}_3$ also increases over different combinations of p and q. The overall study indicates that $\hat{\bar{Y}}_3$ and $\hat{\bar{Y}}_4$, based on regression method of imputation, yield better results.

As it has been mentioned earlier that the magnitude of bias of all the five estimators is almost same when the percentage of missing observations is small but it is to be noted here that even in such cases, the estimators $\hat{\bar{Y}}_3$ and $\hat{\bar{Y}}_4$ have smaller mean squared error than other estimators. Also, an inter-comparison between $\hat{\bar{Y}}_3$ and $\hat{\bar{Y}}_4$ reveals that $\hat{\bar{Y}}_4$ is more dominant than $\hat{\bar{Y}}_3$ over other estimators. So the overall performance under both the bias and mean squared error criterion indicates that regression method of imputation provides more efficient results than other approaches of imputation.

6 Some Concluding Remarks

When p values of the study characteristic and q values of the auxiliary characteristic are missing randomly in the sample data of size n, we have considered

the application of the regression method for estimating the population mean of the study characteristic with and without the use of imputation. In all, five estimators are formulated.

An interesting result emerging from our investigations is that the performance properties of all the estimators under study remain unchanged whether the study and auxiliary characteristics have positive correlation or negative correlation.

Comparing the estimators based on the methodology of imputation, it is found that the regression method of imputation for the missing values of the study characteristic invariably provides an estimator having smaller mean squared error at the cost of larger magnitude of bias in comparison to the method of mean imputation provided that the population mean $\bar{X}$ is used for the purpose of imputation of missing values of the auxiliary characteristic. This result does not remain true when $\bar{X}$ is not employed for imputation and instead the complete part of sample data is utilized for finding the imputed values of missing observations of both the characteristics.

Examining the usefulness of the method of mean imputation, it is interesting to find that the imputation for the missing values from the complete part of sample data is not at all a good strategy as it leads to an estimator which is worse on both the fronts of magnitude of bias and mean squared error in comparison to the strategy of no imputation. The poor performance of the mean imputation procedure slightly improves when the population mean $\bar{X}$ rather than the sample mean $\bar{x}$ is used for the imputation of the missing values of auxiliary characteristic. It continues to provide an estimator with larger magnitude of bias but now the mean squared error may decline in some cases.

Similarly, when we study the impact of the method of regression imputation on the performance of estimators for $\bar{Y}$, no neat inferences could be drawn under the criterion of neither the magnitude of bias nor the mean squared error. However, we could get clear evidence for the superiority of the strategy of regression imputation over the strategy of no imputation under the criterion of mean squared error when missingness pertains to only one characteristic, i. e., the sample data contains either the p missing values of the study characteristic or the q missing values of the auxiliary characteristic.

The Monte-Carlo simulation results also reveals the estimators based on regression method of imputation to be the winner in overall performance under the criterion of bias and mean squared error. The replacement of missing values by the sample mean and population mean of available observations does not come out to be the good strategies. The results obtained under different combinations of missing percentage of observations in study and explanatory variables may give some guidelines to applied workers for choosing a good strategy for imputing the missing values.

Table 6.1: Expected value of estimators with $\sigma_X^2 = 0.6, \sigma_\epsilon^2 = 0.4$

n	p	q	$\hat{\bar{Y}}_1$	$\hat{\bar{Y}}_2$	$\hat{\bar{Y}}_3$	$\hat{\bar{Y}}_4$	$\hat{\bar{Y}}_5$
20	2	2	5.1398	5.1376	4.9671	4.9455	5.1617
20	2	5	5.3445	5.2844	5.1643	5.2801	5.3796
20	2	10	5.9984	5.2517	5.5706	5.8974	5.7013
20	5	2	3.6619	3.6046	5.2503	5.2839	3.4828
20	5	5	5.6166	5.4147	5.6263	5.9265	5.5267
20	10	2	4.0816	4.1431	4.5725	4.8124	3.9284
40	2	2	4.8137	4.8213	4.8522	4.8149	4.8127
40	2	4	4.8364	4.8367	4.8738	5.0207	4.8282
40	2	8	4.8639	4.9325	4.8393	4.922	4.8826
40	4	2	4.9241	4.9247	4.9532	4.6825	4.9349
40	4	4	5.0646	5.0552	5.0084	4.8424	5.0817
40	4	8	5.1394	5.1845	5.153	4.9823	5.1539
40	8	2	4.8098	4.8211	5.0019	4.8107	4.825
40	8	4	5.1355	5.141	5.1538	5.0397	5.1506
40	8	8	4.7949	4.7442	4.9412	4.9247	4.7582
40	10	10	5.3047	5.1658	5.2823	5.3813	5.282
40	10	15	4.5868	4.8661	4.9082	4.77	4.5358
40	10	20	5.5527	4.9342	5.6356	6.1353	4.8645
40	15	10	4.6178	4.9684	3.9284	3.6471	4.9526
40	20	10	5.508	5.7974	4.9705	4.6107	6.3741
100	5	5	5.3306	4.9471	5.6601	6.2604	4.9707
100	10	10	4.9491	4.9493	4.9604	4.9979	4.9473
100	10	20	5.1524	5.144	5.0914	4.8895	5.1676
100	10	30	5.1403	5.1032	5.0956	5.0388	5.1473
100	10	40	4.7979	4.851	4.8599	4.7289	4.7976
100	10	50	5.1046	5.0381	5.1109	5.1733	5.0564
100	20	10	5.3755	5.1187	5.1913	4.958	5.3285
100	20	20	4.9019	4.9132	4.9173	5.0585	4.8795
100	20	30	4.7601	4.8335	4.8126	4.906	4.7441
100	20	40	5.2835	5.1958	5.1328	4.9988	5.34
100	30	10	5.0666	5.0197	5.0439	4.9943	5.0596
100	30	20	4.9416	4.9441	4.8468	4.8966	4.9437
100	30	30	4.8769	4.911	4.8062	4.7623	4.9052
100	30	30	5.1777	5.1814	5.1165	4.983	5.231
100	40	10	4.6988	4.729	4.8038	4.8375	4.6741
100	40	20	4.7793	4.8203	4.8897	4.8935	4.7461
100	50	10	4.9052	4.9244	4.896	4.9568	4.8694
200	10	10	5.0362	5.0327	5.015	5.0846	5.0354
200	10	20	4.9695	4.9728	4.9731	4.9023	4.9712
200	10	40	4.9566	4.9483	4.9809	4.9497	4.9494
200	10	60	5.0508	5.0201	5.0406	5.0547	5.0424
200	10	80	5.1508	5.0991	5.1068	5.0218	5.1497
200	10	100	5.1612	5.1398	5.137	4.9925	5.1669
200	10	120	5.0233	5.0433	5.0368	5.0058	5.0333
200	30	10	5.0263	5.0291	5.0344	5.0191	5.0269
200	30	30	5.0797	5.0729	5.0506	4.9553	5.0928
200	30	50	4.8459	4.8713	4.8672	4.9592	4.8363
200	30	70	5.0455	5.0279	5.0031	5.1228	5.0499
200	30	90	5.1108	5.0411	5.0679	4.9921	5.0998
200	50	10	5.0169	5.0121	5.011	5.0166	5.0168
200	50	30	5.0928	5.077	5.1138	5.0768	5.0923
200	50	50	4.9998	4.9731	5.0268	5.0403	4.9837
200	50	70	5.1621	5.1411	5.0845	4.7093	5.2456
200	70	10	4.9134	4.9104	4.9763	4.9477	4.9177
200	70	30	5.0548	5.0708	5.0153	5.0025	5.0639
200	70	50	5.102	5.092	5.1309	5.0693	5.1049
200	90	10	5.0287	5.0275	5.0051	4.9567	5.0483
200	90	30	4.9582	4.9664	5.0115	5.0035	4.9501
200	100	10	4.942	4.9438	4.9299	4.9185	4.9481
200	100	20	4.8393	4.8447	4.8832	4.8943	4.8293

Table 6.2: Expected value of estimators with $\sigma_X^2 = 1, \sigma_\epsilon^2 = 1$

n	p	q	$\hat{\hat{Y}}_1$	$\hat{\hat{Y}}_2$	$\hat{\hat{Y}}_3$	$\hat{\hat{Y}}_4$	$\hat{\hat{Y}}_5$
20	2	2	5.6782	5.5814	5.4989	5.1473	5.7089
20	2	5	5.4774	5.5138	5.3307	5.0263	5.5562
20	2	10	6.4876	5.8861	5.6253	6.5352	6.7302
20	5	2	5.3423	5.3925	5.0989	5.0432	5.3754
20	5	5	4.9326	4.7611	4.9306	5.3827	4.844
20	10	2	6.1109	6.0642	5.7437	5.5357	6.2714
40	2	2	5.0973	5.1117	5.1148	5.0231	5.0998
40	2	4	5.2573	5.2204	5.1375	5.2659	5.264
40	2	8	4.8071	4.8434	4.7822	4.8683	4.8202
40	4	2	4.9154	4.929	4.9524	4.9797	4.9123
40	4	4	4.7353	4.7471	4.7483	5.0346	4.7166
40	4	8	4.7155	4.6703	4.8253	5.1715	4.6568
40	8	2	4.8468	4.8421	4.918	4.9023	4.8455
40	8	4	4.7543	4.7573	4.7374	5.0571	4.709
40	8	8	4.7882	4.8782	4.7864	4.8868	4.784
40	10	10	6.2258	5.9448	5.8521	5.7036	6.3099
40	10	15	5.3452	5.2346	5.0138	5.0852	5.4875
40	10	20	4.5949	4.5083	5.077	4.6944	4.0655
40	15	10	4.1718	4.4934	4.4528	4.8077	4.0837
40	20	10	5.8561	5.4299	5.9121	6.1855	5.7788
100	5	5	5.0371	5.0386	5.0168	5.0007	5.0387
100	10	10	5.1796	5.1722	5.1173	5.0559	5.1889
100	10	20	4.9605	4.9632	4.9456	4.9263	4.9659
100	10	30	5.2895	5.267	5.2558	5.1023	5.3033
100	10	40	5.1566	5.1337	5.1381	5.193	5.151
100	10	50	4.5797	4.7289	4.7826	5.0748	4.506
100	20	10	5.0917	5.0936	4.9634	4.9449	5.1079
100	20	20	4.9582	4.9798	4.9842	4.8911	4.9667
100	20	30	4.9838	4.8771	4.9753	4.9131	4.9645
100	20	40	5.1305	4.9877	5.0196	4.7789	5.1513
100	30	10	4.7703	4.7751	4.8265	5.1039	4.7139
100	30	20	4.6723	4.7175	4.802	5.0037	4.6134
100	30	30	4.9316	4.8952	4.8655	5.2208	4.8889
100	30	30	4.9381	4.8677	5.0558	4.975	4.8848
100	40	10	5.2302	5.2192	5.33	5.1214	5.3023
100	40	20	5.1403	5.0913	4.9343	5.008	5.1702
100	50	10	4.9037	4.8829	5.0724	4.9681	4.9255
200	10	10	5.0636	5.0592	5.045	5.1209	5.0626
200	10	20	5.1493	5.1545	5.1444	5.1263	5.1508
200	10	40	5.0313	5.0172	5.0002	4.8657	5.0394
200	10	60	4.8669	4.8654	4.8697	5.0253	4.8608
200	10	80	4.9544	4.9929	5.0172	5.2334	4.9314
200	10	100	4.9243	5.0006	4.969	4.9543	4.9542
200	10	120	5.0017	5.0796	5.0901	5.1122	4.9833
200	30	10	5.0459	5.0496	5.0137	5.0104	5.0478
200	30	30	5.0619	5.0986	5.0831	4.9677	5.0697
200	30	50	4.952	4.9718	4.9751	5.0102	4.9461
200	30	70	5.1712	5.1681	5.1809	4.8909	5.1948
200	30	90	4.8545	4.814	4.7895	5.0131	4.8584
200	50	10	5.1489	5.1432	5.1089	5.0975	5.1536
200	50	30	5.0222	5.0287	5.048	4.9857	5.0268
200	50	50	4.9091	4.9243	4.9772	5.1979	4.8465
200	50	70	5.1534	5.1683	5.1102	5.1036	5.1828
200	70	10	5.1478	5.1427	5.1183	5.2178	5.1238
200	70	30	5.174	5.1856	5.1186	4.9023	5.2342
200	70	50	5.072	5.0569	5.0194	5.0816	5.0717
200	90	10	4.8483	4.8569	4.8032	4.8737	4.824
200	90	30	4.9578	4.9437	5.0115	5.0297	4.9439
200	100	10	4.8827	4.8819	4.9251	4.9842	4.848
200	100	20	5.1727	5.161	5.2118	5.1012	5.2128

Table 6.3: Expected value of mean squared error of estimators with $\sigma_X^2 = 0.6, \sigma_\epsilon^2 = 0.4$ (All MSEs are expressed in the order of 10^{-3}).

n	p	q	$MSE(\hat{\bar{Y}}_1)$	$MSE(\hat{\bar{Y}}_2)$	$MSE(\hat{\bar{Y}}_3)$	$MSE(\hat{\bar{Y}}_4)$	$MSE(\hat{\bar{Y}}_5)$
20	2	2	19.5	18.9	1.1	3	26.1
20	2	5	118.7	80.9	27	78.5	144.1
20	2	10	996.8	63.4	325.6	805.2	491.9
20	5	2	1790.6	1947	62.6	80.6	2302
20	5	5	380.2	171.9	392.3	858.4	277.4
20	10	2	843.5	734.3	182.8	35.2	1148.3
40	2	2	34.7	31.9	21.8	34.3	35.1
40	2	4	26.8	26.7	15.9	0.4	29.5
40	2	8	18.5	4.6	25.8	6.1	13.8
40	4	2	5.8	5.7	2.2	100.8	4.2
40	4	4	4.2	3.1	0.1	24.8	6.7
40	4	8	19.4	34.1	23.4	0.3	23.7
40	8	2	36.2	32	0.1	35.8	30.6
40	8	4	18.4	19.9	23.7	1.6	22.7
40	8	8	42.1	65.4	3.5	5.7	58.5
40	10	10	92.9	27.5	79.7	145.4	79.5
40	10	15	170.7	17.9	8.4	52.9	215.5
40	10	20	305.4	4.3	404	1289	18.4
40	15	10	146.1	1	1148.2	1830.4	2.2
40	20	10	258.1	635.8	0.9	151.6	1888
100	5	5	109.3	2.8	435.7	1588.7	0.9
100	10	10	2.6	2.6	1.6	0.1	2.8
100	10	20	23.2	20.7	8.3	12.2	28.1
100	10	30	19.7	10.6	9.1	1.5	21.7
100	10	40	40.8	22.2	19.6	73.5	41
100	10	50	11	1.4	12.3	30	3.2
100	20	10	141	14.1	36.6	1.8	107.9
100	20	20	9.6	7.5	6.8	3.4	14.5
100	20	30	57.6	27.7	35.1	8.8	65.5
100	20	40	80.4	38.4	17.6	0.1	115.6
100	30	10	4.4	0.4	1.9	0.1	3.6
100	30	20	3.4	3.1	23.5	10.7	3.2
100	30	30	15.2	7.9	37.6	56.5	9
100	30	30	31.6	32.9	13.6	0.3	53.3
100	40	10	90.7	73.4	38.5	26.4	106.2
100	40	20	48.7	32.3	12.2	11.4	64.5
100	50	10	9	5.7	10.8	1.9	17.1
200	10	10	1.3	1.1	0.2	7.2	1.3
200	10	20	0.9	0.7	0.7	9.5	0.8
200	10	40	1.9	2.7	0.4	2.5	2.6
200	10	60	2.6	0.4	1.7	3	1.8
200	10	80	22.7	9.8	11.4	0.5	22.4
200	10	100	26	19.5	18.8	0.1	27.9
200	10	120	0.5	1.9	1.4	0.1	1.1
200	30	10	0.7	0.8	1.2	0.4	0.7
200	30	30	6.3	5.3	2.6	2	8.6
200	30	50	23.8	16.6	17.6	1.7	26.8
200	30	70	2.1	0.8	0.1	15.1	2.5
200	30	90	12.3	1.7	4.6	0.1	10
200	50	10	0.3	0.1	0.1	0.3	0.3
200	50	30	8.6	5.9	12.9	5.9	8.5
200	50	50	0.1	0.7	0.7	1.6	0.3
200	50	70	26.3	19.9	7.1	84.5	60.3
200	70	10	7.5	8	0.6	2.7	6.8
200	70	30	3	5	0.2	0.1	4.1
200	70	50	10.4	8.5	17.1	4.8	11
200	90	10	0.8	0.8	0.1	1.9	2.3
200	90	30	1.7	1.1	0.1	0.1	2.5
200	100	10	3.4	3.2	4.9	6.6	2.7
200	100	20	25.8	24.1	13.6	11.2	29.1

Table 6.4: Expected value of mean squared error of estimators with $\sigma_X^2 = 1, \sigma_\epsilon^2 = 1$ (All MSEs are expressed in the order of 10^{-3}).

n	p	q	$MSE(\hat{\bar{Y}}_1)$	$MSE(\hat{\bar{Y}}_2)$	$MSE(\hat{\bar{Y}}_3)$	$MSE(\hat{\bar{Y}}_4)$	$MSE(\hat{\bar{Y}}_5)$
20	2	2	460	338	248.9	21.7	502.5
20	2	5	227.9	264	109.3	0.7	309.4
20	2	10	2213	785.2	391	2356.8	2993.6
20	5	2	117.2	154	9.8	1.9	141
20	5	5	4.5	57.1	4.8	146.4	24.3
20	10	2	1234.2	1132.6	553.1	286.9	1616.5
40	2	2	9.5	12.5	13.2	0.5	10
40	2	4	66.2	48.6	18.9	70.7	69.7
40	2	8	37.2	24.5	47.4	17.3	32.3
40	4	2	7.2	5	2.3	0.4	7.7
40	4	4	70.1	63.9	63.4	1.2	80.3
40	4	8	80.9	108.7	30.5	29.4	117.8
40	8	2	23.5	24.9	6.7	9.5	23.9
40	8	4	60.4	58.9	69	3.3	84.7
40	8	8	44.9	14.8	45.6	12.8	46.6
40	10	10	1502.5	892.7	726.2	495.1	1715.7
40	10	15	119.2	55	0.2	7.3	237.7
40	10	20	164.1	241.8	5.9	93.4	873.3
40	15	10	685.9	256.7	299.5	37	839.7
40	20	10	732.9	184.8	831.9	1405.3	606.5
100	5	5	1.4	1.5	0.3	0.1	1.5
100	10	10	32.3	29.6	13.8	3.1	35.7
100	10	20	1.6	1.4	3	5.4	1.2
100	10	30	83.8	71.3	65.4	10.5	92
100	10	40	24.5	17.9	19.1	37.2	22.8
100	10	50	176.7	73.5	47.3	5.6	244
100	20	10	8.4	8.8	1.3	3	11.6
100	20	20	1.7	0.4	0.2	11.9	1.1
100	20	30	0.3	15.1	0.6	7.5	1.3
100	20	40	17	0.2	0.4	48.9	22.9
100	30	10	52.8	50.6	30.1	10.8	81.9
100	30	20	107.4	79.8	39.2	0.1	149.4
100	30	30	4.7	11	18.1	48.8	12.3
100	30	30	3.8	17.5	3.1	0.6	13.3
100	40	10	53	48.1	108.9	14.7	91.4
100	40	20	19.7	8.3	4.3	0.1	29
100	50	10	9.3	13.7	5.2	1	5.5
200	10	10	4	3.5	2	14.6	3.9
200	10	20	22.3	23.9	20.8	15.9	22.7
200	10	40	1	0.3	0.1	18	1.6
200	10	60	17.7	18.1	17	0.6	19.4
200	10	80	2.1	0.1	0.3	54.5	4.7
200	10	100	5.7	0.1	1	2.1	2.1
200	10	120	0.1	6.3	8.1	12.6	0.3
200	30	10	2.1	2.5	0.2	0.1	2.3
200	30	30	3.8	9.7	6.9	1	4.9
200	30	50	2.3	0.8	0.6	0.1	2.9
200	30	70	29.3	28.2	32.7	11.9	37.9
200	30	90	21.2	34.6	44.3	0.2	20
200	50	10	22.2	20.5	11.9	9.5	23.6
200	50	30	0.5	0.8	2.3	0.2	0.7
200	50	50	8.3	5.7	0.5	39.2	23.6
200	50	70	23.5	28.3	12.1	10.7	33.4
200	70	10	21.9	20.4	14	47.4	15.3
200	70	30	30.3	34.5	14.1	9.5	54.8
200	70	50	5.2	3.2	0.4	6.7	5.1
200	90	10	23	20.5	38.7	15.9	31
200	90	30	1.8	3.2	0.1	0.9	3.1
200	100	10	13.8	13.9	5.6	0.3	23.1
200	100	20	29.8	25.9	44.9	10.2	45.3

Appendix

Let us write

$$\begin{aligned}
u_x &= \left(\frac{\bar{x}-\bar{X}}{\bar{X}}\right) \\
u_y &= \left(\frac{\bar{y}-\bar{Y}}{\bar{Y}}\right) \\
u_x^* &= \left(\frac{\bar{x}^*-\bar{Y}}{\bar{Y}}\right) \\
u_y^{**} &= \left(\frac{\bar{y}^{**}-\bar{Y}}{\bar{Y}}\right) \\
v_x &= \left(\frac{s_{xx}-\bar{X}^2\theta^2C_Y^2}{\bar{X}^2\theta^2C_Y^2}\right) \\
v_y &= \left(\frac{s_{yy}-\bar{Y}^2C_Y^2}{\bar{Y}^2C_Y^2}\right) \\
w &= \left(\frac{s_{xy}-\bar{X}\bar{Y}\rho\theta C_Y^2}{\bar{X}\bar{Y}\rho\theta C_Y^2}\right)
\end{aligned}$$

Using these notations, we can express

$$\begin{aligned}
\left(\frac{\hat{\bar{Y}}_1-\bar{Y}}{\bar{Y}}\right) &= U_y - \frac{\rho}{n\theta}[(n-p)u_x + pu_x^*](1+w)(1+v_x)^{-1} \\
&= U_y - \frac{\rho}{n\theta}[(n-p)u_x + pu_x^*](1+w)(1-v_x+\ldots) \\
&= \left(U_y - \frac{\rho}{n\theta}[(n-p)u_x + pu_x^*]\right) \\
&\quad + \frac{\rho}{n\theta}[(n-p)u_x + pu_x^*] * (v_x - w) + \ldots \qquad \text{(A.1)}
\end{aligned}$$

where

$$U_y = \frac{1}{n}[(n-q)u_y + qu_y^{**}]. \qquad \text{(A.2)}$$

Thus the relative bias of $\hat{\bar{Y}}_1$ to the first order of approximation is

$$\begin{aligned}
RB(\hat{\bar{Y}}_1) &= E\left(\frac{\hat{\bar{Y}}_1-\bar{Y}}{\bar{Y}}\right) \\
&= E(U_y) - \frac{\rho}{n\theta}E[(n-p)u_x + pu_x^*] \\
&\quad + \frac{\rho}{n\theta}E[(n-p)(u_xv_x - u_xw) + pu_x^*(v_x - w)] \\
&= KE_{p,q}\left[\frac{(n-p)}{n(n-p-q)}\right] \qquad \text{(A.3)}
\end{aligned}$$

where use has been made of the following results:

$$\begin{aligned} E(u_y \mid p,q) &= E(u_y^{**} \mid p,q) = E(u_x \mid p,q) = E(u_x^{*} \mid p,q) = 0 \\ E(u_x v_x \mid p,q) &= \frac{1}{(n-p-q)\theta^2 C_Y^2 N} \sum^{N} \left(\frac{X_i - \bar{X}}{\bar{X}} \right)^3 \\ E(u_x w \mid p,q) &= \frac{1}{(n-p-q)\rho\theta C_Y^2 N} \sum^{N} \left(\frac{X_i - \bar{X}}{\bar{X}} \right)^2 \left(\frac{Y_i - \bar{Y}}{\bar{Y}} \right) \end{aligned} \tag{A.4}$$

to the order of our approximation.

Similarly, utilizing the results

$$\begin{aligned} E(u_y^2 \mid p,q) &= \frac{C_Y^2}{(n-p-q)} \\ E(u_y^{**2} \mid p,q) &= \frac{C_Y^2}{q} \\ E(u_x^2 \mid p,q) &= \frac{\theta^2 C_Y^2}{(n-p-q)} \\ E(u_y^{*2} \mid p,q) &= \frac{\theta^2 C_Y^2}{p} \\ E(u_x u_y \mid p,q) &= \frac{\rho\theta C_Y^2}{(n-p-q)} \end{aligned} \tag{A.5}$$

we have

$$\begin{aligned} RMSE(\hat{\bar{Y}}_1) &= E\left(\frac{\hat{\bar{Y}}_1 - \bar{Y}}{\bar{Y}} \right)^2 \\ &= E\left(U_y^2 - \frac{\rho}{n\theta}[(n-p)u_x + pu_x^*] \right)^2 \end{aligned} \tag{A.6}$$

yielding the result (3.15) of Theorem II.

In a similar manner, we see that

$$\begin{aligned} \left(\frac{\hat{\bar{Y}}_2 - \bar{Y}}{\bar{Y}} \right) &= \left(U_y - \frac{\rho}{n\theta}[(n-p-q)u_x + pu_x^*] \right) \\ &\quad + \frac{\rho}{n\theta}[(n-p-q)u_x + pu_x^*](v_x - w) + \ldots \end{aligned} \tag{A.7}$$

$$\left(\frac{\hat{\bar{Y}}_3 - \bar{Y}}{\bar{Y}} \right) = \left(U_y - \frac{(n-q)\rho}{n\theta} u_x \right) + \frac{(n-q)\rho}{n\theta} u_x(v_x - w) + \ldots \tag{A.8}$$

Using these, the expressions (3.3) and (3.4) of Theorem I and (3.15) and (3.17) of Theorem II can be easily obtained.

Likewise, for the estimator $\hat{\bar{Y}}_4$, we have

$$\begin{aligned} \left(\frac{\hat{\bar{Y}}_4 - \bar{Y}}{\bar{Y}} \right) &= \left[U_y - \frac{\rho}{\theta} u_x + \frac{qp^2}{n}(u_y - u_y^{**}) \right] + \frac{\rho}{\theta} u_x(w - v_x) \\ &\quad - \frac{qp^2}{n}(u_y - u_y^{*})(v_x + v_y - 2w) + \ldots \end{aligned} \tag{A.9}$$

Employing the results (A.4) and (A.5) along with

$$\begin{aligned} E(u_y v_x \mid p,q) &= \frac{1}{(n-p-q)\theta^2 C_Y^2 N}\sum^{N}\left(\frac{X_i-\bar{X}}{\bar{X}}\right)^2\left(\frac{Y_i-\bar{Y}}{\bar{Y}}\right) \\ E(u_y v_y \mid p,q) &= \frac{1}{(n-p-q)C_Y^2 N}\sum^{N}\left(\frac{Y_i-\bar{Y}}{\bar{Y}}\right)^3 \qquad \text{(A.10)} \\ E(u_y w \mid p,q) &= \frac{1}{(n-p-q)\rho\theta C_Y^2 N}\sum^{N}\left(\frac{X_i-\bar{X}}{\bar{X}}\right)\left(\frac{Y_i-\bar{Y}}{\bar{Y}}\right)^2 \end{aligned}$$

to the given order of approximation, we find from (A.9) the expression (3.5) of Theorem I and (3.18) of Theorem II.

Lastly, from (2.9), we can write

$$\begin{aligned} \left(\frac{\tilde{\bar{Y}}-\bar{Y}}{\bar{Y}}\right) &= \frac{(n-p-q)u_y + q\bar{y}^{**}}{(n-p)} - \frac{\rho}{\theta(n-q)}[(n-p-q)u_x + pu_x^*] \\ &\quad + \frac{\rho}{\theta(n-q)}[(n-p-q)u_x + pu_x^*](v_x - w) + \ldots \qquad \text{(A.11)} \end{aligned}$$

whence the results stated in Theorem III can be easily found.