STATISTICA, anno LXXIX, n. 4, 2019

AN EFFICIENT ESTIMATION PROCEDURE FOR THE POPULATION MEAN UNDER NON-RESPONSE

Shashi Bhushan
Department of Mathematics and Statistics, Dr Shakuntala Misra National Rehabilitation University, Lucknow, India

Abhay Pratap Pandey [1]
Department of Statistics, Ramanujan College University of Delhi, New Delhi, India

1. Introduction

Usually almost all surveys covering human populations suffer from the problem of non-response. Lack of information, absence at the time of survey, and refusal of the respondents are main reason of the non-response. Non-respondents differ significantly from the respondents. Hansen and Hurwitz (1946) suggested a procedure of taking a sample from the non-respondent and collecting information by a more expensive method like first attempt by mail questionnaire and second attempt by personal interview. It is well known that utilizes the auxiliary information to neutralize the effect of non-response for estimating the population mean. The auxiliary information closely related to the main characteristics plays a very important role in estimation of population characteristics. The parameters can be estimated more accurately by making use such information on auxiliary variable. The ratio, product and regression methods of estimation and their generalizations are good examples in this context. Using the procedure envisaged by Hansen and Hurwitz (1946) various authors including Singh (1965, 1967), Shukla (1966), Ray and Singh (1981), Srivastava and Jhajj (1981), Diana and Tomasi (2003), Singh *et al.* (1994), Bahl and Tuteja (1991), Bhushan and Gupta (2015), Cochran (1977), Rao (1983), Khare and Srivastava (1995, 1997), Okafor and Lee (2000), Singh *et al.* (2010), Singh and Kumar (2008) have suggested improvement in the estimation procedure for population mean in presence of non-response.

An important finding of all these papers was that the difference or the corresponding regression type estimators were found to be best in terms of MSE and any ratio type estimator can at best attains the MSE of regression (difference) estimator. In this paper,

[1] Corresponding Author. E-mail: abhay.pandey@ramanujan.du.ac.in

we have proposed some improvement over these estimators proposed by various authors in their earlier works.

This paper is arranged as follows: Section 2 provides a brief review of some existing estimators along with some other important results; in Section 3, we propose seven new improved estimators using auxiliary variable and the results are reported; Section 4 deals with a comparative study of the proposed estimators using auxiliary variable in comparison to conventional estimators includes an empirical study and in Section 5, some concluding remarks are made.

2. NOTATIONS AND EXISTING ESTIMATORS

Consider a finite population mean $U = (U_1, U_2, ..., U_N$ of N identifiable units in the sense that they are uniquely labeled from 1 to N and the label on each unit is known. Let (y, x) be the study and the auxiliary variables taking values (yi, xi) on the i-th population units U_i, $i = 1, 2, ..., N$. Let n be the size of a sample drawn from the population of size N by using simple random sampling without replacement (SRSWOR), where n_1 of n units respond and $n_2(= n - n_1)$ sample units do not respond. From the n_2 non-response units, $r(n_2/k, k > 1)$ units are selected by making extra effort and thus giving $n_1 + r$ observations of the study variable y in place of n. In this procedure, the population Ω of size of N is supposed to be composed of two strata, namely respondents stratum S_1 and non-respondents stratum S_2 such that $\Omega = S_1 \cup S_2$; having sizes $N_1$and $N_2(= N - N_1)$ respectively. Without loss of generality, we label the data on study variable as $\{y_i, i \in S_1\}$for the response stratum, and as $\{y_i, i \in S_2\}$ for the non-response stratum.

Let $\bar{Y} = \sum_{\Omega} y_i/N$ and $S_y^2 = \sum_{\Omega} \left(y_i - \bar{Y}\right)^2 / (N-1)$ denote the population mean and population variance, respectively. Let $\bar{Y}_1 = \sum_{S_1} y_i/N_1$ and $S_{y_1}^2 = \sum_{S_1} \left(y_i - \bar{Y}_1\right)^2 / (N_1 - 1)$ denote the mean and variance of the response stratum, respectively. Similarly, let $\bar{Y}_2 = \sum_{S_2} y_i/N_2$ and $S_{y_2}^2 = \sum_{S_2} \left(y_i - \bar{Y}_2\right)^2 / (N_2 - 1)$ denote the mean and variance of the non-response stratum respectively. The population mean can be written as $\bar{Y} = W_1\bar{Y}_1 + W_2\bar{Y}_2$, $W_1 = N_1/N$ and $W_2 = N_2/N$. The sample mean $\bar{y}_1 = \sum_{i=1}^{n_1} y_i/n_1$ is unbiased for $\bar{Y}_1$ and hence biased for $\bar{Y}$ having bias $W_1\left(\bar{Y}_1 - \bar{Y}_2\right)$.

The Hansen and Hurwitz (1946) procedure is actually double sampling when strata sizes are not known (see Lohr, 1999) by effectively using the sample mean $\bar{y}_{2r} = \sum_{i=1}^{r} y_i/r$ which is unbiased for the mean $\bar{y}_2$ of the n_2 units resulting in an unbiased estimator for the population mean $\bar{Y}$ given by

$$\bar{y}^* = w_1\bar{y}_1 + w_2\bar{y}_{2r}, \tag{1}$$

where $w_1 = n_1/n$ and $w_2 = n_2/n$.The variance of Hansen and Hurwitz (HH) mean $\bar{y}^*$ is given by

$$\text{Var}(\bar{y}^*) = \left(\frac{1-f}{n}\right) S_y^2 + \frac{W_2(k-1)}{n} S_{y_2}^2, \tag{2}$$

where $f = n/N$ & $k = n_2/r$ are sampling fraction and inverse sub-sampling fraction respectively.

Let $x_i (i = 1, 2, ..., N)$ denote an auxiliary variate correlated with study variate $y_i (i = 1, 2, ..., N)$. The population mean of the auxiliary variable x is $\bar{X} = \sum_{i=1}^{N} x_i/N$. Let $\bar{X}_1$ and $\bar{X}_2$ denote the means of the response and non-response groups. Let $\bar{x}$ denote the mean of all the n units. Let $\bar{x}_1$ and $\bar{x}_2$ denote the means of the n_1 responding units and the n_2 non-responding units. Further let $\bar{x}_{2r} = \sum_{i=1}^{r} x_i/r$ denote the mean of the sub-sampled units. The population variances of x and y are denoted by S_x^2 and S_y^2, and the population covariance by S_{xy}. The population correlation coefficient is $\rho = S_{xy}/S_x^2\, S_y^2$. The unbiased estimator of the population mean $\bar{X}$ of the auxiliary variable x is

$$\bar{x}^* = w_1 \bar{x}_1 + w_2 \bar{x}_2. \tag{3}$$

The variance of $\bar{x}^*$ is given by

$$\text{Var}(\bar{x}^*) = P S_x^2 + Q S_{x_2}^2, \tag{4}$$

where $S_{x_2}^2 = \sum_{i=N_1+1}^{N_1+N_2} \left(x_i - \bar{X}_2\right)^2 / (N_2 - 1)$.

Taking inspiration from some important works of Khare and Srivastava (1993, 1995, 1997), Okafor and Lee (2000), Singh and Kumar (2008, 2010), Cochran (1977), and Rao (1986) we have considered some regression (difference) type estimators under non-response. Due to paucity of space, we have not considered ratio, product or ratio product type estimators and have considered only the regression (difference) estimators as they are the BLUE within each category. Also, for better understanding, we have divided them under four different strategies given below.

- **Strategy I**: $\bar{y}^*$, $\bar{x}^*$ and $\bar{X}$ are used. The non-response occurs on both the study variable y and auxiliary variable x, and the population mean $\bar{X}$ of the auxiliary variable is known. The difference and regression type estimators are

$$t_1 = \bar{y}^* + K_1 \left(\bar{X} - \bar{x}^*\right) \tag{5}$$

$$t_{lr_1} = \bar{y}^* + b^* \left(\bar{X} - \bar{x}^*\right). \tag{6}$$

- **Strategy II**: $\bar{y}^*$, $\bar{x}$ and $\bar{X}$ are used. The non-response occurs on the study variable y, and information on the auxiliary variable x is available from all the sample units along the population mean $\bar{X}$ of the auxiliary variable is known. The difference and regression type estimators are

$$t_2 = \bar{y}^* + K_2 \left(\bar{X} - \bar{x}\right) \tag{7}$$

$$t_{lr_2} = \bar{y}^* + b \left(\bar{X} - \bar{x}\right). \tag{8}$$

- **Strategy III**: $\bar{y}^*$, $\bar{x}^*$and $\bar{x}$ are used. The non-response occurs on the study variable y, and the information on the auxiliary variable x is obtained from all the sample units, but the population mean $\bar{X}$ of the auxiliary variable is not known. The difference and regression type estimators are

$$t_3 = \bar{y}^* + K_3(\bar{x} - \bar{x}^*). \tag{9}$$

$$t_{lr_3} = \bar{y}^* + b_{2r}(\bar{x} - \bar{x}^*). \tag{10}$$

- **Strategy IV**: $\bar{y}^*$, $\bar{x}$, $\bar{x}^*$ and $\bar{X}$ are used. The difference and regression type estimators are

$$t_4 = \bar{y}^* + d_1(\bar{x} - \bar{x}^*) + d_2(\bar{X} - \bar{x}). \tag{11}$$

The above mentioned estimators under strategy I, II and III were improved by Singh and Kumar (2008, 2010) by incorporating all possible auxiliary information that might be available at the disposal of a survey statistician and proposed t_4 under strategy IV. Singh and Kumar (2008, 2010) proved that their estimator minimized both overall variance component as well as the non-response variances component in (2), as evident by (13). It is important to note that t_4 is a generalization over t_3 and t_2, given in Equations (7) and (9) for $d_2 = 0$ and $d_1 = 0$, respectively.

The variances of the above estimators, up to the first order of approximation, are given by

$$\text{Var}\left(t_{lr_1}\right) = \left[\frac{(1-f)}{n}S_y^2(1-\rho^2) + \frac{W_2(k-1)}{n}\left(S_{y_2}^2 + \beta^2 S_{x_2}^2 - 2\beta S_{xy2}\right)\right] \tag{12}$$

$$\text{min.MSE}(t_1) = \left\{\frac{(1-f)}{n}s_y^2 + \frac{W_2(k-1)}{n}S_{y_2}^2\right\} - \frac{\left\{\frac{(1-f)}{n}S_{xy} + \frac{W_2(k-1)}{n}S_{xy_2}\right\}^2}{\left\{\frac{(1-f)}{n}S_x^2 + \frac{W_2(k-1)}{n}S_{x_2}^2\right\}} \tag{13}$$

$$\text{Var}\left(t_{lr_2}\right) = \min.MSE(t_2) = \left[\frac{(1-f)}{n}s_y^2(1-\rho^2) + \frac{W_2(k-1)}{n}S_{y_2}^2\right] \tag{14}$$

$$\text{Var}\left(t_{lr_3}\right) = \min.MSE(t_3) = \left[\frac{(1-f)}{n}s_y^2 + \frac{W_2(k-1)}{n}S_{y_2}^2(1-\rho_2^2)\right] \tag{15}$$

$$\text{min.MSE}(t_4) = \left[\frac{(1-f)}{n}s_y^2(1-\rho^2) + \frac{W_2(k-1)}{n}S_{y_2}^2(1-\rho_2^2)\right]. \tag{16}$$

It is important to notice that MSE of any strategy under non-response can be split as $M = M_1 + M_2$, where M_1 is the over all sampling variance without non-response and M_2 is the contribution of sub-sampling due to non-response.

Also, if we observe the construction of t_2 which utilizes $\bar{y}^*$, $\bar{X}$ and zero function $\omega_1 = \bar{x}^* - \bar{X}$ (zero function is a function whose average value is zero) under optimum conditions minimises its M_1 component but not M_2. Now, if we observe the construction of t_3 which uses $\bar{y}^*$ and zero function $\omega_2 = \bar{x} - \bar{x}^*$ under optimum conditions minimises its M_2 component but not M_1. While, if we observe the construction of t_4 which uses both ω_1 and ω_2 under optimum conditions minimises both M_1 and M_2 components. This was the result due to chaining of ω_1 and ω_2 or chaining all the available information about x with $\bar{x}$ as the chaining statistic.

Also, Khare and Srivastava (1993, 1995, 1997) and Okafor and Lee (2000) proposed a double sampling scheme for ratio and regression estimation with sub sampling the non-respondent while dealing with non-response problem. Due to economy of space, we have considered only the difference and regression type estimators similar to the estimators defined in (6) to (11). Some generalisation of these estimators was proposed by Singh and Bhushan (2012). In this paper we have also considered as follow the double sampling estimators under four strategies. The double sampling strategies are given below.

- **Strategy V**: $\bar{y}^*$, $\bar{x}^*$ and $\bar{x}'$ are used. The difference and regression type estimators are

$$t_5 = \bar{y}^* + K_5\left(\bar{x}' - \bar{x}^*\right) \tag{17}$$

$$t_{lr_5} = \bar{y}^* + b^*\left(\bar{x}' - \bar{x}^*\right). \tag{18}$$

- **Strategy VI**: $\bar{y}^*$, $\bar{x}$ and and $\bar{x}'$ are used. The difference and regression type estimators are

$$t_6 = \bar{y}^* + K_6\left(\bar{x}' - \bar{x}\right) \tag{19}$$

$$t_{lr_6} = \bar{y}^* + b^{**}\left(\bar{x}' - \bar{x}\right). \tag{20}$$

  Singh and Kumar (2010) proposed estimators for the population mean $\bar{Y}$ by using double sampling scheme under non-response.

- **Strategy VII**: $\bar{y}^*$, $\bar{x}^*$, $\bar{x}$ and $\bar{x}'$ are used. The difference (regression) type estimators is given by

$$t_7 = \bar{y}^* + d_3(\bar{x} - \bar{x}^*) + d_4\left(\bar{x}' - \bar{x}\right). \tag{21}$$

Again, it is important to note that t_7 is an extension of t_4 under double sampling. Also, t_7 is a generalisation over t_6 and t_5 for $d_3 = 0$ and $d_4 = 0$ respectively. Further t_7 minimises both M_1 and M_2 components due to double sampling and sub-sampling of non-respondents.

The MSEs of the above estimators, up to the first order of approximation, are given by

$$\mathrm{Var}\left(t_{lr_5}\right) = \left[\left(\frac{1}{n'} - \frac{1}{N}\right)s_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right)s_y^2\left(1-\rho^2\right) + \frac{W_2(k-1)}{n}S^2_{\beta_2}\right], \tag{22}$$

where $S^2_{\beta_2} = S^2_{y_2} + \beta S^2_{x_2}(\beta - 2\beta_2)$.

$$\text{min.MSE}(t_5) = \left(\frac{1-f}{n}\right) S_y^2 + \frac{W_2(k-1)}{n} S^2_{y_2} - \frac{\left\{\left(\frac{1}{n}-\frac{1}{n'}\right) S_{xy} + \frac{W_2(k-1)}{n} S_{xy_2}\right\}^2}{\left\{\left(\frac{1}{n}-\frac{1}{n'}\right) S_x^2 + \frac{W_2(k-1)}{n} S^2_{x_2}\right\}} \quad (23)$$

$$\begin{aligned}\text{Var}\left(t_{lr_6}\right) &= \text{min.MSE}(t_6) = \left[\left(\frac{1}{n'} - \frac{1}{N}\right) s_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right) s_y^2 (1-\rho^2)\right. \\ &\quad \left. + \frac{W_2(k-1)}{n} S^2_{y_2}\right] \end{aligned} \quad (24)$$

$$\begin{aligned}\text{min.MSE}(t_7) &= \left[\left(\frac{1}{n'} - \frac{1}{N}\right) s_y^2 + \left(\frac{1}{n} - \frac{1}{n'}\right) s_y^2 (1-\rho^2) + \frac{W_2(k-1)}{n}\right. \\ &\quad \left. S^2_{y_2}(1-\rho_2^2)\right], \end{aligned} \quad (25)$$

where $b^* = \left(s^*_{xy}/s^{*2}_x\right)$, $b = \left(s^*_{xy}/s^2_x\right)$, $b_{2r} = \left(s_{xy(2r)}/s^2_{x(2r)}\right)$, $b^{**} = \left(s^*_{xy}/s^2_x\right)$,
$s^*_{xy} = \frac{1}{(n-1)}\left(\sum_{u_1} x_j y_j + r \sum_{u_{2m}} x_j y_j - n\bar{x}\bar{y}^*\right)$, $s^{*2}_x = \frac{1}{(n-1)}\left(\sum_{u_1} x_j^2 + r \sum_{u_{2m}} x_j^2 - n\bar{x}\bar{x}^*\right)$,
$s^2_x = \sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1)$, $s_{xy(2r)} = \sum_{i=1}^{r} (x_i - \bar{x}_{2r})(y_i - \bar{y}_{2r})/(r-1)$ and
$s^2_{x(2r)} = \sum_{i=1}^{r} (x_i - \bar{x}_{2r})^2/(r-1)$.

In this paper, we have proposed the number of improved regression (difference) estimators motivated by estimators suggested by Searls (1964), Cochran (1977), Rao (1986), Okafor and Lee (2000) and Singh and Kumar (2008) under single phase and two phase sampling classified under seven different strategies with deterministic non-response setup.

3. Proposed estimators

In this section, we propose to use Searls (1964) type transformation (STT) under different strategies. The STT can be defined as

$$T = \alpha \bar{y}^*, \quad (26)$$

where α is a suitable chosen scalar. The following estimators under the strategies described in the previous section are proposed using the STT.

The proposed estimator under *Strategy I*, when $\bar{y}^*$, $\bar{x}^*$ and $\bar{X}$ are used, is given by

$$T_1 = \alpha_1 \bar{y}^* + \beta_1\left(\bar{X} - \bar{x}^*\right). \quad (27)$$

The proposed estimator under *Strategy II*, when $\bar{y}^*$, $\bar{x}$ and $\bar{X}$ are used, is given by

$$T_2 = \alpha_2 \bar{y}^* + \beta_2 \left(\bar{X} - \bar{x}\right). \tag{28}$$

The proposed estimator under *Strategy III*, when $\bar{y}^*$, $\bar{x}$ and $\bar{x}^*$ are used, is given by

$$T_3 = \alpha_3 \bar{y}^* + \beta_3 \left(\bar{x}^* - \bar{x}\right). \tag{29}$$

Similarly, another proposed estimator obtained by chaining all the available auxiliary information. The proposed estimator under *Strategy IV*, when $\bar{y}^*$, $\bar{x}$, $\bar{x}^*$ and $\bar{X}$ are used, is given by

$$T_4 = \alpha_4 \bar{y}^* + \beta_4 \left(\bar{x}^* - \bar{x}\right) + \gamma_4 \left(\bar{X} - \bar{x}\right). \tag{30}$$

Also, under the two phase sampling scheme, we propose the following improved difference estimators classified under different strategies. The proposed estimator under *Strategy V*, when $\bar{y}^*$, $\bar{x}^*$ and $\bar{x}'$ are used, is given by

$$T_5 = \alpha_5 \bar{y}^* + \beta_5 \left(\bar{x}' - \bar{x}^*\right). \tag{31}$$

The proposed estimator under *Strategy VI*, when $\bar{y}^*$, $\bar{x}$ and $\bar{x}'$ are uses, is given by

$$T_6 = \alpha_6 \bar{y}^* + \beta_6 \left(\bar{x}' - \bar{x}\right). \tag{32}$$

The proposed estimator under *Strategy VII*, when $\bar{y}^*$, $\bar{x}$, $\bar{x}^*$ and $\bar{x}'$ are used, is given by

$$T_7 = \alpha_7 \bar{y}^* + \beta_7 \left(\bar{x} - \bar{x}^*\right) + \gamma_7 \left(\bar{x}' - \bar{x}\right), \tag{33}$$

where α_i, β_i and γ_j are suitably chosen constant.

It is important to note that all these estimators T_i are generalizations over t_i and we get t_i if we put $\alpha_i = 1$ in $T_i\,(i = 1, 2, ...7)$. In this study, we have focussed our attention over improvement in conventional difference and regression type estimators, which are the BLUE.

THEOREM 1. *The bias and minimum MSE of the STD estimator T_i $(i = 1, 2, ..., 7)$ is given by*

$$Bias(T_i) = \bar{Y}(\alpha_i - 1). \tag{34}$$

and

$$\min MSE_{\alpha_i}(T_i) = \frac{\bar{Y}^2 MSE(t_i)}{\bar{Y}^2 + MSE(t_i)}. \tag{35}$$

where $MSE_{\alpha_i}(T_i)$ is the first order MSE with parameter α_i and $MSE(t_i)$ is the first order MSE of the t_i or T_i when $\alpha_i = 1$.

PROOF. The outline of derivation is given in Appendix A along with the optimum values of different scalars. □

THEOREM 2. *Under the optimum values of scalars, the STD estimators T_i have always lesser MSE than the conventional estimators t_i, $i = 1,2,...,7$. Alternatively,*

$$\min MSE(T_i) < \min MSE(t_i) \ (i = 1,2,...,7).$$

PROOF. Using (35), proof is obvious. □

THEOREM 3. *T_4 provides minimum MSE in comparison to single phase estimators T_1, T_2 and T_3 under optimal conditions.*

PROOF. Since min.MSE(T_i) can be written as

$$\text{min.MSE}(T_i) = \bar{Y}^2(1-\alpha_i), i = 1,2,3,4$$

On the basis optimal value of α_i, we observed that α_4 has greater value in comparison to α_1, α_2 and α_3. Therefore,

$$\text{min.MSE}(T_4) < \text{min.MSE}(T_i), i = 1,2,3$$

□

THEOREM 4. *T_7 provides minimum MSE in comparison to two phase estimators T_5 and T_6 under optimal conditions.*

PROOF. Since min.MSE(T_i) can be written as

$$\text{min.MSE}(T_i) = \bar{Y}^2(1-\alpha_i), i = 5,6,7$$

On the basis optimal value of α_i, we observe that α_7 has greater value in comparison to α_5 and α_6. Therefore,

$$\text{min.MSE}(T_7) < \text{min.MSE}(T_i), i = 5,6$$

□

Therefore, the estimators T_4 and T_7 provides the way to minimize the MSE and to improve the efficiency.

4. EMPIRICAL STUDY

In order to have a better understanding about the efficiency of the proposed estimators, we have conducted an empirical study on the data by Srivastava (1993, p. 50) and compared the proposed estimators with $\bar{y}_r$ and the results are reported.

A list of 70 villages in a Tehsil of India along with their population in 1981 and cultivated area (in acres) in the same year is taken in to consideration. Here the cultivated

TABLE 1
MSE and PRE of the existing estimators and proposed estimators for the first dataset.

Estimator	$k = 2$	$k = 3$	$k = 4$	$k = 5$
$\bar{y}^*$	10160.17(100)	10636.88(100)	11113.60(100)	11590.32(100)
T	10054.08(101.05)	10520.67(101.10)	10986.79(101.15)	11452.47(101.20)
Strategy I				
t_{lr_1}	4294.06(236.61)	4765.90(223.18)	5237.7533(212.18)	5709.60(202.10)
t_1	4288.77(236.90)	4745.94(224.12)	5195.20(213.92)	5637.74(205.58)
T_1	4269.75(**237.95**)	4722.66(**225.23**)	5167.32(**215.07**)	5604.92(**206.78**)
Strategy II				
$t_{lr_2} = t_2$	4298.93(236.34)	4775.64(222.73)	5252.36(211.60)	5729.08(202.30)
T_2	4279.82(**237.39**)	4752.08(**223.83**)	5223.87(**212.74**)	5695.19(**203.51**)
Strategy III				
$t_{lr_3} = t_3$	10065.76(100.93)	10448.08(101.80)	10830.39(102.61)	11212.71(103.36)
T_3	9961.63(**101.99**)	10335.93(**102.91**)	10709.93(**103.77**)	11083.65(**104.57**)
Strategy IV				
t_{lr_4}	4204.52(241.65)	4586.8435(231.99)	4969.1591(223.65)	5351.4746(216.58)
T_4	4186.25(**242.70**)	4565.09(**233.00**)	4961.16(**224.80**)	5321.89(**217.78**)
Strategy V				
t_{lr_5}	6736.24(150.82)	7208.08(147.57)	7679.93(144.71)	8151.78(142.18)
t_5	6727.54(151.02)	7176.38(148.22)	7614.40(145.95)	8044.04(144.08)
T_5	6680.87(**152.07**)	7123.29(**149.32**)	7554.66(**147.11**)	7977.40(**145.29**)
Strategy VI				
$t_{lr_6} = t_6$	6741.11(150.72)	7217.83(147.37)	7694.54(144.43)	8171.26(141.84)
T_6	6694.24(**151.77**)	7164.13(**148.47**)	7633.55(**145.58**)	8102.50(**143.04**)
Strategy VII				
t_7	6646.71(152.86)	7029.02(151.32)	7411.34(149.95)	7793.65(148.71)
T_7	6601.14(**153.91**)	6978.08(**152.43**)	7354.73(**151.10**)	7731.08(**149.91**)

TABLE 2
MSE and PRE of the existing estimators and proposed estimators for the second dataset.

Estimator	$k=2$	$k=3$	$k=4$	$k=5$
$\bar{y}^*$	0.2067(100)	0.2464(100)	0.2860(100)	0.3256(100)
T	0.2066(100.054)	0.2462(100.064)	0.2857(100.075)	0.3253(100.085)
Strategy I				
t_1	0.0664(311.050))	0.0853(288.828)	0.1040(274.838)	0.1227(265.221)
t_{lr_1}	0.0665(310.625)	0.0856(287.794)	0.1046(273.273)	0.1237(263.223)
T_1	0.0664(**311.104**)	0.0852(**288.893**)	0.1040(**274.913**)	0.1227(**265.307**)
Strategy II				
$t_{lr_2}=t_2$	0.0871(237.288)	0.1267(194.376)	0.1663(171.902)	0.2060(158.072)
T_2	0.0871(**237.342**)	0.1267(**194.441**)	0.1663(**171.977**)	0.2058(**158.158**)
Strategy III				
$t_{lr_3}=t_3$	0.1857(111.337)	0.2042(120.615))	0.2228(128.346)	0.2414(134.888))
T_3	0.1856(**111.392**)	0.2041(**120.680**)	0.2227(**128.421**)	0.2412(**134.974**)
Strategy IV				
t_{lr_4}	0.0660(312.896)	0.0846(291.079)	0.1032(277.110)	0.1217(267.401)
T_4	0.0660(**312.950**)	0.0846(**291.144**)	0.1031(**277.186**)	0.1217(**267.486**)
Strategy V				
t_{lr_5}	0.0914(225.982)	0.1105(222.865)	0.1296(220.665)	0.1486(219.029)
t_5	0.0913(226.276)	0.1101(223.627)	0.1289(221.877)	0.1475(220.636)
T_5	0.0913(**226.331**)	0.1101(**223.692**)	0.1288(**221.952**)	0.1475(**220.722**)
Strategy VI				
$t_{lr_6}=t_6$	0.1120(184.514)	0.1516(162.438)	0.1913(149.506)	0.2309(141.012)
T_6	0.1120(**184.568**)	0.1516(**162.503**)	0.1912(**149.581**)	0.2307(**141.098**)
Strategy VII				
t_7	0.0910(227.205)	0.1095(224.869)	0.1281(223.210)	0.1467(221.971)
T_7	0.0909(**227.259**)	0.1095(**224.934**)	0.1280(**223.286**)	0.1466(**222.057**)

area (in acres) is taken as main study character and the population of village is taken as auxiliary character. The parameters of the population are as follows.
$N = 70$, $n' = 40$, $n = 25$, $\bar{Y} = 981.29$, $\bar{X} = 1755.53$, $S_y = 613.66$, $S_x = 1406.13$, $\bar{Y}_2 = 597.29$, $\bar{X}_2 = 1100.24$, $S_{y_2} = 244.11$, $S_{x_2} = 631.51$, $\rho = 0.778$, $\rho_2 = 0.445$, $R = 0.5589$, $\beta = 0.3395$, $\beta_2 = 0.1720$, $W_2 = 0.20$.

We have conducted an empirical study on the data by Khare and Sinha (2004, p. 53). The data belongs to the data on physical growth of upper-socio-economic group of 95 school children of Varanasi under an ICMR study. The first 25 (i.e. 24 children) units have been considered as non-response units. The values of the parameters related to the study variate y (the weight in kg) and the auxiliary variate x (the chest circumference in cm) have been given below:
$N = 95$, $n' = 70$, $n = 35$, $\bar{Y} = 19.497$, $\bar{X} = 55.8611$, $S_y = 3.0435$, $S_x = 3.2735$, $S_{y_2} = 2.3552$, $S_{x_2} = 2.5137$, $\rho = 0.8460$, $\rho_2 = 0.7290$, $R = 0.3490$, $\beta = 0.7865$, $\beta_2 = 0.6829$, $W_2 = 0.25$, $N_2 = 24$, $N_1 = 71$.

The MSE and percent relative efficiency (PRE) of the estimators with respect to $\bar{y}_r$ at different values of k are given in Tables 1 and 2.

It can be easily seen that the proposed estimators are better in comparison than the conventional estimators. From perusal of above results, it is observed that the STD estimators T_i are always better than the respective conventional difference or regression type counterparts t_i, $i = 1, 2, ..., 7$ both under single phase and two phase sampling. Hence, we conclude that all the proposed estimators have higher efficiency in comparison to the conventional regression (BLUE) estimators. A comparison of STD (regression) estimators T_i $(i = 1, 2, ..., 7)$ within themselves and with t_i $(i = 1, 2, 3, 5, 6, 7)$ shows that STD estimator T_4 is superior under single phase sampling and T_7 is superior under two phase sampling.

5. Simulation study

In this section, simulation is conducted to evaluate the performance of the proposed class of estimators with respect to traditional estimators. For this study we have generated a population size $N = 1,000$ from standard normal distribution using the MVRNORM package in software R, where study and auxiliary variable are correlated with correlation $\rho = 0.7$, draw a large sample of size $n' = 400$ from the population then again select a sample of size $n = 200$ from n' with 35% non-response. The whole simulation process starting from the drawing sample from variable Y and auxiliary variable X from normal population and calculating the estimates was repeated 50,000 times.

It can be easily seen that in simulation study the proposed estimators are better in comparison than the conventional estimators. A comparison of STD (regression) estimators T_i $(i = 1, 2, ..., 7)$ within themselves and with t_i $(i = 1, 2, 3, 5, 6, 7)$ shows that STD estimator T_4 is superior under single phase sampling and T_7 is superior under two phase sampling.

TABLE 3
Percentage relative efficiency (PRE) of the proposed estimators with respect to $\bar{y}^*$ using simulation.

	Estimator	PRE
	$\bar{y}^*$	100
	T	101.20
Strategy I	t_1	249.66
	t_{lr_1}	249.64
	T_1	250.85
Strategy II	$t_2 = t_{lr_2}$	129.31
	T_2	130.51
Strategy III	$t_3 = t_{lr_3}$	159.45
	T_3	160.64
Strategy IV	t_4	249.67
	T_4	**250.87**
Strategy V	t_5	205.96
	t_{lr_5}	205.94
	T_5	207.16
Strategy VI	$t_6 = t_{lr_6}$	116.50
	T_6	117.70
Strategy VII	t_7	205.97
	T_7	207.16

6. CONCLUSIONS

Diana and Perri (2013) in their study opined that the regression and difference type estimator provide best possible improvement and cannot be improved upon. In this paper, we have proposed various improved difference type estimators, which provide an improvement over the corresponding regression estimators in their respective strategies. This result is significant as it provides an improvement over BLUE. The results are proved both theoretically as well as empirically. The results of conventional estimators can be obtained as a special case of the respective proposed estimators by setting $\alpha_i = 1$, $i = 1, 2, ..., 7$.

ACKNOWLEDGEMENTS

The authors would like to thank the Editor-in-Chief, and the Referees for their careful reading and constructive comments and suggestions which greatly improved the presentation of the paper.

Appendix

A. Outline of the derivation of Theorem 1

The MSE of T_1 is given by

$$\text{MSE}(T_1)=(\alpha_1-1)^2\bar{Y}^2+\alpha_1^2\left(AS_y^2+BS_{y_2}^2\right)+\beta_1^2\left(AS_x^2+BS_{x_2}^2\right)-2\alpha\beta\left(AS_{xy}+BS_{xy_2}\right),$$

for optimum value of α_1 and β_1. Differentiating above equation partially with respect to α_1 and β_1, we get

$$\alpha_1=\frac{\bar{Y}^2}{\left\{\bar{Y}^2+AS_y^2+BS_{y_2}^2-\frac{\left(AS_{xy}+BS_{xy_2}\right)^2}{\left(AS_x^2+BS_{x_2}^2\right)}\right\}}$$

$$\beta_1=\frac{\bar{Y}^2\left(AS_{xy}+BS_{xy_2}\right)}{\left(AS_x^2+BS_{x_2}^2\right)\left\{\bar{Y}^2+AS_y^2+BS_{y_2}^2-\frac{\left(AS_{xy}+BS_{xy_2}\right)^2}{\left(AS_x^2+BS_{x_2}^2\right)}\right\}}.$$

After putting these values in $\text{MSE}(T_1)$, we get

$$\min.\text{MSE}(T_1)=\frac{\bar{Y}^2\left[\left\{AS_y^2+BS_{y_2}^2\right\}-\frac{\left\{AS_{xy}+BS_{xy_2}\right\}^2}{\left\{AS_x^2+BS_{x_2}^2\right\}}\right]}{\bar{Y}^2+\left[\left\{AS_y^2+BS_{y_2}^2\right\}-\frac{\left\{AS_{xy}+BS_{xy_2}\right\}^2}{\left\{AS_x^2+BS_{x_2}^2\right\}}\right]}=\frac{\bar{Y}^2\left[MSE\left(t_1\right)\right]}{\left[\bar{Y}^2+MSE\left(t_1\right)\right]},$$

where $A=\left(\frac{1}{n}-\frac{1}{N}\right)$ and $B=\frac{W_2(k-1)}{n}$.

Similarly, we get the min.MSE such that

$$\text{min.MSE}(T_i)=\frac{\bar{Y}^2\left[MSE\left(t_i\right)\right]}{\left[\bar{Y}^2+MSE\left(t_i\right)\right]},$$

for $i=1,2,...,7$.

The optimum values of scalars for different estimators involved are

$\alpha=\frac{\bar{Y}^2}{\bar{Y}^2+AS_y^2+BS_{y_2}^2}$

$\alpha_2=\frac{\bar{Y}^2}{\left\{\bar{Y}^2+AS_y^2(1-\rho^2)+BS_{y_2}^2\right\}}$, $\beta_2=\frac{\bar{Y}^2S_{xy}}{S_x^2\left\{\bar{Y}^2+AS_y^2(1-\rho^2)+BS_{y_2}^2\right\}}$

$$\alpha_3 = \frac{\bar{Y}^2}{\left\{\bar{Y}^2 + AS_y^2 + BS_{y_2}^2(1-\rho_2^2)\right\}},\ \beta_3 = \frac{\bar{Y}^2 S_{xy_2}}{S_{x_2}^2\left\{\bar{Y}^2 + AS_y^2 + BS_{y_2}^2(1-\rho_2^2)\right\}}$$

$$\alpha_4 = \frac{\bar{Y}^2}{\left\{\bar{Y}^2 + AS_y^2(1-\rho^2) + BS_{y_2}^2(1-\rho_2^2)\right\}},\ \beta_4 = \frac{\bar{Y}^2 S_{xy_2}}{S_{x_2}^2\left\{\bar{Y}^2 + AS_y^2(1-\rho^2) + BS_{y_2}^2(1-\rho_2^2)\right\}}$$

$$\gamma_4 = \frac{\bar{Y}^2 S_{xy}}{S_x^2\left\{\bar{Y}^2 + AS_y^2(1-\rho^2) + BS_{y_2}^2(1-\rho_2^2)\right\}}$$

$$\alpha_5 = \frac{\bar{Y}^2}{\left\{\bar{Y}^2 + AS_y^2 + BS_{y_2}^2 - \frac{\left(CS_{xy}+BS_{xy_2}\right)^2}{\left(CS_x^2+BS_{x_2}^2\right)}\right\}}$$

$$\beta_5 = \frac{\bar{Y}^2\left(CS_{xy}+BS_{xy_2}\right)}{\left(CS_x^2+BS_{x_2}^2\right)\left\{\bar{Y}^2 + AS_y^2 + BS_{y_2}^2 - \frac{\left(CS_{xy}+BS_{xy_2}\right)^2}{\left(CS_x^2+BS_{x_2}^2\right)}\right\}}$$

$$\alpha_6 = \frac{\bar{Y}^2}{\left\{\bar{Y}^2 + DS_y^2 + CS_y^2(1-\rho^2) + BS_{y_2}^2\right\}},\ \beta_2 = \frac{\bar{Y}^2 S_{xy}}{S_x^2\left\{\bar{Y}^2 + DS_y^2 + CS_y^2(1-\rho^2) + BS_{y_2}^2\right\}}$$

$$\alpha_7 = \frac{\bar{Y}^2}{\left\{\bar{Y}^2 + DS_y^2 + CS_y^2(1-\rho^2) + BS_{y_2}^2(1-\rho_2^2)\right\}},\ \beta_7 = \frac{\bar{Y}^2 S_{xy_2}}{S_{x_2}^2\left\{\bar{Y}^2 + DS_y^2 + CS_y^2(1-\rho^2) + BS_{y_2}^2(1-\rho_2^2)\right\}}$$

$$\gamma_7 = \frac{\bar{Y}^2 S_{xy}}{S_x^2\left\{\bar{Y}^2 + DS_y^2 + CS_y^2(1-\rho^2) + BS_{y_2}^2(1-\rho_2^2)\right\}},$$

where $C = \left(\frac{1}{n} - \frac{1}{n'}\right)$, $D = \left(\frac{1}{n'} - \frac{1}{N}\right)$.

References

S. Bahl, R. K. Tuteja (1991). *Ratio and product type exponential estimators*. Information and Optimization Sciences, 12, no. 1, pp. 159–163.

S. Bhushan, R. Gupta (2015). *Some log type classes of estimators using auxiliary*. International Journal of Agricultural and Statistical Sciences, 2, no. 2, pp. 487–491.

W. G. Cochran (1977). *Sampling Techniques*. John Wiley & Sons, New York, 3rd ed.

G. Diana, P. F. Perri (2013). *A class of estimators in two-phase sampling with subsampling the non-respondents*. Applied Mathematics and Computation, 219, pp. 10033–10043.

G. Diana, C. Tomasi (2003). *Optimal estimation for finite population mean in two phase sampling*. Statistical Methods and Applications, 12, pp. 41–48.

M. H. Hansen, W. N. Hurwitz (1946). *The problem of non-response in sample surveys*. Journal of American Statistical Association, 41, pp. 517–529.

B. B. Khare, R. R. Sinha (2004). *Estimation of finite population ratio using two phase sampling scheme in the presence of non-response*. Aligarh Journal of Statistics, 24, pp. 43–56.

B. B. KHARE, S. SRIVASTAVA (1993). *Estimation of population mean uing auxiliary character in presence of non-response*. National Academy Science Letters - India, 16, pp. 111–114.

B. B. KHARE, S. SRIVASTAVA (1995). *Study of conventional and alternative two-phase sampling ratio, product and regression estimators in presence of non-response*. Proceedings of the Indian National Science Academy, 65, pp. 195–203.

B. B. KHARE, S. SRIVASTAVA (1997). *Transformed ratio type estimators for the population mean in the presence of non-response*. Communication in Statistics - Theory and Methods, 26, pp. 1779–1791.

S. L. LOHR (1999). *Sampling-Design & Analysis*. Duxbury Press, New York.

F. C. OKAFOR, H. LEE (2000). *Double sampling for ratio and regression estimation with sub sampling the non respondents*. Survey Methodology, 26, no. 2, pp. 183–188.

P. S. R. S. RAO (1983). *Randomization approach*. In W. G. MADOW, I. OLKIN, D. B. RUBIN (eds.), *Incomplete Data in Sample Surveys*, Academic Press, New York, vol. 2, pp. 33–44.

P. S. R. S. RAO (1986). *Ratio estimation with sub sampling the non-respondents*. Survey Methodology, 12, pp. 217–230.

S. K. RAY, R. K. SINGH (1981). *Difference-cum-ration type estimators*. Journal of the Indian Statistical Association, 19, pp. 147–151.

D. T. SEARLS (1964). *The utilization of a known coefficient of variation in the estimation procedure*. Journal of the American Statistical Association, 59, pp. 1225–1226.

G. K. SHUKLA (1966). *An alternative multivariate ratio estimate for finite population*. Calcutta Statistical Association Bulletin, 15, pp. 127–134.

H. P. SINGH, S. KUMAR (2008). *A regression approach to the estimation of the finite population mean in the presence of non-response*. Australian & New Zealand Journal of Statistics, 50, no. 4, pp. 395–408.

H. P. SINGH, S. KUMAR (2010). *Estimation of mean in presence of non-response using two phase sampling scheme*. Statistical Papers, 51, pp. 559–582.

H. P. SINGH, S. KUMAR, M. KOZAK (2010). *Improved estimation of finite population mean using sub sampling to deal with non-response in two phase sampling scheme*. Communication in Statistics, Theory and Methods, 39, no. 5, pp. 791–802.

M. P. SINGH (1965). *On estimation of ratio and product of a population parameters*. Sankhya B, 27, pp. 321–328.

M. P. SINGH (1967). *Ratio cum product method of estimation*. Metrika, 112, no. 1, pp. 34–43.

R. K. SINGH, S. BHUSHAN (2012). *Generalized classes of two phase sampling estimators of population mean in presence of non-response*. In *Proceeding of VII ISOS Aligarh Muslim University*. Aligarh Muslim University, Aligarh.

V. K. SINGH, H. P. SINGH, H. P. SINGH, D. SHUKLA (1994). *A general class of chain estimator for ratio and product of two means of a finite population*. Communication in Statistics - Theory and Methods, 23, no. 5, pp. 1341–1365.

S. SRIVASTAVA (1993). *Some problems on the estimation of population mean using auxiliary character in presence of non-response in sample survey*. Ph.D. thesis, Banaras Hindu University, Varanasi, India.

S. K. SRIVASTAVA, H. S. JHAJJ (1981). *A class of estimators of the population mean in survey sampling using auxiliary information*. Biometrika, 68, pp. 341–343.

SUMMARY

This paper introduces an efficient estimation procedure for the population mean in the presence of non-response. The proposed estimators of population mean provides an improvement over the corresponding conventional estimators proposed by Cochran (1977), Rao (1983, 1986) and Singh and Kumar (2008, 2010) under the deterministic non-response in terms of efficiency. A comparative study has been performed and it has been shown that the proposed estimators perform better in comparison to the conventional estimators. The theoretical findings are supported by an empirical study.

Keywords: Auxiliary information; Non-response; Mean square error.