Brazilian Journal of Probability and Statistics
2011, Vol. 25, No. 2, 205–217
DOI: 10.1214/10-BJPS117

Combination of regression and ratio estimate in presence of nonresponse

Housila P. Singh and Sunil Kumar

Vikram University

Abstract. This article addresses the problem of estimating the population mean of the study variable y using information on two auxiliary variables x and z in presence of nonresponse. Two classes of combined regression and ratio estimators are defined in two different situations along with their properties. An empirical study is carried out to judge the merits of the suggested estimators over usual unbiased estimator, ratio estimator and regression estimators. Both theoretical and empirical results are encouraging.

1 Introduction

Consider a finite population $U = \{U_1, \dots, U_N\}$ of size N from which a simple random sample of size n is drawn without replacement. Let the characteristic under study, say y take value y_i on the unit U_i $(i = 1, \dots, N)$. In surveys covering human populations, n_1 units respond on the first attempt while remaining n_2 $(=n - n_1)$ units do not provide any response. Kadilar and Cingi (2008) and Singh (2009) proposed estimators for the single auxiliary variable when some observations are missing. An estimate obtained from such incomplete data may be misleading especially when the respondents differ from the nonrespondents because the estimate can be biased. The work of Hansen and Hurwitz (1946) pioneering the treatment of nonresponse, they suggested a double sampling scheme for estimating population mean comprising the following steps:

(i) a simple random sample of size n is drawn and the questionnaire is mailed to the sample units;
(ii) a subsample of size $r = (n_2/k)$, $(k \geq 1)$ from the n_2 nonresponding units in the initial step attempt is contacted through personal interviews.

It is to be mentioned that Hansen and Hurwitz (1946) considered the mail surveys at the first attempt and the personal interviews at the second attempt. In Hansen and Hurwitz (1946) procedures the population is supposed to be consisting of the response stratum of size N_1 and the nonresponse stratum of size $N_2 = (N - N_1)$. Let $\overline{Y} = \sum_{i=1}^{N} y_i/N$ and $S_y^2 = \sum_{i=1}^{N}(y_i - \overline{Y})^2/(N-1)$ denote the mean and the population variance of the study variable y. Let $\overline{Y}_1 = \sum_{i=1}^{N_1} y_i/N_1$

Key words and phrases. Study variable, auxiliary variable, variance, nonresponse.
Received September 2009; accepted February 2010.

and $S^2_{y(1)} = \sum_{i=1}^{N_1} (y_i - \overline{Y})^2/(N_1 - 1)$ denote the mean and variance of response group. Similarly, let $\overline{Y}_2 = \sum_{i=1}^{N_2} y_i/N_2$ and $S^2_{y(2)} = \sum_{i=1}^{N_2} (y_i - \overline{Y})^2/(N_2 - 1)$ denote the mean and variance of the nonresponse group. The population mean $\overline{Y}$ of the study variable y can be defined as $\overline{Y} = W_1\overline{Y}_1 + W_2\overline{Y}_2$, where $W_1 = (N_1/N)$ and $W_2 = (N_2/N)$. Besides, let $(\bar{y}, s^2_y)$, $(\bar{y}_1, s^2_{y(1)})$, $(\bar{y}_2, s^2_{y(2)})$ and $(\bar{y}_{r(2)}, s^2_{yr(2)})$ be the means together with variances based on n, n_1, n_2 and r units respectively where

$$s^2_y = \sum_{i=1}^{n} (y_i - \bar{y})^2/(n-1), \qquad s^2_{y(1)} = \sum_{i=1}^{n_1} (y_i - \bar{y}_1)^2/(n_1 - 1),$$

$$s^2_{y(2)} = \sum_{i=1}^{n_2} (y_i - \bar{y}_2)^2/(n_2 - 1) \quad \text{and} \quad s^2_{yr(2)} = \sum_{i=1}^{r} (y_i - \bar{y}_{r2})^2/(r - 1).$$

Hansen and Hurwitz (1946) suggested an unbiased estimator for the population mean $\overline{Y}$ is given by

$$\bar{y}^* = w_1\bar{y}_1 + w_2\bar{y}_{r2},$$

where $w_1 = (n_1/n)$ and $w_2 = (n_2/n)$ are responding and nonresponding proportions in the sample. The variance of $\bar{y}^*$ is given by

$$\mathrm{Var}(\bar{y}^*) = \left(\frac{1-f}{n}\right) S^2_y + \frac{W_2(k-1)}{n} S^2_{y(2)}, \tag{1.1}$$

where $f = (n/N)$.

In sample surveys precision in estimating the population mean $\overline{Y}$ can be increased by utilizing information on the auxiliary variable x which is correlated with y whose population mean $\overline{X}$ is known. Using Hansen and Hurwitz (1946) technique, Cochran (1977) suggested the ratio and regression estimators of the population mean $\overline{Y}$ of the study variable y in which information on the auxiliary variable is obtained from all the sample units, and the population mean $\overline{X}$ of the auxiliary variable x is known, while some sample units failed to supply information on study variable y. Rao (1986, 1987), Khare and Srivastava (1993, 1995, 1997), Okafor and Lee (2000), Särndal and Lundstrom (2005), Tabasum and Khan (2004, 2006) and Singh and Kumar (2008a, 2008b, 2008c, 2009a, 2009b) have suggested some estimators for population mean $\overline{Y}$ of the study variable y using auxiliary information in presence of nonresponse and studied their properties.

Let x_i, $i = (1, \ldots, N)$ denote an auxiliary characteristic with population mean $\overline{X} = \sum_{i=1}^{N} x_i/N$. Let $\overline{X}_1$ and $\overline{X}_2$ denote the population means of the response and nonresponse groups. Let $\bar{x} = \sum_{i=1}^{n} x_i/n$ denote the mean of all the n units. Let $\bar{x}_1 = \sum_{i=1}^{n_1} x_i/n_1$ and $\bar{x}_2 = \sum_{i=1}^{n_2} x_i/n_2$ denote the means of the n_1 responding units and the n_2 nonresponding units. Further, let $\bar{x}_{r2} = \sum_{i=1}^{r} x_i/r$ denote the mean of the $r = (n_2/k)$ subsampled units. When there is nonresponse on both the variables study as well as auxiliary and the population mean $\overline{X}$ of the auxiliary variable x is known, the conventional ratio and regression estimators for population mean $\overline{Y}$ are respectively defined by

$$t_1^* = \frac{\bar{y}^*}{\bar{x}^*}\overline{X} \qquad \text{(ratio estimator)}$$

and

$$t_2^* = \bar{y}^* + \hat{\beta}_{yx}^*(\overline{X} - \bar{x}^*) \qquad \text{(regression estimator)},$$

where

$$\bar{x}^* = w_1\bar{x}_1 + w_2\bar{x}_{r2}, \qquad \hat{\beta}_{yx}^* = (s_{xy}^*/s_x^{*2}),$$

$$s_{xy}^* = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i + k\sum_{i=1}^{r} x_i y_i - n\bar{x}\bar{y}^*\right),$$

$$\bar{x} = \sum_{i=1}^{n} x_i/n \quad \text{and} \quad s_x^{*2} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 + k\sum_{i=1}^{r} x_i^2 - n\bar{x}\bar{x}^*\right).$$

When there is incomplete information on the study variable y and complete information on the auxiliary variable x are available. In addition, the population mean $\overline{X}$ of the auxiliary variable x is known. The conventional ratio and regression estimators for population mean $\overline{Y}$ of the study variable y are respectively defined by

$$t_1 = \bar{y}^*\left(\frac{\overline{X}}{\bar{x}}\right) \qquad \text{(ratio estimator)}$$

and

$$t_2 = \bar{y}^* + \hat{\beta}_{yx}(\overline{X} - \bar{x}) \qquad \text{(regression estimator)},$$

where $\hat{\beta}_{yx} = (s_{xy}^*/s_x^2)$, $s_x^2 = \sum_{i=1}^{n}(x_i - \overline{X})^2/(n-1)$.

In the present paper motivated by Mohanty (1967) in presence of nonresponse we have suggested a class of combined ratio and regression estimators for estimating the population mean $\overline{Y}$ of the study variable y when there two auxiliary variable x and z which are correlated to the study variable y and the population means $\overline{X}$ and $\overline{Z}$ of x and z respectively are known. The auxiliary variable z may be the value of y at some previous time when a complete census was taken, and x is another auxiliary variable which we come across when y is being measured and which is highly correlated with the study variable y. In support of the present study an empirical study is carried out.

2 The suggested class of combined ratio and regression estimators

In this section utilizing information on two auxiliary variables x and z with known population means, we have suggested two different classes of estimators for population mean $\overline{Y}$ in two different situations which are as follows:

Case I: *Population means* $\overline{X}$ *and* $\overline{Z}$ *are known, incomplete information on* y, x *and* z.

In this case we observe that n_1 units respond for y, x and z from the sample of size n and population means $\overline{X}$ and $\overline{Z}$ are known. Motivated by Reddy (1974, 1978) and Walsh (1970) we define a class of estimators for population mean $\overline{Y}$ of y in presence of nonresponse as

$$t^*_{(\alpha)} = \{\bar{y}^* + \hat{\beta}^*_{yx}(\overline{X} - \bar{x}^*)\} \frac{\overline{Z}}{\{\overline{Z} + \alpha(\bar{z}^* - \overline{Z})\}}, \tag{2.1}$$

where α is a suitably chosen scalar and $\bar{z}^* = w_1\bar{z}_1 + w_2\bar{z}_{r2}$ is an unbiased estimator of the population mean $\overline{Z}$ defined using Hansen and Hurwitz (1946) technique, $\overline{Z}_1$ and $\overline{Z}_{r2}$ are the sample means of the auxiliary variable z based on n_1 and r units, respectively.

To obtain the bias and variance of the estimator $t^*_{(\alpha)}$ we write

$$\bar{y}^* = \overline{Y}(1+\varepsilon_0), \qquad \bar{x}^* = \overline{X}(1+\varepsilon_1), \qquad \bar{z}^* = \overline{Z}(1+\varepsilon_2),$$
$$s^*_{xy} = S_{xy}(1+\varepsilon_3), \qquad s^{*2}_x = S^2_x(1+\varepsilon_4)$$

such that

$$E(\varepsilon_i) = 0 \qquad \forall i = 0 \text{ to } 4,$$
$$E(\varepsilon_0^2) = \left(\frac{1-f}{n}\right)C_y^2 + \frac{W_2(k-1)}{n}C^2_{y(2)},$$
$$E(\varepsilon_1^2) = \left(\frac{1-f}{n}\right)C_x^2 + \frac{W_2(k-1)}{n}C^2_{x(2)},$$
$$E(\varepsilon_2^2) = \left(\frac{1-f}{n}\right)C_z^2 + \frac{W_2(k-1)}{n}C^2_{z(2)},$$
$$E(\varepsilon_0\varepsilon_1) = \left(\frac{1-f}{n}\right)\rho_{xy}C_yC_x + \frac{W_2(k-1)}{n}\rho_{xy(2)}C_{y(2)}C_{x(2)},$$
$$E(\varepsilon_0\varepsilon_2) = \left(\frac{1-f}{n}\right)\rho_{yz}C_yC_z + \frac{W_2(k-1)}{n}\rho_{yz(2)}C_{y(2)}C_{z(2)},$$
$$E(\varepsilon_1\varepsilon_2) = \left(\frac{1-f}{n}\right)\rho_{xz}C_xC_z + \frac{W_2(k-1)}{n}\rho_{xz(2)}C_{x(2)}C_{z(2)},$$
$$E(\varepsilon_1\varepsilon_3) = \frac{N(N-n)}{(N-1)(N-2)}\frac{\mu_{21}}{n\overline{X}S_{xy}} + \frac{W_2(k-1)}{n}\frac{\mu_{21(2)}}{\overline{X}S_{xy}},$$

$$E(\varepsilon_1\varepsilon_4) = \frac{N(N-n)}{(N-1)(N-2)}\frac{\mu_{30}}{n\overline{X}S_x^2} + \frac{W_2(k-1)}{n}\frac{\mu_{30(2)}}{\overline{X}S_x^2},$$

$$C_y = S_y/\overline{Y}, \qquad C_{y(2)} = S_{y(2)}/\overline{Y}, \qquad C_x = S_x/\overline{X},$$

$$C_{x(2)} = S_{x(2)}/\overline{X}, \qquad C_z = S_z/\overline{Z}, \qquad C_{z(2)} = S_{z(2)}/\overline{Z},$$

$$\rho_{xy} = S_{xy}/S_xS_y, \qquad \rho_{xy(2)} = S_{xy(2)}/S_{x(2)}S_{y(2)},$$

$$\rho_{yz} = S_{yz}/S_yS_z, \qquad \rho_{yz(2)} = S_{yz(2)}/S_{y(2)}S_{z(2)},$$

$$\rho_{xz} = S_{xz}/S_xS_z, \qquad \rho_{xz(2)} = S_{xz(2)}/S_{x(2)}S_{z(2)},$$

$$\mu_{\nu s} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{X})^{\nu}(y_i - \overline{Y})^{s},$$

$$\mu_{\nu s(2)} = \frac{1}{N_2}\sum_{i=1}^{N_2}(x_i - \overline{X}_2)^{\nu}(y_i - \overline{Y}_2)^{s},$$

$$\overline{U} = \frac{1}{N}\sum_{i=1}^{N}u_i, \qquad \overline{U}_2 = \frac{1}{N_2}\sum_{i=1}^{N_2}u_i, \qquad u = (x, y, z),$$

(ν, s) being nonnegative integers.

Expressing $t^*_{(\alpha)}$ in terms of ε's we have

$$\begin{aligned} t^*_{(\alpha)} &= \left\{\overline{Y}(1+\varepsilon_0) + \frac{S_{xy}(1+\varepsilon_3)}{S_x^2(1+\varepsilon_4)}(-\overline{X}\varepsilon_1)\right\}(1+\alpha\varepsilon_2)^{-1} \\ &= \overline{Y}\left\{1+\varepsilon_0 - \left(\frac{\beta_{yx}\overline{X}}{\overline{Y}}\right)\varepsilon_1(1+\varepsilon_3)(1+\varepsilon_4)^{-1}\right\}(1+\alpha\varepsilon_2)^{-1}, \end{aligned} \tag{2.2}$$

where $\beta_{yx} = (S_{xy}/S_x^2)$ is the population regression coefficient of y on x.

We assume that $|\alpha\varepsilon_2| < 1$ and $|\varepsilon_4| < 1$ so that $(1+\alpha\varepsilon_2)^{-1}$ and $(1+\varepsilon_4)^{-1}$ are expandable in terms of ε's. Expanding the right-hand side (r.h.s.) of (2.2), multiplying out and neglecting terms of ε's having power greater than two we have

$$\begin{aligned} (t^*_{(\alpha)} - \overline{Y}) = \overline{Y}[&\varepsilon_0 - A_0\varepsilon_1 - \alpha\varepsilon_2 \\ &+ \alpha(\alpha\varepsilon_2^2 - \varepsilon_0\varepsilon_2 + A_0\varepsilon_1\varepsilon_2) - A_0(\varepsilon_1\varepsilon_3 - \varepsilon_1\varepsilon_4)], \end{aligned} \tag{2.3}$$

where $A_0 = (\beta_{yx}/R)$ and $R = (\overline{Y}/\overline{X})$.

Taking expectations of both sides of (2.3) we get the bias of $t^*_{(\alpha)}$ to the first degree of approximation as

$$\begin{aligned} B(t^*_{(\alpha)}) = \Big[&\left(\frac{1-f}{n}\right)\overline{Y}\alpha C_z^2(\alpha - K_{yz} + K_{xy}K_{xz}) \\ &+ \frac{W_2(k-1)}{n}C_{z(2)}^2\overline{Y}\alpha\big(\alpha - K_{yz(2)} + K_{xy}K_{xz(2)}\big) \end{aligned} \tag{2.4}$$

$$-\beta_{yx}\left\{\left(\frac{N-n}{N-2}\right)\frac{1}{n}\left(\frac{\mu_{21}}{\mu_{11}}-\frac{\mu_{30}}{\mu_{20}}\right)+\frac{W_2(k-1)}{n}\left(\frac{\mu_{21(2)}}{\mu_{11}}-\frac{\mu_{30(2)}}{\mu_{20}}\right)\right\}\Bigg],$$

where $K_{yz}=\rho_{yz}(C_y/C_z)$, $K_{xy}=\rho_{xy}(C_y/C_x)$, $K_{xz}=\rho_{xz}(C_x/C_z)$, $K_{yz(2)}=\rho_{yz(2)}(C_{y(2)}/C_{z(2)})$ and $K_{xz(2)}=\rho_{xz(2)}(C_{x(2)}/C_{z(2)})$.

Squaring both sides of (2.3) and neglecting terms of ε's having power greater than two we have

$$\begin{aligned}(t^*_{(\alpha)}-\overline{Y})^2&=\overline{Y}^2(\varepsilon_0-A_0\varepsilon_1-\alpha\varepsilon_2)^2\\&=\overline{Y}^2\{\varepsilon_0^2+A_0^2\varepsilon_1^2+\alpha^2\varepsilon_2^2-2A_0\varepsilon_0\varepsilon_1-2\alpha(\varepsilon_0\varepsilon_2-A_0\varepsilon_1\varepsilon_2)\}.\end{aligned}\tag{2.5}$$

Taking expectations of both sides of (2.5) we get the variance of $t^*_{(\alpha)}$ to the first degree of approximation as

$$\begin{aligned}\mathrm{Var}(t^*_{(\alpha)})=\Bigg[\left(\frac{1-f}{n}\right)&\{S_y^2(1-\rho_{xy}^2)+\alpha R^*(\alpha R^*-2A)S_z^2\}\\&+\frac{W_2(k-1)}{n}\{S^2_{y(2)}+\beta_{yx}S^2_{x(2)}(\beta_{yx}-2\beta_{yx(2)})\\&\qquad+\alpha R^*(\alpha R^*-2B)S^2_{z(2)}\}\Bigg],\end{aligned}\tag{2.6}$$

where

$$A=(\beta_{yz}-\beta_{yx}\beta_{xz}),\qquad B=(\beta_{yz(2)}-\beta_{yx}\beta_{xz(2)}),\qquad R^*=\overline{Y}/\overline{Z},$$
$$\beta_{xz}=S_{xz}/S_z^2,\qquad \beta_{xz(2)}=S_{xz(2)}/S^2_{z(2)},\qquad \beta_{yz}=S_{yz}/S_z^2,$$
$$\beta_{yz(2)}=S_{yz(2)}/S^2_{z(2)},\qquad \beta_{yx}=S_{yx}/S_x^2\quad\text{and}\quad \beta_{yx(2)}=S_{yx(2)}/S^2_{x(2)},$$

which is minimum when

$$\alpha=\{N^*/(R^*D^*)\}=\alpha_0\qquad\text{(say)},$$

where

$$N^*=\left\{\left(\frac{1-f}{n}\right)AS_z^2+\frac{W_2(k-1)}{n}BS^2_{z(2)}\right\},$$
$$D^*=\left\{\left(\frac{1}{n}-\frac{1}{N}\right)S_z^2+\frac{W_2(k-1)}{n}S^2_{z(2)}\right\}.$$

Thus the resulting minimum variance of $t^*_{(\alpha)}$ is given by

$$\min\mathrm{Var}(t^*_{(\alpha)})=\mathrm{Var}(t^*_2)-(N^{*2}/D^*),\tag{2.7}$$

where

$$\begin{aligned}\mathrm{Var}(t^*_2)=\Bigg[\left(\frac{1-f}{n}\right)&S_y^2(1-\rho_{xy}^2)\\&+\frac{W_2(k-1)}{n}\{S^2_{y(2)}+\beta_{yx}S^2_{x(2)}(\beta_{yx}-2\beta_{yx(2)})\}\Bigg]\end{aligned}\tag{2.8}$$

is approximate variance of the regression estimator $t_2^* = \bar{y}^* + \hat{\beta}_{yx}^*(\overline{X} - \bar{x}^*)$.

To the first degree of approximation, the variance of the ratio estimator t_1^* is given by

$$\operatorname{Var}(t_1^*) = \left[\left(\frac{1-f}{n}\right)\{S_y^2 + S_x^2 R(R - 2\beta_{yx})\} + \frac{W_2(k-1)}{n}\{S_{y(2)}^2 + RS_{x(2)}^2(R - 2\beta_{yx(2)})\}\right]. \tag{2.9}$$

If we set $\alpha = 1$ in (2.1) we get an estimator for $\overline{Y}$ as

$$t_{(1)}^* = \{\bar{y}^* + \hat{\beta}_{yx}^*(\overline{X} - \bar{x}^*)\}\left(\frac{\overline{Z}}{\bar{z}^*}\right),$$

with the approximate variance

$$\operatorname{Var}(t_{(1)}^*) = \left[\left(\frac{1-f}{n}\right)\{S_y^2(1-\rho_{xy}^2) + R^*(R^* - 2A)S_z^2\} + \frac{W_2(k-1)}{n}\{S_{y(2)}^2 + \beta_{yx}S_{x(2)}^2(\beta_{yx} - 2\beta_{yx(2)}) + R^*(R^* - 2B)S_{z(2)}^2\}\right]. \tag{2.10}$$

From (1.1), (2.7), (2.8) and (2.10) we have

$$\operatorname{Var}(\bar{y}^*) - \min \operatorname{Var}(t_{(\alpha)}^*) = \left\{\left(\frac{1-f}{n}\right)S_y^2\rho_{xy}^2 + \frac{W_2(k-1)}{n}\beta_{yx}S_{x(2)}^2(2\beta_{yx(2)} - \beta_{yx}) + \frac{N^{*2}}{D^*}\right\} > 0 \quad \text{if } \beta_{yx(2)} > \beta_{yx}/2, \tag{2.11}$$

$$\operatorname{Var}(t_2^*) - \min \operatorname{Var}(t_{(\alpha)}^*) = \frac{N^{*2}}{D^*} > 0, \tag{2.12}$$

$$\operatorname{Var}(t_{(1)}^*) - \min \operatorname{Var}(t_{(\alpha)}^*) = \frac{(R^*D^* - N^*)^2}{D^*} > 0. \tag{2.13}$$

It is observed from (2.11), (2.12) and (2.13) that the proposed estimator $t_{(\alpha)}^*$ with $\alpha = \alpha_0$ is:

(i) better than $\bar{y}^*$ if $\beta_{yx(2)} > \beta_{yx}/2$.
(ii) better than regression estimator t_2^* and the ratio estimator $t_{(1)}^*$.

If α does not coincide with its exact optimum value, that is, $\alpha \neq \alpha_0$ then from (1.1), (2.8) and (2.10) we note that the suggested class of estimators $t_{(\alpha)}^*$ is better than:

(i) the usual unbiased estimator $\bar{y}^*$ if

$$\frac{N^* - \sqrt{N^{*2} + D^*C^*}}{R^*D^*} < \alpha < \frac{N^* + \sqrt{N^{*2} + D^*C^*}}{R^*D^*},$$

where

$$C^* = \left\{ \left(\frac{1-f}{n}\right) S_y^2 \rho_{yx}^2 + \frac{W_2(k-1)}{n} \beta_{yx} S_{x(2)}^2 (2\beta_{yx(2)} - \beta_{yx}) \right\}.$$

(ii) the regression estimator t_2^* if

$$0 < \alpha < 2\{N^*/(R^*D^*)\},$$

(iii) the ratio estimator $t_{(1)}^*$ if

$$\frac{N^* - \sqrt{N^{*2} + D^*E^*}}{R^*D^*} < \alpha < \frac{N^* + \sqrt{N^{*2} + D^*E^*}}{R^*D^*},$$

where

$$E^* = \left\{ C^* + \left(\frac{1-f}{n}\right) R S_x^2 (R - 2\beta_{yx}) + \frac{W_2(k-1)}{n} R S_{x(2)}^2 (R - \beta_{yx(2)}) \right\}.$$

From (2.8) and (2.10), we note that $\mathrm{Var}(t_{(1)}^*) < \mathrm{Var}(t_2^*)$ if

$$A > R^*/2 \quad \text{and} \quad B > R^*/2.$$

Case II: *Population means $\overline{X}$ and $\overline{Z}$ of the auxiliary variables x and z are known, incomplete information on the study variable y and complete information on both auxiliary variables x and z.*

We consider the situation in which information on auxiliary variables x and z are obtained from all the sample units n, and the population means $\overline{X}$ and $\overline{Z}$ are known, while some units are failed to supply information on the study variable y, that is, in this case, we use information on $(n_1 + r)$ responding units on the study variable y and complete information on both auxiliary variables x and z from the sample of size n. Thus we define a class of estimators for the population mean $\overline{Y}$ of the study variable y as

$$t_{(\alpha)} = \{\bar{y}^* + \hat{\beta}_{yx}(\overline{X} - \bar{x})\} \frac{\overline{Z}}{\{\overline{Z} + \alpha^*(\bar{z} - \overline{Z})\}} = t_2 \frac{\overline{Z}}{\{\overline{Z} + \alpha^*(\bar{z} - \overline{Z})\}}, \qquad (2.14)$$

where α^* is a suitably chosen scalar and t_2 is defined as $t_2 = \bar{y}^* + \hat{\beta}_{yx}(\overline{X} - \bar{x})$.

To obtain the bias and variance of the estimator $t_{(\alpha)}$ we write

$$\bar{y}^* = \overline{Y}(1 + \varepsilon_0), \qquad \bar{x} = \overline{X}(1 + e_1), \qquad \bar{z} = \overline{Z}(1 + e_2),$$
$$s_{xy} = S_{xy}(1 + e_3), \qquad s_x^2 = S_x^2(1 + e_4)$$

such that

$$E(e_i) = 0 \qquad \forall i = 0 \text{ to } 4,$$
$$E(e_1^2) = \left(\frac{1-f}{n}\right)C_x^2, \qquad E(e_2^2) = \left(\frac{1-f}{n}\right)C_z^2,$$
$$E(\varepsilon_0 e_1) = \left(\frac{1-f}{n}\right)\rho_{xy}C_yC_x,$$
$$E(\varepsilon_0 e_2) = \left(\frac{1-f}{n}\right)\rho_{yz}C_yC_z, \qquad E(e_1e_2) = \left(\frac{1-f}{n}\right)\rho_{xz}C_xC_z,$$
$$E(e_1e_3) = \frac{N(N-n)}{(N-1)(N-2)}\left(\frac{\mu_{21}}{n\overline{X}S_{xy}}\right),$$
$$E(e_1e_4) = \frac{N(N-n)}{(N-1)(N-2)}\left(\frac{\mu_{30}}{n\overline{X}S_x^2}\right).$$

Expressing $t_{(\alpha)}$ in terms of ε's and e's we have

$$\begin{aligned} t_{(\alpha)} &= \left\{\overline{Y}(1+\varepsilon_0) + \frac{S_{xy}(1+e_3)}{S_x^2(1+e_4)}(-\overline{X}e_1)\right\}(1+\alpha^* e_2)^{-1} \\ &= \overline{Y}\left\{1+\varepsilon_0 - \left(\frac{\beta_{yx}\overline{X}}{\overline{Y}}\right)e_1(1+e_3)(1+e_4)^{-1}\right\}(1+\alpha^* e_2)^{-1}. \end{aligned} \tag{2.15}$$

We assume that $|\alpha^* e_2| < 1$ and $|e_4| < 1$ so that $(1+\alpha^* e_2)^{-1}$ and $(1+e_4)^{-1}$ are expandable in terms of ε's and e's. Expanding the right-hand side (r.h.s.) of (2.15), multiplying out and neglecting terms of ε's and e's having power greater than two we have

$$\begin{aligned} (t_{(\alpha)} - \overline{Y}) = \overline{Y}(&\varepsilon_0 - A_0e_1 - \alpha^* e_2 + A_0e_1e_4 \\ &- A_0e_1e_3 - \alpha^*\varepsilon_0e_2 + \alpha^* A_0e_1e_2 + \alpha^{*2}e_2^2). \end{aligned} \tag{2.16}$$

Taking expectations of both sides of (2.16) we get the bias of $t_{(\alpha)}$ to the first degree of approximation as

$$\begin{aligned} B(t_{(\alpha)}) = \Bigg[&\left(\frac{1-f}{n}\right)\{\overline{Y}\alpha^* C_z^2(\alpha^* - K_{yz} + K_{xy}K_{xz})\} \\ &- \frac{\beta_{yx}}{n}\left(\frac{N-n}{N-2}\right)\left(\frac{\mu_{21}}{\mu_{11}} - \frac{\mu_{30}}{\mu_{20}}\right)\Bigg]. \end{aligned} \tag{2.17}$$

Squaring both sides of (2.16) and neglecting terms of ε's and e's having power greater than two, we have

$$\begin{aligned} (t_{(\alpha)} - \overline{Y})^2 &= \overline{Y}^2(\varepsilon_0 - A_0e_1 - \alpha^* e_2)^2 \\ &= \overline{Y}^2(\varepsilon_0^2 + A_0^2e_1^2 + \alpha^{*2}e_2^2 - 2A_0\varepsilon_0e_1 - 2\alpha^*\varepsilon_0e_2 + 2A_0\alpha^* e_1e_2). \end{aligned} \tag{2.18}$$

Taking expectations of both sides of (2.18) we get the variance of $t_{(\alpha)}$ to the first degree of approximation as

$$\begin{aligned}\mathrm{Var}(t_{(\alpha)}) = \Big[\Big(\frac{1-f}{n}\Big)\{S_y^2(1-\rho_{xy}^2) + \alpha^* R^*(\alpha^* R^* - 2A)S_z^2\} \\ + \frac{W_2(k-1)}{n} S_{y(2)}^2\Big],\end{aligned} \tag{2.19}$$

which is minimum when

$$\alpha^* = (A/R^*) = \alpha_0^* \qquad \text{(say)}.$$

Thus the resulting minimum variance of $t_{(\alpha)}$ is given by

$$\min \mathrm{Var}(t_{(\alpha)}) = \mathrm{Var}(t_2) - \Big(\frac{1-f}{n}\Big)A^2 S_z^2, \tag{2.20}$$

where

$$\mathrm{Var}(t_2) = \Big\{\Big(\frac{1-f}{n}\Big)S_y^2(1-\rho_{xy}^2) + \frac{W_2(k-1)}{n}S_{y(2)}^2\Big\} \tag{2.21}$$

is approximate variance of the regression estimator t_2 defined as $t_2 = \bar{y}^* + \hat{\beta}_{yx}(\overline{X} - \bar{x})$.

To the first degree of approximation, the variance of the ratio estimator t_1 is given by

$$\mathrm{Var}(t_1) = \Big[\Big(\frac{1-f}{n}\Big)\{S_y^2 + S_x^2 R(R - 2\beta_{yx})\} + \frac{W_2(k-1)}{n}S_{y(2)}^2\Big]. \tag{2.22}$$

Putting $\alpha = 1$ in (2.14) we get an estimator for population mean $\overline{Y}$ of the study variable y as

$$t_{(1)} = \{\bar{y}^* + \hat{\beta}_{yx}(\overline{X} - \bar{x})\}\Big(\frac{\overline{Z}}{\bar{z}}\Big),$$

with the approximate variance

$$\begin{aligned}\mathrm{Var}(t_{(1)}) &= \Big[\Big(\frac{1-f}{n}\Big)\{S_y^2(1-\rho_{xy}^2) + R^*(R^* - 2A)S_z^2\} + \frac{W_2(k-1)}{n}S_{y(2)}^2\Big] \\ &= \mathrm{Var}(t_2) + \Big(\frac{1-f}{n}\Big)\{R^*(R^* - 2A)S_z^2\},\end{aligned} \tag{2.23}$$

which shows that $\mathrm{Var}(t_{(1)}) < \mathrm{Var}(t_2)$ if

$$A > (R^*/2).$$

From (1.1), (2.20), (2.21) and (2.23) we have

$$\mathrm{Var}(\bar{y}^*) - \min \mathrm{Var}(t_{(\alpha)}) = \left[\left(\frac{1-f}{n}\right)\{S_y^2\rho_{xy}^2 + A^2 S_z^2\}\right] > 0, \tag{2.24}$$

$$\mathrm{Var}(t_2) - \min \mathrm{Var}(t_{(\alpha)}) = \left(\frac{1-f}{n}\right) A^2 S_z^2 > 0, \tag{2.25}$$

$$\mathrm{Var}(t_{(1)}) - \min \mathrm{Var}(t_{(\alpha)}) = \left(\frac{1-f}{n}\right)(A - R^*)^2 S_z^2 > 0. \tag{2.26}$$

It is observed from (2.24), (2.25) and (2.26) that the proposed estimator $t_{(\alpha)}$ with $\alpha = \alpha_0^*$ is better than the usual unbiased estimator $\bar{y}^*$, the regression estimator t_2 and the ratio estimator $t_{(1)}$.

If α does not coincide with its exact optimum value i.e. $\alpha \neq \alpha_0^*$ then from (1.1), (2.21) and (2.23) we note that the suggested class of estimators $t_{(\alpha)}$ is better than:

(i) the usual unbiased estimator $\bar{y}^*$ if

$$\frac{A - \sqrt{A^2 + R^{*2}K_{xy}^2}}{R^*} < \alpha < \frac{A + \sqrt{A^2 + R^{*2}K_{xy}^2}}{R^*},$$

(ii) the regression estimator t_2 if

$$0 < \alpha < 2(A/R^*),$$

(iii) the ratio estimator $t_{(1)}$ if

$$\frac{A - \sqrt{A^2 + R^*(R^* - 2A)}}{R^*} < \alpha < \frac{A + \sqrt{A^2 + R^*(R^* - 2A)}}{R^*}.$$

3 Empirical study

To illustrate the results we consider the data earlier consider by Khare and Sinha (2007). The description of the population is given below:

The data on physical growth of upper socioeconomic group of 95 school children of Varanasi under an ICMR study, Department of Pediatrics, B. H. U., during 1983–84 has been taken under study. The first 25% (i.e., 24 children) units have been considered as nonresponding units. Here we have taken the study characters and the auxiliary characters as follows:

y: weight (in kg) of the children,
x: skull circumference (in cm) of the children,
z: chest circumference (in cm) of the children.

$$\overline{Y} = 19.4968, \quad \overline{X} = 51.1726, \quad \overline{Z} = 55.8611,$$
$$C_y = 0.15613, \quad C_x = 0.03006, \quad C_z = 0.05860,$$
$$C_{y(2)} = 0.12075, \quad C_{x(2)} = 0.02478, \quad C_{z(2)} = 0.05402,$$

Table 1 *Percent Relative Efficiencies (PREs) of estimators with respect to $\bar{y}^*$ for different values of k*

		$(1/k)$			
	Estimators	(1/5)	(1/4)	(1/3)	(1/2)
Case I	$\bar{y}^*$	100.00	100.00	100.00	100.00
	t_1^*	113.78	113.18	112.40	111.34
	t_2^*	118.77	117.75	116.38	114.58
	$t_{(1)}^*$	191.02	193.83	197.67	203.24
	$t_{(\alpha)}^*$	203.69	211.46	223.68	246.43
Case II	t_1	104.83	105.54	106.48	107.81
	t_2	105.86	106.72	107.88	109.52
	$t_{(1)}$	137.44	144.72	156.13	174.79
	$t_{(\alpha)}$	152.49	164.39	183.24	217.72

$$\rho_{yx} = 0.328, \qquad \rho_{yx(2)} = 0.477, \qquad \rho_{yz} = 0.846,$$
$$\rho_{yz(2)} = 0.729, \qquad \rho_{xz} = 0.297, \qquad \rho_{xz(2)} = 0.570,$$
$$N = 95, \qquad n = 35, \qquad W_2 = 0.25.$$

We have computed the percent relative efficiencies (PREs) of different estimators of population mean $\overline{Y}$ with respect to usual unbiased estimator $\bar{y}^*$ for varying values of k. Findings are presented in Table 1.

It is observed from Table 1 that in Case I, the percent relative efficiencies (PREs) of $t_{(1)}^*$ and $t_{(\alpha)}^*$ increase as the value of k increases while the PREs of t_1^* and t_2^* decrease as the value of k increases. We also note that the proposed estimator $t_{(\alpha)}^*$ is the best among all estimators $\bar{y}^*$, t_1^*, t_2^*, $t_{(1)}^*$ and $t_{(\alpha)}^*$. In Case II, the percent relative efficiencies (PREs) of the estimators t_1, t_2, $t_{(1)}$ and $t_{(\alpha)}$ increase as the value of k increases. Again it is observed that the proposed estimator $t_{(\alpha)}$ is the best among all estimators $\bar{y}^*$, t_1, t_2, $t_{(1)}$ and $t_{(\alpha)}$. Thus we conclude that the proposed estimators $t_{(\alpha)}^*$ and $t_{(\alpha)}$ are to be recommended for their use in practice.

Acknowledgment

Authors are thankful to the referee for his valuable constructive suggestions regarding improvement of the paper.

References

Cochran, W. G. (1977). *Sampling Techniques*, 3rd ed. New York: John Wiley and Sons. MR0474575

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

Kadilar, C. and Cingi, H. (2008). Estimators for the population mean in the case of missing data. *Communications in Statistics: Theory and Methods* **37**, 2226–2236. MR2526677

Khare, B. B. and Sinha, R. R. (2007). Estimation of the ratio of the two population means using multi auxiliary characters in the presence of non-response. *Statistical Techniques in Life Testing, Reliability, Sampling Theory and Quality Control*, 163–171.

Khare, B. B. and Srivastava, S. (1993). Estimation of population mean using auxiliary character in presence of non-response. *National Academy Science Letters, India* **16**, 111–114.

Khare, B. B. and Srivastava, S. (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**, II, 195–203.

Khare, B. B. and Srivastava, S. (1997). Transformed ratio type estimators for the population mean in the presence of non-response. *Communications in Statistics: Theory and Methods* **26**, 1779–1791. MR1457675

Mohanty, S. (1967). Combination of regression and ratio estimates. *Journal of the Indian Statistical Association* **5**, 1–14.

Okafor, F. C. and Lee, H. (2000). Double sampling for ratio and regression estimation with sub-sampling the non-respondents. *Survey Methodology* **26**, 183–188.

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

Rao, P. S. R. S. (1987). Ratio and regression estimates with sub sampling the non respondents. In *Paper presented at a special contributed session of the International Statistical Association Meeting*, Sept. 2–16, Tokyo, Japan.

Reddy, V. N. (1974). A transformed ratio method of estimation. *Sankhya* **36**, 59–70.

Reddy, V. N. (1978). A study on the use of prior knowledge on certain population parameters in estimation. *Sankhya* **40**, 29–34.

Särndal, C. E. and Lundstrom, S. (2005). *Estimation in Surveys with non-response*. New York: Wiley.

Singh, S. (2009). A new method of imputation in survey sampling. *Statistics* **43**, 499–511. MR2566697

Singh, H. P. and Kumar, S. (2008a). Estimation of mean in presence of non-response using two phase sampling scheme. *Statistical Papers*, DOI:10.1007/s00362-008-0140-5.

Singh, H. P. and Kumar, S. (2008b). A regression approach to the estimation of finite population mean in presence of non-response. *Australian and New Zealand Journal of Statistics* **50**, 395–408. MR2474199

Singh, H. P. and Kumar, S. (2008c). A general family of estimators of finite population ratio, product and mean using two phase sampling scheme in the presence of non-response. *Journal of Statistical Theory and Practice* **2**, 677–692.

Singh, H. P. and Kumar, S. (2009a). A general class of estimators of the population mean in survey sampling using auxiliary information with sub sampling the non-respondents. *Korean Journal of Applied Statistics* **22**, 387–402.

Singh, H. P. and Kumar, S. (2009b). A general procedure of estimating the population mean in the presence of non-response under double sampling using auxiliary information. *SORT* **33**, 71–84. MR2532232

Tabasum, R. and Khan, I. A. (2004). Double sampling for ratio estimation with non response. *Journal of the Indian Society of Agricultural Statistics* **58**, 300–306. MR2275109

Tabasum, R. and Khan, I. A. (2006). Double sampling ratio estimator for the population mean in presence of non-response. *Assam Statistical Review* **20**, 73–83.

Walsh, J. E. (1970). Generalization of ratio estimate for population total. *Sankhyā* **32**, 99–106.

School of Studies in Statistics
Vikram University
Ujjain-456010, M. P.
India
E-mail: hpsujn@rediffmail.com
sunilbhougal06@gmail.com