Hacettepe Journal of Mathematics and Statistics
Volume 40 (5) (2011), 757 – 765

# A GENERAL CLASS OF ESTIMATORS IN TWO-STAGE SAMPLING WITH TWO AUXILIARY VARIABLES

L. N. Sahoo*†, R. K. Sahoo*, S. C. Senapati‡ and A. K. Mangaraj§

Received 08 : 03 : 2009 : Accepted 10 : 03 : 2011

**Abstract**

This paper presents a general class of estimators for a finite population total with the aid of two auxiliary variables in a two-stage sampling with varying probabilities. The methodology developed can be extended readily to three-stage and stratified two-stage sampling designs.

**Keywords:** Asymptotic variance, Auxiliary variable, Two-stage sampling.

*2000 AMS Classification:* 62 D 05.

## 1. Introduction

Consider $U$, a finite population consisting of $N$ first stage units (*fsu*) $U_1, U_2, \dots, U_N$, such that $U_i$ contains $M_i$ second stage units (*ssu*) and $M = \sum_{i=1}^{N} M_i$. Let $Y_i$, $X_i$ and $Z_i$ be the totals of $U_i$ in respect of the study variable $y$, and two auxiliary variables $x$ and $z$ respectively with corresponding overall totals $Y = \sum_{i=1}^{N} Y_i$, $X = \sum_{i=1}^{N} X_i$ and $Z = \sum_{i=1}^{N} Z_i$. To estimate $Y$, let us consider a general class of two-stage sampling designs: At stage one, a sample $s$ ($s \subset U$) of $n$ *fsu*s is drawn from $U$ according to any design with $\pi_i$ and $\pi_{ij}$ as the known first and second order inclusion probabilities. Then for every $i \in s$, a sample $s_i$ of $m_i$ *ssu*s is drawn from $U_i$ ($s_i \subset U_i$) with suitable selection probabilities at the second stage. More detailed accounts of two-stage sampling procedure are given in general survey sampling books (cf. Cochran [1], Sarndal *et al* [9]).

Let $E_1, E_2$ ($V_1, V_2; \mathrm{Cov}_1, \mathrm{Cov}_2$) denote the expectation (variance, covariance) operators over repeated sampling in the first and second stages; by $E$ ($V$ or Cov) we denote the overall expectation (variance or covariance). It is assumed that from the second

*Department of Statistics, Utkal University, Bhubaneswar 751004, India.
E-mail: (L. N. Sahoo, R. K. Sahoo) `lnsahoostatuu@rediffmail.com`
†Corresponding Author.
‡Department of Statistics, Ravenshaw College, Cuttack 753003, India.
E-mail: `scsenapati2002@rediffmail.com`
§Department of Statistics, R. D. Women's College, Bhubaneswar 751004, India.
E-mail: `akmangaraj@gmail.com`

stage sample $s_i$, $i \in s$, unbiased estimates $t_{iy}, t_{ix}$ and $t_{iz}$ respectively for $Y_i, X_i$ and $Z_i$ are available. Then, $V_2(t_{iy}) = \sigma_{iy}^2$, $V_2(t_{ix}) = \sigma_{ix}^2$, $V(t_{iz}) = \sigma_{iz}^2$, $\text{Cov}_2(t_{iy}, t_{ix}) = \sigma_{iyx}$, $\text{Cov}_2(t_{iy}, t_{iz}) = \sigma_{iyz}$, $\text{Cov}_2(t_{ix}, t_{iz}) = \sigma_{ixz}$.

Given the first stage sample $s$, we define $\pi$-estimators $t_y = \sum_{i \in s} \frac{t_{iy}}{\pi_i}$, $t_x = \sum_{i \in s} \frac{t_{ix}}{\pi_i}$ and $t_z = \sum_{i \in s} \frac{t_{iz}}{\pi_i}$ so that $\mathrm{E}(t_y) = Y$, $\mathrm{E}(t_x) = X$, $E(t_z) = Z$, $V(t_y) = \sigma_y^2 + \sum_{i=1}^N \frac{\sigma_{iy}^2}{\pi_i}$, $V(t_x) = \sigma_x^2 + \sum_{i=1}^N \frac{\sigma_{ix}^2}{\pi_i}$, $V(t_z) = \sigma_z^2 + \sum_{i=1}^N \frac{\sigma_{iz}^2}{\pi_i}$, $\text{Cov}(t_y, t_x) = \sigma_{yx} + \sum_{i=1}^N \frac{\sigma_{iyx}}{\pi_i}$, $\text{Cov}(t_y, t_z) = \sigma_{yz} + \sum_{i=1}^N \frac{\sigma_{iyz}}{\pi_i}$ and $\text{Cov}(t_x, t_z) = \sigma_{xz} + \sum_{i=1}^N \frac{\sigma_{ixz}}{\pi_i}$, where

$$\sigma_y^2 = \frac{1}{2} \sum_{i \neq j=1}^{N} (\pi_i \pi_j - \pi_{ij}) \left( \frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right)^2,$$

$$\sigma_{yx} = \frac{1}{2} \sum_{i \neq j=1}^{N} (\pi_i \pi_j - \pi_{ij}) \left( \frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j} \right) \left( \frac{X_i}{\pi_i} - \frac{X_j}{\pi_j} \right), \text{etc.}$$

We have seen that many two-stage sampling estimation techniques require advance knowledge on overall population totals $X$ and $Z$ . However, these techniques do not fully exhaust the information content of the auxiliary variables. In many surveys when the clusters are selected, it is more likely that the totals $X_i$ and $Z_i$ are known (or can be found easily or cheaply) for $i \in s$, *i.e.*, for the selected clusters. For instance, in a crop survey if $y, x$ and $z$ are respectively yield of the crop, area under the crop and area under cultivation, then information on total area under the crop $(X_i)$ and total area under cultivation $(Z_i)$ for the $i$th selected block (cluster of villages) can be obtained easily from the block records. Detail studies on the profitability of using knowledge on the auxiliary variable values at the level of first stage units, are provided by several authors, see, for example Sahoo [4], Sahoo and Panda ([5,6]), Sahoo and Sahoo [7], Sahoo *et al* [8], Hansen *et al* [2] and Smith [11]. In this context, we may also refer to Zheng and Little [14] who used penalized spline nonparametric regression models on the selection of probabilities of the first stage units, and Kim *et al* [3] who considered nonparametric regression estimation of the population total in which complete auxiliary information is available for the first stage units. Works of Singh *et al* [10], and Tracy and Singh [13] also focus on the use of various kinds of auxiliary information at different phases of a survey operation under a two-phase sampling set-up.

In this paper, we are aimed at constructing a general class of estimators for $Y$ with explicit involvement of $x$ and $z$ motivated by the assumption that $X$, $Z$, $X_i$ and $Z_i$, $i \in s$, are known.

## 2. The class of estimators

For given $s_i$, following the work of Srivastava [12] let us define a class of estimators for $Y_i$ by $\hat{Y}_i = g_i(t_{iy}, t_{ix}, t_{iz})$, $i \in s$, where $g_i(t_{iy}, t_{ix}, t_{iz})$ is a known function of $t_{iy}, t_{ix}$ and $t_{iz}$, which may depend on $X_i$ and $Z_i$ but is independent of $Y_i$ such that $g_i(t_{iy}, X_i, Z_i) = t_{iy}$, which implies that $g_i(Y_i, X_i, Z_i) = Y_i$. Also, for given $s$, let $t'_y = \sum_{i \in s} \frac{\hat{Y}_i}{\pi_i}$, $t'_x = \sum_{i \in s} \frac{X_i}{\pi_i}$, $t'_z = \sum_{i \in s} \frac{Z_i}{\pi_i}$, and $g(t'_y, t'_x, t'_z)$ be a function of $t'_y$, $t'_x$ and $t'_z$, which may depend on $X$ and $Z$ but is independent of $Y$, such that $g(t'_y, X, Z, ) = t'_y$, which implies that $g(Y, X, Z) = Y$. Further, following Srivastava [12], let us consider the following assumptions:

(a) $(t_{iy}, t_{ix}, t_{iz})$, $i \in s$, and $(t'_y, t'_x, t'_z)$ assume values in a bounded convex subspace $R_3$ of 3-dimensional real space containing the points $(Y_i, X_i, Z_i)$ and $(Y, X, Z)$, and

(b) The functions $g_i(t_{iy}, t_{ix}, t_{iz})$ and $g(t'_y, t'_x, t'_z)$ are continuous having first and second order partial derivatives *w.r.t.* their arguments which are also continuous in $R_3$.

Then, motivated by the work of Srivastava [12], we propose a class of estimators of $Y$ defined by

$$t_g = g(t'_y, t'_x, t'_z).$$

Here, $\hat{Y}_i = g_i(t_{iy}, t_{ix}, t_{iz})$ covers both linear and nonlinear functions of the statistics $t_{iy}, t_{ix}$ and $t_{iz}$. So, it is impossible to obtain an exact general expression for the conditional variance $V_2(\hat{Y}_i)$ due to the fact that the expectation operator is a linear operator. However, for simplicity we use Taylor linearization technique (cf. Sarndal *et al* [9]) to approximate $\hat{Y}_i$ by a more easily handled linear function, so that an approximate expression for $V_2(\hat{Y}_i)$ can be obtained. Hence, on considering the first order Taylor approximation of the function $g_i$ after expanding around the point $(Y_i, X_i, Z_i)$ and neglecting the remainder term, we obtain

$$\hat{Y}_i \approx Y_i + g_{i0}(t_{iy} - Y_i) + g_{i1}(t_{ix} - X_i) + g_{i2}(t_{iz} - Z_i), \tag{1}$$

where $g_{i0}, g_{i1}$ and $g_{i2}$ are respectively the first order differential coefficients of $g_i$ *w.r.t.* $t_{iy}, t_{ix}$ and $t_{iz}$, when evaluated at $(Y_i, X_i, Z_i)$. Hence, from (1) and noting that $g_{i0} = 1$, to a first order of approximation we have $\mathrm{E}_2(\hat{Y}_i) \approx Y_i$ and

$$V_2(\hat{Y}_i) \approx \sigma^2_{iy} + g^2_{i1}\sigma^2_{ix} + g^2_{i2}\sigma^2_{iz} + 2g_{i1}\sigma_{iyx} + 2g_{i2}\sigma_{iyz} + 2g_{i1}g_{i2}\sigma_{ixz}. \tag{2}$$

In light of the above discussion, once again using the linear approximation

$$t_g \approx Y + (t'_y - Y) + g_1(t'_x - X) + g_2(t'_z - Z), \tag{3}$$

the asymptotic variance of $t_g$ is obtained as

$$\begin{aligned} V(t_g) \approx V(t'_y) + g_1^2 V(t'_x) + g_2^2 V(t'_z) + 2g_1\mathrm{Cov}(t'_y, t'_x) + 2g_2\mathrm{Cov}(t'_y, t'_z) \\ + 2g_1g_2\mathrm{Cov}(t'_x, t'_z), \end{aligned} \tag{4}$$

where $g_1$ and $g_2$ are respectively the first derivatives of $g$ *w.r.t.* $t'_x$ and $t'_z$ around $(Y, X, Z)$.

Verifying that $V(t'_y) = V_1E_2(t'_y) + E_1V_2(t'_y)$, on simplification we get

$$V(t'_y) = \sigma^2_y + \sum_{i=1}^{N} V_2(\hat{Y}_i)\Big/\pi_i. \tag{5}$$

Similarly, under the conditional argument we also have $\mathrm{Cov}(t'_y, t'_x) = \sigma_{yx}$, $\mathrm{Cov}(t'_y, t'_z) = \sigma_{yz}$, $\mathrm{Cov}(t'_x, t'_z) = \sigma_{xz}$, $V(t'_x) = \sigma^2_x$ and $V(t'_z) = \sigma^2_z$. Finally, the formula for the asymptotic variance of $t_g$ is obtained as

$$\begin{aligned} V(t_g) = \sigma^2_y + g_1^2\sigma^2_x + g_2^2\sigma^2_z + 2g_1\sigma_{yx} + 2g_2\sigma_{yz} + 2g_1g_2\sigma_{xz} \\ + \sum_{i=1}^{N}\left(\sigma^2_{iy} + g^2_{i1}\sigma^2_{ix} + g^2_{i2}\sigma^2_{iz} + 2g_{i1}\sigma_{iyx} + 2g_{i2}\sigma_{iyz} + 2g_{i1}g_{i2}\sigma_{ixz}\right)\Big/\pi_i. \end{aligned} \tag{6}$$

This variance is minimized when

$$g_{i1} = -\frac{\beta_{iyx} - \beta_{iyz}\beta_{izx}}{1 - \beta_{izx}\beta_{ixz}} = -\hat{g}_{i1} \text{ (say)}, \quad g_{i2} = -\frac{\beta_{iyz} - \beta_{iyx}\beta_{ixz}}{1 - \beta_{izx}\beta_{ixz}} = -\hat{g}_{i2} \text{ (say)},$$

$$g_1 = -\frac{\beta_{yx} - \beta_{yz}\beta_{zx}}{1 - \beta_{zx}\beta_{xz}} = -\hat{g}_1 \text{ (say), and } g_2 = -\frac{\beta_{yz} - \beta_{yx}\beta_{xz}}{1 - \beta_{zx}\beta_{xz}} = -\hat{g}_2 \text{ (say)},$$

where $\beta_{iyx} = \sigma_{iyx}\big/\sigma^2_{ix}, \beta_{yx} = \sigma_{yx}\big/\sigma^2_x$, etc.

The optimum values of $g_{i1}, g_{i2}, g_1$ and $g_2$ determined are unique in the sense that they do not depend on each other for their computation. Using these optimum values in (6), we obtain a minimum asymptotic variance (which may be called the asymptotic minimum variance bound of the class) as

$$\min V(t_g) = \sigma_y^2(1-\rho^2) + \sum_{i=1}^{N} \sigma_{iy}^2(1-\rho_i^2)/\pi_i, \tag{7}$$

where $\rho_i^2 = \frac{\rho_{iyx}^2+\rho_{iyz}^2-2\rho_{iyx}\rho_{iyz}\rho_{ixz}}{1-\rho_{ixz}^2}$ and $\rho^2 = \frac{\rho_{yx}^2+\rho_{yz}^2-2\rho_{yx}\rho_{yz}\rho_{xz}}{1-\rho_{xz}^2}$ are such that $\rho_{iyx} = \sigma_{iyx}/\sigma_{iy}\sigma_{ix}$, $\rho_{yx} = \sigma_{yx}/\sigma_y\sigma_x$, etc. The estimator attaining this bound (which may be called a minimum variance bound (MVB) estimator) is a regression-type estimator defined by

$$t_{RG} = \sum_{i\in s} \left[t_{iy} - \hat{g}_{i1}(t_{ix}-X_i) - \hat{g}_{i2}(t_{iz}-Z_i)\right]/\pi_i - \hat{g}_1(t'_x - X) - \hat{g}_2(t'_z - Z).$$

## 3. Some specific cases of the class

If there is no use for $x$ and $z$, $\hat{Y}_i = t_{iy} \implies t_g = t_y$, the simple expansion estimator of $Y$. On the other hand, if the emphasis is laid on the use of either $x$ or $z$ or both, $t_g$ defines a wide class of estimators. For various choices of $g_i$ and $g$, it also reduces to many other classes. Let us now examine a few specific cases.

**3.1.** When the values of $X$ and $Z$ are not taken into consideration $t_g = t'_y$, producing a family of separate variety of estimators whose asymptotic variance structure is shown in (5). The minimum variance bound and the corresponding MVB estimator of the class are given by

$$\begin{aligned} \min V(t_g) &= \sigma_y^2 + \sum_{i=1}^{N} \sigma_{iy}^2(1-\rho_i^2)/\pi_i, \\ t_{RG}^{(s)} &= \sum_{i\in s} \left[t_{iy} - \hat{g}_{i1}(t_{ix}-X_i) - \hat{g}_{i2}(t_{iz}-Z_i)\right]/\pi_i. \end{aligned} \tag{8}$$

**3.2.** When $X$ is unknown but $X_i, Z_i$ for $i \in s$, and $Z$ are known, we have $\hat{Y}_i = g_i(t_{iy}, t_{ix}, t_{iz})$ and $t_g = g(t'_y, t'_z)$. Then the asymptotic MVB of the class is given by

$$\min V(t_g) = \sigma_y^2(1-\rho_{yz}^2) + \sum_{i=1}^{N} \sigma_{iy}^2(1-\rho_i^2)/\pi_i, \tag{9}$$

and the MVB estimator is defined by

$$t_{RG}^{(1)} = \sum_{i\in s} \left[t_{iy} - \hat{g}_{i1}(t_{ix}-X_i) - \hat{g}_{i2}(t_{iz}-Z_i)\right]/\pi_i - \beta_{yz}(t'_z - Z).$$

**3.3.** Assuming that $X_i (i \in s)$, $Z$ are known and $X$, $Z_i (i \in s)$ are unknown, and then defining $\hat{Y}_i = g_i(t_{iy}, t_{ix})$ and $t_g = g(t'_y, t_z)$, we see that the class of estimators considered by Sahoo and Panda [6] is a particular case of $t_g$. Here, the MVB of the class and the resulting MVB estimator are given by

$$\begin{aligned} \min V(t_g) &= \sigma_y^2 - \gamma^2 \left\{\sigma_z^2 + \sum_{i=1}^{N} \sigma_{iz}^2\left(1-\rho_{ixz}^2\right)/\pi_i\right\} + \sum_{i=1}^{N} \sigma_{iy}^2\left(1-\rho_{iyx}^2\right)/\pi_i, \\ t_{RG}^{(2)} &= \sum_{i\in s} \left[t_{iy} - \gamma_i\left(t_{ix}-X_i\right)\right]/\pi_i - \gamma\left(t_z - Z\right), \end{aligned} \tag{10}$$

where $\gamma_i = (\beta_{iyx} + \gamma\beta_{izx})$ and $\gamma = \frac{\sigma_{yz}+\sum_{i=1}^{N}\sigma_{iz}^2\left(\beta_{iyz}-\beta_{iyx}\beta_{ixz}\right)/\pi_i}{\sigma_z^2+\sum_{i=1}^{N}\sigma_{iz}^2\left(1-\rho_{ixz}^2\right)/\pi_i}$.

**3.4.** Assume that $Z$ is unknown but $X_i, Z_i$ for $i \in s$, and $X$ are known. Then, considering $\hat{Y}_i = g_i(t_{iy}, t_{iz})$ , the class of estimators is defined by $t_g = g(t'_y, t'_x)$ as studied by Sahoo and Sahoo [7]. In this case the minimum variance bound is

$$\min V(t_g) = \sigma_y^2(1-\rho_{yx}^2) + \sum_{i=1}^{N} \sigma_{iy}^2(1-\rho_{iyz}^2)/\pi_i, \tag{11}$$

and the corresponding MVB estimator is

$$t_{RG}^{(3)} = \sum_{i \in s} \left[t_{iy} - \beta_{iyz}(t_{iz} - Z_i)\right]/\pi_i - \beta_{yx}(t'_x - X).$$

**3.5.** If the estimation procedure is carried out with the involvement of $x$ only, then $\hat{Y}_i = g_i(t_{iy}, t_{ix})$ and $t_g = g(t'_y, t'_x)$, a class of estimators considered by Sahoo and Panda [5]. The asymptotic MVB of the class is

$$\min V(t_g) = \sigma_y^2(1-\rho_{yx}^2) + \sum_{i=1}^{N} \sigma_{iy}^2(1-\rho_{iyx}^2)/\pi_i, \tag{12}$$

and the corresponding MVB estimator is of the form

$$t_R^{(4)} = \sum_{i \in s} \left[t_{iy} - \beta_{iyx}(t_{ix} - X_i)\right]/\pi_i - \beta_{yx}(t'_x - X),$$

which can be transformed to the regression-type estimator developed by Sahoo [4] on adopting the simple random sampling without replacement (SRSWOR) design at different stages.

**3.6.** If $s = U$, then $\pi_i = \pi_{ij} = 1$ for all $i$ and $j$ . In this case $t_g$ simply defines a class of estimators for a stratified sampling with $N$ *fsu*s as a set of strata.

**3.7.** If $s_i = U_i \forall\, i$, $t_g$ defines a class of estimators for single-stage cluster sampling.

## 4. Precision of the class

In order to study precision of $t_g$ compared to other classes of estimators utilizing information on two auxiliary variables, let us now consider the classes of estimators developed by Srivastava [12] and Sahoo *et al*, [8]. These classes are respectively defined by

$$t_c = h(t_y, t_x, t_z)$$

and

$$t_l = f(\tilde{Y}, \tilde{X}),$$

where $\tilde{Y} = \sum_{i \in s} \phi_i(t_{iy}, t_{ix})/\pi_i$ and $\tilde{X} = \phi(t'_x, t'_z)$, are such that the functions involved in composing the classes admit regularity conditions. It may be remarked here that $t_l$ makes use of pre-assigned values of $X, Z, X_i$ and $Z_i$ $(i \in s)$, whereas $t_c$ makes use of only $X$ and $Z$.

The asymptotic expressions for $V(t_c)$ and $V(t_l)$ are given by

$$V(t_c) = \sigma_y^2 + h_1^2\sigma_x^2 + h_2^2\sigma_z^2 + 2h_1\sigma_{yx} + 2h_2\sigma_{yz} + 2h_1h_2\sigma_{xz} + \sum_{i=1}^{N}\left[\sigma_{iy}^2 + h_1^2\sigma_{ix}^2 + h_2^2\sigma_{iz}^2 + 2h_1\sigma_{iyx} + 2h_2\sigma_{iyz} + 2h_1h_2\sigma_{ixz}\right]/\pi_i, \tag{13}$$

$$V(t_l) = \sigma_y^2 + f_1^2\left(\sigma_x^2 + \phi_2^2\sigma_z^2 + 2\phi_2\sigma_{xz}\right) + 2f_1\left(\sigma_{yx} + \phi_2\sigma_{yz}\right) + \sum_{i=1}^{N}\left(\sigma_{iy}^2 + \phi_{i1}^2\sigma_{ix}^2 + 2\phi_{i1}\sigma_{iyx}\right)/\pi_i, \tag{14}$$

where $h_1 = \left.\frac{\partial h(t_y,t_x,t_z)}{\partial t_x}\right|_{(Y,X,Z)}$, $h_2 = \left.\frac{\partial h(t_y,t_x,t_z)}{\partial t_z}\right|_{(Y,X,Z)}$, $\phi_{i1} = \left.\frac{\partial \phi_i(t_{iy},t_{ix})}{\partial t_{ix}}\right|_{(Y_i,X_i)}$, $\phi_2 = \left.\frac{\partial \phi(t'_x,t'_z)}{\partial t'_z}\right|_{(X,Z)}$ and $f_1 = \left.\frac{\partial f(\tilde{Y},\tilde{X})}{\partial \tilde{X}}\right|_{(Y,X)}$.

The MVB and the corresponding MVB estimators of $t_c$ and $t_l$ are

$$\min V(t_c) = \sigma_y^2(1-R^2) + \sum_{i=1}^{N}\sigma_{iy}^2(1-R^2)/\pi_i, \tag{15}$$

$$\min V(t_l) = \sigma_y^2(1-\rho^2) + \sum_{i=1}^{N}\sigma_{iy}^2(1-\rho_{iyx}^2)/\pi_i, \tag{16}$$

$$t_{RG}^{(c)} = t_y - \beta_1(t_x - X) - \beta_2(t_z - Z),$$
$$t_{RG}^{(l)} = \sum_{i\in s}\left[t_{iy} - \beta_{iyx}(t_{ix} - X_i)\right]/\pi_i - \hat{f}_1(t'_x - X) - \hat{\phi}_2(t'_z - Z),$$

where $R$ is the multiple correlation coefficient of $t_y$ on $t_x$ and $t_z$; $\beta_1$ and $\beta_2$ are respectively the partial regression coefficients of $t_y$ on $t_x$ and $t_y$ on $t_z$; $\hat{f}_1 = \hat{g}_1$, $\hat{\phi}_2 = \frac{\beta_{yz} - \beta_{yx}\beta_{xz}}{\beta_{yx} - \beta_{yz}\beta_{xz}}$.

As $t_c$ and $t_l$ can be taken to be potential competitors of $t_g$, one should naturally be interested to compare their precisions. But, comparing (6) with (13) and (14), we can derive only some sufficient conditions under which an estimator of $t_g$ is asymptotically more precise than an estimator of $t_c$ or $t_l$. However, these conditions are extremely complicated and mainly depend on the choices of the functions, $h, f, \phi, \phi_i, g$ and $g_i$, and cannot lead to any straightforward conclusion unless the nature of these functions are known. But, for simplicity, if we accept MVB as an intrinsic measure of the precision of a class, the problem of precision comparison seems to be easier and our attention will be concentrated on the MVB estimators only. Thus,

- $\min V(t_g) \le \min V(t_c)$ *i.e.*, $t_{RG}$ is more precise than $t_{RG}^{(c)}$ if $R \le \rho$ and $\rho_i\ \forall i$, and,
- $\min V(t_g) \le \min V(t_l)$ *i.e.*, $t_{RG}$ is always more precise than $t_{RG}^{(l)}$.

On these grounds, we also find that $t_{RG}$ is more precise than $t_{RG}^{(s)}$, $t_{RG}^{(1)}$, $t_{RG}^{(3)}$ and $t_{RG}^{(4)}$, whereas no conclusion can be drawn regarding the precision of $t_{RG}$ over $t_{RG}^{(2)}$.

## 5. A simulation study

As seen above, a theoretical comparison is not very useful in showing the merits of the suggested estimation procedure over others. Therefore, as a counterpart to the theoretical comparison, we carry out a simulation study. In this study, we do not limit ourselves to the MVB estimators only.

The simulation study reported here involves repeated draws of independent samples from a natural population consisting of 198 blocks (*ssu*s) divided into $N = 27$ wards (*fsu*s) of Berhampur City of Orissa (India). The number of blocks ($M_i$) in the 27 wards are 6, 6, 12, 5, 6, 6, 10, 5, 6, 6, 6, 6, 6, 12, 6, 7, 7, 7, 10, 6, 6, 7, 10, 11, 9, 8 and 6. Three variables *viz.*, number of educated females, number of households and the female population are used as $y, x$ and $z$ respectively, data on which is readily available from the Census of India (1971) document.

The estimators under consideration are the eight MVB estimators *viz.*, $t_{RG}^{(s)}$, $t_{RG}^{(1)}$, $t_{RG}^{(2)}$, $t_{RG}^{(3)}$, $t_{RG}^{(4)}$, $t_{RG}^{(c)}$, $t_{RG}^{(l)}$ and $t_{RG}$, and their respective ratio counterparts defined by

$$t_R^{(s)} = \sum_{i\in s} t_{iy}\frac{X_i}{t_{ix}}\frac{Z_i}{t_{iz}}\Big/\pi_i, \qquad t_R^{(1)} = \left(\sum_{i\in s} t_{iy}\frac{X_i}{t_{ix}}\frac{Z_i}{t_{iz}}\Big/\pi_i\right)\frac{Z}{t'_z},$$

$$t_R^{(2)} = \left(\sum_{i\in s} t_{iy}\frac{X_i}{t_{ix}}\Big/\pi_i\right)\frac{Z}{t_z}, \qquad t_R^{(3)} = \left(\sum_{i\in s} t_{iy}\frac{Z_i}{t_{iz}}\Big/\pi_i\right)\frac{X}{t'_x},$$

$$t_R^{(4)} = \left(\sum_{i\in s} t_{iy}\frac{X_i}{t_{ix}}\Big/\pi_i\right)\frac{X}{t'_x}, \qquad t_R^{(c)} = t_y\frac{X}{t_x}\frac{Z}{t_z},$$

$$t_R^{(l)} = \left(\sum_{i\in s} t_{iy}\frac{X_i}{t_{ix}}\Big/\pi_i\right)\frac{X}{t'_x}\frac{Z}{t'_z}, \quad \text{and} \quad t_R = \left(\sum_{i\in s} t_{iy}\frac{X_i}{t_{ix}}\frac{Z_i}{t_{iz}}\Big/\pi_i\right)\frac{X}{t'_x}\frac{Z}{t'_z}.$$

We did not touch the product or product-type estimators as $y$ is positively correlated with $x$ and $z$. It may also be noted here that, for simplicity, we compute $t_{RG}$ and the other MVB estimators by considering population values of their respective coefficients, and in this case the estimators are unbiased.

The following performance measures of an estimator $t$ are taken into consideration:

(1) Relative absolute bias (RAB) $= 100\,|B(t)|\,/Y$, where $B(t) = \mathrm{E}(t) - Y$ is the bias of $t$.
(2) Percentage relative efficiency (PRE) compared to the direct estimator $t_y = \sum_{i\in s}\frac{t_{iy}}{\pi_i}$ i.e., PRE $= 100V(t_y)/V(t)$, where $V(t)$ is the variance of $t$.

Our simulation consisted in the selection of 1000 independent first stage samples each of size $n = 10$ *fsu*s from the population by SRSWOR. From every selected $U_i$, $(i = 1, 2, \ldots, 10)$ in a first stage sample, a second stage sample of size $m_i = 2$ or 3 *ssu*s is again selected by SRSWOR. Thus, we now have 1000 independent samples each of size 20 or 30 *ssu*s. Considering these independent samples simulated biases and variances of the comparable estimators are calculated. If $r$ indexes the $r$-th sample, the simulated bias and variance of an estimator $t$ are given by

$$B(t) = \frac{1}{1000}\sum_{r=1}^{1000} t^{(r)} - Y$$

and

$$V(t) = \frac{1}{1000}\sum_{r=1}^{1000}\left(t^{(r)} - \frac{1}{1000}\sum_{r=1}^{1000} t^{(r)}\right)^2,$$

respectively, where $t^{(r)}$ is the value of $t$ for the $r$-th realized sample. The simulated values of RAB and PRE of different estimators are then calculated as suggested above and their values are displayed in Table 1. But, we see that the simulated values of RAB for the MVB estimators are not equal to zero as these values are computed from a limited number of independent samples.

**Table 1. RAB and PRE of Different Estimators**

| Estimator | RB | | PRE | |
|---|---|---|---|---|
| | $m_i = 2$ | $m_i = 3$ | $m_i = 2$ | $m_i = 3$ |
| $t_{RG}^{(s)}$ | 4.875 | 4.177 | 101 | 105 |
| $t_{RG}^{(1)}$ | 2.965 | 2.390 | 109 | 110 |
| $t_{RG}^{(2)}$ | 3.501 | 3.188 | 120 | 121 |
| $t_{RG}^{(3)}$ | 6.347 | 5.910 | 109 | 111 |
| $t_{RG}^{(4)}$ | 2.956 | 2.163 | 107 | 108 |
| $t_{RG}^{(c)}$ | 5.885 | 4.632 | 118 | 119 |
| $t_{RG}^{(l)}$ | 2.122 | 2.119 | 116 | 118 |
| $t_{RG}$ | 2.136 | 2.042 | 121 | 123 |
| $t_{R}^{(s)}$ | 49.991 | 33.135 | 105 | 105 |
| $t_{R}^{(1)}$ | 61.116 | 55.246 | 108 | 109 |
| $t_{R}^{(2)}$ | 58.123 | 41.765 | 113 | 115 |
| $t_{R}^{(3)}$ | 55.448 | 46.654 | 110 | 112 |
| $t_{R}^{(4)}$ | 29.567 | 28.769 | 109 | 112 |
| $t_{R}^{(c)}$ | 62.987 | 54.944 | 107 | 111 |
| $t_{R}^{(l)}$ | 27.576 | 26.323 | 117 | 118 |
| $t_R$ | 25.653 | 22.444 | 119 | 120 |

Simulation results show that $t_R$ is superior to the other ratio-type estimators on the grounds of RAB and PRE. On the other hand, $t_{RG}$ is superior to the other regression-type (MVB) estimators with respect to RAB for $m_i = 3$ and PRE. But, it is slightly inferior to $t_{RG}^{(l)}$ with respect to RAB for $m_i = 2$. This imperfection is probably caused by our restriction to 1000 independent samples. An increase in the number of samples and their size may improve the degree of performance of $t_{RG}$ considerably. However, our simulation study, though of limited scope, clearly indicates that there are practical situations which can favor the application of the suggested estimation methodology.

**Acknowledgement**

The authors are grateful to the referees whose constructive comments led to an improvement in the paper.

## References

[1] Cochran, W. G. *Sampling Techniques*, 3rd edition (John Wiley & Sons, New York, 1977).

[2] Hansen, M. H., Hurwitz, W. N. and Madow, W. G. *Sampling Survey Methods and Theory* vol.I (John Wiley & Sons, New York, 1953).

[3] Kim, J.-Y., Breidt, F. J. and Opsomer, J. D. *Nonparametric regression estimation of finite population totals under two-stage sampling* (Technical Report **4**, Department of Statistics, Colorado State University, 2009).

[4] Sahoo, L. N. *A regression-type estimator in two-stage sampling*, Calcutta Statistical Association Bulletin **36**, 97–100, 1987.

[5] Sahoo, L. N. and Panda, P. *A class of estimators in two-stage sampling with varying probabilities*, South African Statistical Journal **31**, 151–160, 1997.

[6] Sahoo, L. N. and Panda, P. *A class of estimators using auxiliary information in two-stage sampling*, Australian & New Zealand Journal of Statistics **41**, 405–410, 1999.

[7] Sahoo, L. N. and Sahoo, R. K. *On the construction of a class of estimators for two-stage sampling*, Journal of Applied Statistical Science **14**, 317–322, 2005.

[8] Sahoo, L. N., Senapati, S. C. and Singh, G. N. *An alternative class of estimators in two-stage sampling with two auxiliary variables*, Journal of the Indian Statistical Association **43**, 147–156, 2005.

[9] Sarndal, C. E., Swensson, B. and Wretman, J. *Model Assisted Survey Sampling* (Springer-Verlag, New York, 1992).

[10] Singh, V. K., Singh, Hari P. and Singh, Housila, P. *Estimation of ratio and product of two finite population means in two-phase sampling*, Journal of Statistical Planning and Inferences **41**, 163–171, 1994.

[11] Smith, T. M. F. *A note on ratio estimates in multi-stage sampling*, Journal of the Royal Statistical Society **A132**, 426–430, 1969.

[12] Srivastava, S. K. *A class of estimators using auxiliary information in sample surveys*, Canadian Journal of Statistics **8**, 253–254, 1980.

[13] Tracy, D. S. and Singh, H. P. *A general class of chain regression estimators in two-phase sampling*, Journal of Applied Statistical Science **8**, 205–216, 1999.

[14] Zheng, H. and Little, R. J. *Penalized spline nonparametric mixed models for inference about finite population mean from two-stage sampling*, Survey Methodology **30**, 209–218, 2004.