International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2019; 4(2): 01-06

www.mathsjournal.com
Received: 01-01-2019
Accepted: 04-02-2019

NR Das
Research Scholar, Department of Statistics Utkal University, Bhubaneswar, Odisha, India

LN Sahoo
Visiting Faculty, Institute of Mathematics and Applications Andharua, Bhubaneswar, Odisha, India

Correspondence
NR Das
Research Scholar, Department of Statistics Utkal University, Bhubaneswar, Odisha, India

A class of estimators using auxiliary information in two-stage sampling for two phases of sample selection

NR Das and LN Sahoo

Abstract
In this paper, we compose a class of estimators for the mean of a finite population using information on an auxiliary variable in a two-stage sampling when samples at each stage are drawn under two phase selections. The novel feature of the class is that it makes an attempt to gather available auxiliary information both at primary and secondary levels.

Keywords: Auxiliary variable, regression estimator, two-phase sampling, two-stage sampling

1. Introduction

Consider a finite population U, partitioned into N clusters called primary sampling units (PSU) denoted by $U_1, U_2, \ldots, U_N$ such that the number of secondary sampling units (SSU) in U_i is M_i and $M = \sum_{i=1}^{N} M_i$. Let y_{ij} and x_{ij} respectively be the values of the study variable y and an auxiliary variable x for the jth SSU of U_i $(j = 1, 2, \ldots, M_i; i = 1, 2, \ldots, N)$. Define

$$Y_i = \sum_{j=1}^{M_i} y_{ij},\ X_i = \sum_{j=1}^{M_i} x_{ij} \text{ and } \bar{Y}_i = \frac{Y_i}{M_i},\ \bar{X}_i = \frac{X_i}{M_i}$$

As the totals and means of U_i; and

$$Y = \sum_{i=1}^{N} Y_i,\ X = \sum_{i=1}^{N} X_i \text{ and } \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} u_i \bar{Y}_i,\ \bar{X} = \frac{1}{N}\sum_{i=1}^{N} u_i \bar{X}_i$$

As the overall totals and means of U, where $u_i = \frac{NM_i}{M}$.

To estimate $\bar{Y}$, we consider a two-stage sampling design. At stage one, a sample s of n PSUs is drawn from U. Then, for every $i \in s$, a sample s_i of m_i SSUs is drawn from U_i at the second stage. Assume that units at each stage are sampled according to the design simple random sampling without replacement (SRSWOR). Let us define the following sample means:

$$\bar{y}_i = \frac{1}{m_i}\sum_{j\in s_i} y_{ij},\ \bar{x}_i = \frac{1}{m_i}\sum_{j\in s_i} x_{ij},\ \bar{y} = \frac{1}{n}\sum_{i\in s} u_i\bar{y}_i,\ \bar{x} = \frac{1}{n}\sum_{i\in s} u_i\bar{x}_i \text{ and } \bar{x}' = \frac{1}{n}\sum_{i\in s} u_i\bar{X}_i.$$

Sampling theory provides many successful attempts on the use of auxiliary information about the population for improving precision of the estimates under a given sampling design. However, precision of an estimator in a two-stage sampling can be improved considerably when the kind and extent of available auxiliary information at different stages/levels are effectively utilized. According to the nature of the available auxiliary information, three different cases can be taken into consideration *viz.*, (i) The values of X_i or $\bar{X}_i$ *i.e.*, the cluster totals or means of the auxiliary variable is available for all PSUs in U so that a correct value of X or $\bar{X}$ is available, (ii) The values of x *i.e.*, x_{ij}'s can be observed for all SSUs in U so that a correct value of X or $\bar{X}$ is available, and (iii) The values of x can be observed for SSUs in $U_i, i \in s$, *i.e.*, for the selected SSUs so that a correct value of X_i or $\bar{X}_i$ for $i \in s$ can be available [see, for example, Sarndal, Swensson and Wretman (1992)] [13].

Many two-stage sampling estimation techniques consider case (ii) and use advance knowledge on overall population mean or total of the auxiliary variable. But these techniques not only need up-to-date lists of all SSUs of the entire population but also do not fully exhaust the information contents of the auxiliary variable. Cases (i) and (iii) have many practical applications because they ordinarily use auxiliary information that are easily obtained at the two different stages of the survey operation. Combination of cases (i) and (iii) are also considered in many traditional texts [*cf*., Hansen, Hurwitz, and Madow (1953) [2], Murthy (1977)] [4] and journal articles [*cf.*, Smith (1969) [14], Sahoo and Swain (1983) [11], Sahoo (1987) [5], Zheng and Little (2004) [16], Kim *et al*. (2009)] [3]. Because, in many surveys cluster means or totals of x for the selected clusters are usually known or can be known easily or cheaply. Therefore, a survey statistician can take the advantage of utilizing this type of more extensive available information while developing efficient estimation techniques. As an example to indicate how such a situation arises, we may refer to a crop survey conducted to estimate the yield rate of a crop in a district with blocks (cluster of villages) as the PSUs and villages as the SSUs. If y and x represent respectively yield of the crop and area under cultivation, then information on mean area under cultivation $(\bar{X}_i)$ for the ith selected block can be obtained easily from the block records and information on the mean cultivated area for the entire district *i.e*., $\bar{X}$ can be easily available from the district records. With the spirit of utilizing such type of extensive auxiliary information on x, Sahoo and Panda (1997) [7] and Sahoo *et al*. (2009) [6] formulated some efficient estimation methods under two-stage sampling. Similar problems, with the involvement of two auxiliary variables, are also considered in Sahoo and Panda (1999) [8], Sahoo *et al*. (2005, 2011) [9, 10].

Sahoo and Panda (1997) [7] developed a class of estimators that needs advance knowledge on the overall population mean $\bar{X}$ as well as the cluster means $\bar{X}_i$, $i\epsilon s$. But, in many surveys, no such extensive information is available at the outset. The common procedure in such cases is to use a two-phase sampling or sampling followed by sub-sampling. The main objective of this paper is to construct and study a general class of estimators in two-phase sampling scheme for estimating the population mean $\bar{Y}$ when one fails to gather prior knowledge on $\bar{X}_i$ and $\bar{X}$.

2. Two-phase sampling procedure

Let us consider the following two-phase sampling mechanism under the assumption that sampling at each phase is also done by SRSWOR:

a) In the first phase, a sample s' $(s' \subset U)$ of n' PSUs is drawn out of N in the first stage and a sample s_i' $(s_i' \subset U_i)$ of m_i' SSUs is drawn from M_i SSUs of $U_i, i \in s'$. The sample so selected consists of $\sum_{i=1}^{n'} m_i'$ SSUs used to gather information on x.
b) In the second phase, a sub-sample s $(s \subset s')$ of n PSUs is selected out of n' PSUs selected in the first phase sample s' and then in $U_i, i \in s$, a sub-sample s_i $(s_i \subset s_i')$ of m_i SSUs is selected out of the m_i' SSUs selected in the first phase sample s_i'. The study variable y is then observed for the SSUs selected in the second phase sample.

We now define the following sample means in the context of two-phase sampling:

$$\bar{x}_{di} = \frac{1}{m_i'}\sum_{j\epsilon s_i'} x_{ij}\,, \bar{x}_d = \frac{1}{n}\sum_{i\epsilon s} u_i\,\bar{x}_{di} \text{ and } \bar{x}_d' = \frac{1}{n'}\sum_{i\epsilon s'} u_i\,\bar{x}_{di}.$$

3. The proposed class of estimators

For ith selected PSU U_i, $i \in s'$, and for given s_i' and s_i $(s_i \subset s_i' \subset U_i)$, consider a class of estimators for $\bar{Y}_i$ defined by $\ell_i = g_i(\bar{y}_i, \bar{x}_i, \bar{x}_{di})$, where $g_i(\bar{y}_i, \bar{x}_i, \bar{x}_{di})$ is a known function of $\bar{y}_i$, $\bar{x}_i$ and $\bar{x}_{di}$ such that $g_i(\bar{Y}_i, \bar{X}_i, \bar{X}_i) = \bar{Y}_i$ and admits the following regularity conditions:

a) $(\bar{y}_i, \bar{x}_i, \bar{x}_{di})$ is considered as a point in a bounded and closed convex subspace $\mathcal{S}$ of the three-dimensional real space containing the point $(\bar{Y}_i, \bar{X}_i, \bar{X}_i)$, and
b) The function $g_i(\bar{y}_i, \bar{x}_i, \bar{x}_{di})$ is continuous having the first and second order partial derivatives which are also continuous in $\mathcal{S}$.

Further, for given s' and s $(s \subset s' \subset U)$, define $\ell_s = \frac{1}{n}\sum_{i\epsilon s} u_i\,\ell_i$, and let $g(\ell_s, \bar{x}_d, \bar{x}_d')$ be a known function of $\ell_s, \bar{x}_d$ and $\bar{x}_d'$ such that $g(\bar{Y}, \bar{X}, \bar{X}) = \bar{Y}$ and admits regularity conditions in $\mathcal{S}$. Then, the proposed class of estimators of $\bar{Y}$ is defined by $\ell_g = g(\ell_s, \bar{x}_d, \bar{x}_d')$.

Here $\ell_i = g_i(\bar{y}_i, \bar{x}_i, \bar{x}_{di})$ may be a linear or a nonlinear function of $\bar{y}_i, \bar{x}_i$ and $\bar{x}_{di}$ within the framework of the mentioned regularity conditions. Therefore, for analyzing properties of ℓ_i in a simpler way, we use the Taylor linearization technique [*cf*., Sarndal, Swensson and Wretman (1992)] [13] to approximate ℓ_i by a more easily handled linear function so that approximate expressions for the mean and variance, under the sampling design in cluster i, can be easily found. Hence, considering the first order Taylor approximation of $g_i(\bar{y}_i, \bar{x}_i, \bar{x}_{di})$, expanding around the point $(\bar{Y}_i, \bar{X}_i, \bar{X}_i)$ and neglecting the remainder term, we obtain

$$\ell_i \approx \bar{Y}_i + g_{i0}(\bar{y}_i - \bar{Y}_i) + g_{i1}(\bar{x}_i - \bar{X}_i) + g_{i2}(\bar{x}_{di} - \bar{X}_i) \quad (1)$$

Where g_{i0}, g_{i1} and g_{i2} are respectively the first order partial derivatives of g_i with respect to $\bar{y}_i, \bar{x}_i$ and $\bar{x}_{di}$ about $(\bar{Y}_i, \bar{X}_i, \bar{X}_i)$.

Let E_1, E_2 $(V_1, V_2;\ Cov_1, Cov_2)$ denote the expectation (variance, covariance) operators over repeated sampling in the first and second stages respectively; by E (V or Cov) we denote the overall expectation (variance or covariance). Thus, noting that $g_{i0} = 1$ and $g_{i1} = -g_{i2}$, to a first order of approximation, we have $E_2(\ell_i) \approx \bar{Y}_i$, and $V_2(\ell_i) = V_2(\bar{y}_i) + g_{i1}^2[V_2(\bar{x}_i) + V_2(\bar{x}_{di}) - 2Cov_2(\bar{x}_i, \bar{x}_{di})]$

$$+2g_{i1}[Cov_2(\bar{y}_i, \bar{x}_i) - Cov_2(\bar{y}_i, \bar{x}_{di})] \quad (2)$$

After a considerable simplification, we get

$$V_2(\ell_i) = \left(\frac{1-\emptyset_i}{m_i}\right) S_{iy}^2 + \left(\frac{1-\gamma_i}{m_i}\right) g_{i1}\left(g_{i1} S_{ix}^2 + 2S_{iyx}\right) \quad (3)$$

Where
$\emptyset_i = \frac{m_i}{M_i}$, $\gamma_i = \frac{m_i}{m_i'}$, $S_{iy}^2 = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}\left(y_{ij} - \bar{Y}_i\right)^2$, $S_{ix}^2 = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}\left(x_{ij} - \bar{X}_i\right)^2$ and
$S_{iyx} = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}\left(y_{ij} - \bar{Y}_i\right)\left(x_{ij} - \bar{X}_i\right)$.

Let g_0, g_1 and g_2 be respectively the first order derivatives of the function g with respect to $\ell_s, \bar{x}_d$ and $\bar{x}_d'$ around $(\bar{Y}, \bar{X}, \bar{X})$. For simplicity, once again considering first-order Taylor linearization and noting that $g_0 = 1$ and $g_1 = -g_2$, ℓ_g is approximated by the following serviceable linear form:

$$\ell_g \approx \bar{Y} + (\ell_s - \bar{Y}) + g_1(\bar{x}_d - \bar{x}_d') \quad (4)$$

This linearized estimator ℓ_g looks mysterious, because of that g is a function of $\ell_s, \bar{x}_d$ and $\bar{x}_d'$, and on the other hand in the light of (1), ℓ_s is also a function of $\bar{x}_d$. Hence, ℓ_g is a composite function whose approximate variance is given by

$$V(\ell_g) = V(\ell_s) + g_1^2[V(\bar{x}_d) + V(\bar{x}_d') - 2Cov(\bar{x}_d, \bar{x}_d')] + 2g_1[Cov(\ell_s, \bar{x}_d) - Cov(\ell_s, \bar{x}_d')] \quad (5)$$

Now, we have
$V(\ell_s) = V_1E_2(\ell_s) + E_1V_2(\ell_s)$,
$Cov(\ell_s, \bar{x}_d') = Cov_1\{E_2(\ell_s), E_2(\bar{x}_d')\} + E_1Cov_2(\ell_s, \bar{x}_d')$ Etc.

Using these conditional operators, we find, after simplification and rearrangement of terms,

$$V(\ell_s) = \left(\frac{1-\emptyset}{n}\right) S_{by}^2 + \frac{1}{nN}\sum_{i=1}^{N} u_i^2 \left[\left(\frac{1-\emptyset_i}{m_i}\right) S_{iy}^2 + \left(\frac{1-\gamma_i}{m_i}\right) g_{i1}\left(g_{i1} S_{ix}^2 + 2S_{iyx}\right)\right] \quad (6)$$

$$V(\bar{x}_d) + V(\bar{x}_d') - 2Cov(\bar{x}_d, \bar{x}_d') = \left(\frac{1-\gamma}{n}\right)\left[S_{bx}^2 + \frac{1}{N}\sum_{i=1}^{N} u_i^2 \left(\frac{1-\emptyset_i'}{m_i'}\right) S_{ix}^2\right] \quad (7)$$

And $Cov(\ell_s, \bar{x}_d) - Cov(\ell_s, \bar{x}_d') = \left(\frac{1-\gamma}{n}\right)\left[S_{byx} + \frac{1}{N}\sum_{i=1}^{N} u_i^2 \left(\frac{1-\emptyset_i'}{m_i'}\right) S_{iyx}\right]$ (8)

Where
$\emptyset = \frac{n}{N}$, $\gamma = \frac{n}{n'}$, $\emptyset_i' = \frac{m_i'}{M_i}$, $S_{by}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(u_i\bar{Y}_i - \bar{Y})^2$, $S_{bx}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(u_i\bar{X}_i - \bar{X})^2$ and
$S_{byx} = \frac{1}{N-1}\sum_{i=1}^{N}(u_i\bar{Y}_i - \bar{Y})(u_i\bar{X}_i - \bar{X})$.

Finally, we derive formula for the approximate variance of ℓ_g (in a compact form) as

$$V(\ell_g) = V(\bar{y}) + \frac{1-\gamma}{n}\Delta_x^2\, g_1(g_1 + 2\beta) + \frac{1}{nN}\sum_{i=1}^{N} \nu_i\, S_{ix}^2 g_{i1}\left(g_{i1} + 2\beta_{iyx}\right) \quad (9)$$

Where

$$V(\bar{y}) = \frac{1-\emptyset}{n} S_{by}^2 + \frac{1}{nN}\sum_{i=1}^{N} u_i^2 \frac{1-\emptyset_i}{m_i} S_{iy}^2 \quad (10)$$

is the variance of $\bar{y}$; $\nu_i = u_i^2 \frac{1-\gamma_i}{m_i}$, $\eta_i = u_i^2 \frac{1-\emptyset_i'}{m_i'}$, $\Delta_x^2 = S_{bx}^2 + \frac{1}{N}\sum_{i=1}^{N}\eta_i S_{ix}^2$, $\Delta_{yx} = S_{byx} + \frac{1}{N}\sum_{i=1}^{N}\eta_i S_{iyx}$, $\beta = \Delta_{yx}/\Delta_x^2$ and $\beta_{iyx} = S_{iyx}/S_{ix}^2$ is the regression coefficient of y on x in U_i.
Minimization of $V(\ell_g)$ in (9) with respect to $g_1$and g_{i1} leads to their unique optimum values as $g_1^{(opt)} = -\beta$ and $g_{i1}^{(opt)} = -\beta_{iyx}$. When these optimum values are inserted into (9), we obtain a minimum approximate variance (which may be called approximate minimum variance bound (MVB) of the class) as

$$\min V(\ell_g) = V(\bar{y}) - \frac{1-\gamma}{n}\Delta_y^2\rho^2 - \frac{1}{nN}\sum_{i=1}^{N}\nu_i S_{iy}^2\rho_{iyx}^2, \quad (11)$$

Where
$\Delta_y^2 = S_{by}^2 + \frac{1}{N}\sum_{i=1}^{N}\eta_i S_{iy}^2$, $\rho = \Delta_{yx}/\Delta_y\Delta_x$ and $\rho_{iyx} = S_{iyx}/S_{iy}S_{ix}$ is the correlation coefficient between y and x in U_i. The estimator attaining this bound (which may be called an MVB estimator) is a regression-type estimator of the form

$$\ell_{RG} = \frac{1}{n}\sum_{i\epsilon s} u_i\left[\bar{y}_i - \beta_{iyx}(\bar{x}_i - \bar{x}_{di})\right] - \beta(\bar{x}_d - \bar{x}_d').$$

4. Some particular cases of ℓ_g

4.1 If the use of x is not taken into consideration, then $\ell_i = \bar{y}_i \Longrightarrow \ell_g = \bar{y}$, the simple expansion estimator of $\bar{Y}$. On the other hand, when emphasis is given on the use of x under the allowable sampling model, ℓ_g produces an infinite number of estimators (ratio, product and regression types) for proper selections of the functions g_i and g. For instance, the following estimators are three simple noteworthy cases of ℓ_g:

$\ell_R = \left(\frac{1}{n}\sum_{i\epsilon s} u_i \bar{y}_i \frac{\bar{x}_{di}}{\bar{x}_i}\right)\frac{\bar{x}'_d}{\bar{x}_d}$, $\ell_{RG1} = \frac{1}{n}\sum_{i\epsilon s} u_i \left[\bar{y}_i - \beta_{iyx}(\bar{x}_i - \bar{x}_{di})\right] - \beta_{byx}(\bar{x}_d - \bar{x}'_d)$
and $\ell_P = \left(\frac{1}{n}\sum_{i\epsilon s} u_i \bar{y}_i \frac{\bar{x}_i}{\bar{x}_{di}}\right)\frac{\bar{x}_d}{\bar{x}'_d}$,

Where $\beta_{byx} = S_{byx}/S_{bx}^2$. One can easily verify that the estimators ℓ_R, $\ell_{RG1}$and ℓ_P are two-phase sampling extensions of the chain ratio, chain regression and chain product estimators of Murthy (1977) [4], Sahoo (1987) [5], and Sahoo and Swain (1986) [12]. Das and Sahoo (2015) [1] also evaluated the performance of ℓ_{RG1} relative to other regression-type estimators.

4.2 When the available auxiliary information within the selected PSUs only is taken into consideration (*i.e.*, only case (iii) is considered), then ℓ_g reduces to ℓ_s, producing a family of separate variety of estimators whose approximate variance structure is

$$V(\ell_s) = V(\bar{y}) + \frac{1}{nN}\sum_{i=1}^{N} \nu_i S_{ix}^2 g_{i1}\left(g_{i1} + 2\beta_{iyx}\right) \tag{12}$$

The class of estimators produced by ℓ_s constitutes a subclass of estimators produced by ℓ_g. Two estimators $\ell_R^{(s)} = \frac{1}{n}\sum_{i\epsilon s} u_i \bar{y}_i \frac{\bar{x}_{di}}{\bar{x}_i}$ and $\ell_P^{(s)} = \frac{1}{n}\sum_{i\epsilon s} u_i \bar{y}_i \frac{\bar{x}_i}{\bar{x}_{di}}$ may be identified as members of this subclass where $\ell_R^{(s)}$ is the two-phase sampling analogue of the ratio estimator considered in Hansen *et al*. (1953) [2]. The minimum variance bound and the MVB estimator of ℓ_s are given by

$$\min V(\ell_s) = V(\bar{y}) - \frac{1}{nN}\sum_{i=1}^{N} \nu_i S_{iy}^2 \rho_{iyx}^2 \tag{13}$$
And $\ell_{RG}^{(s)} = \frac{1}{n}\sum_{i\epsilon s} u_i \left[\bar{y}_i - \beta_{iyx}(\bar{x}_i - \bar{x}_{di})\right]$.

4.3 When the available auxiliary information about the clusters as well as about the population are utilized (*i.e.*, only case (i) is considered), then ℓ_g reduces to ℓ_c, a class of estimators (of course a subclass of ℓ_g) defined by

$\ell_c = g(\bar{y}, \bar{x}_d, \bar{x}'_d)$,

With approximate variance

$$V(\ell_c) = V(\bar{y}) + \frac{1-\gamma}{n} A_x^2 g_1(g_1 + 2\beta) \tag{14}$$

Now we see that three estimators defined by $\ell_R^{(c)} = \bar{y}\frac{\bar{x}'_d}{\bar{x}_d}$, $\ell_{RG1}^{(c)} = \bar{y} - \beta_{byx}(\bar{x}_d - \bar{x}'_d)$ and $\ell_P^{(c)} = \bar{y}\frac{\bar{x}_d}{\bar{x}'_d}$ are direct byproducts of ℓ_c, where $\ell_R^{(c)}$ is the Smith's (1969) [14] ratio estimator under our two-phase sampling design.
The approximate MVB of ℓ_c and the corresponding MVB estimator are

$$\min V(\ell_c) = V(\bar{y}) - \frac{1-\gamma}{n} A_y^2 \rho^2 \tag{15}$$
and $\ell_{RG}^{(c)} = \bar{y} - \beta(\bar{x}_d - \bar{x}'_d)$.

4.4 If $s_i = s'_i \; \forall \; i$ and $s = s'$ *i.e.*, if we consider two-stage sampling with single phase selection at each stage, then ℓ_g defines the class of estimators studied earlier by Sahoo and Panda (1997) [7].
4.5 When $s_i = s'_i = U_i \; \forall \; i$, ℓ_g generates a system of estimators for two-phase cluster sampling. On the other hand, when $s = s' = U$, we get a system of separate estimators for two-phase stratified random sampling with PSUs as strata.

5. Efficiency of the class
In order to study efficiency aspect of the formulated estimation technique of the utilization of auxiliary variable at different stages, we consider two cases. In the first case, we compare efficiency of ℓ_g with that of $\bar{y}$, the traditional estimator incorporating no auxiliary variable, and in the second case we compare with that of the traditional class of estimators suggested by Srivastava (1980) incorporating x.

5.1 We note from (9) that $V(\ell_g) \leq V(\bar{y})$ *i.e.*, an estimator of ℓ_g is more efficient than $\bar{y}$ if

$$\beta_{iyz} < -\frac{1}{2} g_{i1} \; \forall \; i \text{ and } \beta < -\frac{1}{2} g_1 \tag{16}$$

These sufficient conditions, although strongly dependent on the selection and nature of the functions g_i and g, clearly indicate that there is scope for improving upon the direct estimation strategy that considers no auxiliary variable through the suggested

estimation strategy that considers one auxiliary variable. From (11) it is also clear that for optimum choices of g_{i1} and g_1, the later strategy always yields higher efficiency gain over the former one.

5.2 Srivastava's (1980) class of estimators for the adopted two-phase sampling mechanism, may be defined by

$t_h = h(\bar{y}, \bar{x}, \bar{x}'_d)$,

Where $(\bar{y}, \bar{x}, \bar{x}'_d)\ \epsilon\ S$ and the function $h(\bar{y}, \bar{x}, \bar{x}'_d)$ involved in composing the class admits regularity conditions as stated in section 3 such that $h(\bar{Y}, \bar{X}, \bar{X}) = \bar{Y}$. As is discussed earlier, here we also have $h_0 = 1$ and we assume that $h_1 = -h_2$, where $h_0, h_1$and h_2 are respectively the first order partial derivatives of $h(\bar{y}, \bar{x}, \bar{x}'_d)$ with respect to $\bar{y}, \bar{x}$ and $\bar{x}'_d$ about $(\bar{Y}, \bar{X}, \bar{X})$. Hence, an approximate expression of $V(t_h)$, derived through the Taylor linearization method, is given by

$$V(t_h) = V(\bar{y}) + \frac{1-\gamma}{n} \Delta_x^2 h_1 (h_1 + 2\beta) + \frac{1}{nN} \sum_{i=1}^{N} \nu_i S_{ix}^2 h_1 \left(h_1 + 2\beta_{iyx}\right) \tag{17}$$

From (9) and (17) one can easily check that $V(\ell_g) \leq V(t_h)$ *i.e.*, an estimator of ℓ_g is more efficient than an estimator of t_h when

$$\beta_{iyx} \geq -\frac{1}{2}(h_1 + g_{i1})\ \forall\ i \text{ and } \beta \geq -\frac{1}{2}(h_1 + g_1) \tag{18}$$

After simplification and rearrangement of terms in (17), an alternative expression for $V(t_h)$ is given by

$$V(t_h) = V(\bar{y}) + \frac{1-\gamma}{n} \Delta_{hx}^2 h_1 (h_1 + 2\beta_h) \tag{19}$$

Where $\alpha_i = \eta_i + \frac{1}{1-\gamma}\nu_i, \Delta_{hx}^2 = S_{bx}^2 + \frac{1}{N}\sum_{i=1}^{N} \alpha_i S_{ix}^2, \Delta_{hyx} = S_{byx} + \frac{1}{N}\sum_{i=1}^{N} \alpha_i S_{iyx}$ and $\beta_h = \Delta_{hyx}/\Delta_{hx}^2$. Hence, $\partial V(t_h)/\partial h_1 = 0$ leads to the optimal value $h_1^{(\text{opt})} = -\beta_h$ so that the MVB of the class and the resulting MVB estimator are given by $\min V(t_h) = V(\bar{y}) - \frac{1-\gamma}{n}\Delta_{hy}^2 \rho_h^2$

(20)

And $t_{RG}^{(h)} = \bar{y} - \beta_h(\bar{x} - \bar{x}_d)$,

Where $\Delta_{hy}^2 = S_{by}^2 + \frac{1}{N}\sum_{i=1}^{N} \alpha_i S_{iy}^2$ and $\rho_h = \Delta_{hyx}/\Delta_{hy}\Delta_{hx}$. From (11) and (20) it appears that for the optimal cases ℓ_g is likely to have a smaller variance than t_h *i.e.*, ℓ_{RG} is likely to be more efficient than $t_{RG}^{(h)}$ when $\Delta_{hy}^2 \rho_h^2 \leq \Delta_y^2 \rho^2$.

Considering ℓ_c and t_h as competing classes for estimating $\bar{Y}$, we also see that $V(\ell_c) \leq V(t_h)$ when $\frac{g_1(g_1+2\beta)}{h_1(h_1+2\beta_h)} \geq \frac{\Delta_{hx}^2}{\Delta_x^2}$, and $\min V(\ell_c) \leq \min V(t_h)$ when $\Delta_{hy}^2 \rho_h^2 \leq \Delta_y^2 \rho^2$.

6. Numerical Study

From the various results derived in the preceding sections we note that an analytical comparison to test the effectiveness of different estimation methods discussed here is not very much useful. One way to avoid this difficulty is to consider a numerical comparison as a counterpart to the analytical comparison. Here we carry out a numerical study to illustrate the performance of different estimators using data on two populations.

Population I: It consists of strip-wise complete enumeration data on timber volume $(= y)$ and length $(= x)$ for 176 strips (SSUs) divided into 10 $(= 10)$ blocks (PSUs) of the Black Mountain Experimental Forest given in Murthy (1977, p.131) [4]. For this population, we select $n' = 6, n = 3, m_i' = 5, 5, 5, 5, 4, 5, 4, 5, 8$ and 8, and $m_i = 2, \forall i \in s$.

Population II: This population is the MU284 population available in Sarndal, Swensson and Wretman (1992, p.652 [13], Appendix B). It consists of 284 municipalities (SSUs) divided into 50 clusters (PSUs) with two variables 1985 population $(= y)$ and 1975 population $(= x)$. Here, we consider $n' = 20, n = 10, m_i' = 3, \forall i \in s'$, and $m_i = 2, \forall i \in s$.

The estimators under consideration are the four MVB estimators *viz.*, $\ell_{RG}, \ell_{RG}^{(s)}, \ell_{RG}^{(c)}$ and $t_{RG}^{(h)}$; three ratio-type estimators: ℓ_R , $\ell_R^{(s)}$and $\ell_R^{(c)}$; and two regression-type estimators: ℓ_{RG1} and $\ell_{RG1}^{(c)}$ already defined in section 4. We also include a ratio-type estimator $t_R^{(h)} = \bar{y}\frac{\bar{x}'_d}{\bar{x}}$ and a regression-type estimator $t_{RG1}^{(h)} = \bar{y} - \beta_{byx}(\bar{x} - \bar{x}'_d)$ as some specific cases of t_h in our numerical comparison. We do not touch product or product-type estimators as in the two populations y is positively correlated with x.

Approximate variance expressions of the ratio-type estimators ℓ_R , $\ell_R^{(s)}$, $\ell_R^{(c)}$ and $t_R^{(h)}$can be easily obtained from (9), (12), (14) and (19) respectively by substituting $g_{i1} = \frac{\bar{Y}_i}{\bar{X}_i}$, $g_1 = \frac{\bar{Y}}{\bar{X}}$ and $h_1 = \frac{\bar{Y}}{\bar{X}}$. However, the approximate variance expressions of the regression-type estimators ℓ_{RG1}, $\ell_{RG1}^{(c)}$ and $t_{RG1}^{(h)}$, in the simplified form, are as follows:

$$V(\ell_{RG1}) = V(\bar{y}) - \frac{1-\gamma}{n}\left[S_{by}^2 \rho_{byx}^2 + \frac{\beta_{byx}}{N}\sum_{i=1}^{N} \eta_i S_{ix}^2 \left(2\beta_{iyx} - \beta_{byx}\right)\right] - \frac{1}{nN}\sum_{i=1}^{N} \nu_i S_{iy}^2 \rho_{iyx}^2 \tag{21}$$

$$V\left(\ell_{RG1}^{(c)}\right) = V(\bar{y}) - \frac{1-\gamma}{n}\beta_{byx}\Delta_x^2\left(2\beta - \beta_{byx}\right) \tag{22}$$

$$V\left(t_{RG1}^{(h)}\right) = V(\bar{y}) - \frac{1-\gamma}{n}\beta_{byx}\Delta_x^2\left(2\beta - \beta_{byx}\right) - \frac{\beta_{byx}}{nN}\sum_{i=1}^{N} \nu_i S_{ix}^2 \left(2\beta_{iyx} - \beta_{byx}\right) \tag{23}$$

Where $\rho_{byx} = S_{byx}/S_{by}S_{bx}$.

To compute relative precision of an estimator ℓ compared to $\bar{y}$, we use its approximate variance which is defined by

$$RP = \frac{V(\bar{y})}{V(\ell)} \times 100 \quad (24)$$

Relative precision of the comparable estimators, as calculated from the data, are compiled in table 6.1.

Table 6.1: Relative Precision of Different Estimators Compared to $\bar{y}$ (in %)

Populations	Estimators										
	ℓ_{RG}	$\ell_{RG}^{(s)}$	$\ell_{RG}^{(c)}$	$t_{RG}^{(h)}$	ℓ_R	$\ell_R^{(s)}$	$\ell_R^{(c)}$	$t_R^{(h)}$	ℓ_{RG1}	$\ell_{RG1}^{(c)}$	$t_{RG1}^{(h)}$
I	201	152	181	169	145	113	136	119	166	145	123
II	368	301	334	315	211	140	185	178	290	232	224

Table 6.1 prompts the following results

i) Among the four MVB estimators, ℓ_{RG} and $\ell_{RG}^{(c)}$ respectively achieve the highest and the next highest precision gains.

ii) Among the three regression-type estimators $\ell_{RG1}, t_{RG1}^{(h)}$ and $\ell_{RG1}^{(c)}$, ℓ_{RG1} is better than others.

iii) Among the comparable ratio estimators, ℓ_R is superior to others followed by $\ell_R^{(c)}$.

In view of these results, ℓ_g turns out to be the best performer among the competing classes. Hence, we conclude that situations may be encountered in practice where the proposed estimation methodology can be effectively employed. However, finding of this numerical study, which of course has some limitations, may not fit to other situations.

7. References

1. Das NR Sahoo LN. A note on the performance of some two-stage regression estimators in two-phase sampling. Research Journal of Mathematical and Statistical Sciences. 2015; 3(10):1-3.
2. Hansen MH, Hurwitz WN, Madow WG. Sampling Survey Methods and Theory. John Wiley & Sons, New York, 1953, I.
3. Kim JY, Breidt FJ, Opsomer JD. Nonparametric regression estimation of finite population totals under two-stage sampling. Technical Report 4, Department of Statistics, Colorado State University, 2009.
4. Murthy MN. Sampling Theory and Methods. Statistical Publishing Society, Calcutta, 1977.
5. Sahoo LN. A regression-type estimator in two-stage sampling, Calcutta Statistical Association Bulletin. 1987; 36:97-100.
6. Sahoo LN, Das BC, Sahoo J. A class of predictive estimators in two-stage sampling. Journal of the Indian Society of Agricultural Statistics. 2009; 63(2):175-180.
7. Sahoo LN, Panda P. A class of estimators in two-stage sampling with varying probabilities. South African Statistical Journal. 1997; 31:151-160.
8. Sahoo LN, Panda P. A class of estimators using auxiliary information in two-stage sampling. Australian and New Zealand Journal of Statistics. 1999; 41(4):405-410.
9. Sahoo LN, Sahoo RK, Senapati SC, Mangaraj AK. A general class of estimators in two-stage sampling with two auxiliary variables. Hacettepe Journal of Mathematics and Statistics. 2011; 40(5):757-765.
10. Sahoo LN, Senapati SC, Singh GN. Analternative class of estimators in two-stage sampling with two auxiliary variables. Journal of Indian Statistical Association. 2005; 43(2):147-156.
11. Sahoo LN, Swain AKPC. Chain ratio estimators. Journal of the Indian Society of Agricultural Statistics. 1983; 35:70-79.
12. Sahoo LN, Swain AKPC. Chain product estimators. The Aligarh Journal of Statistics. 1986; 6:53-58.
13. Sarndal CE, Swensson B, Wretman J. Model Assisted Survey Sampling. Springer-Verlag, 1992.
14. Smith TMF. A note on ratio estimates in multi-stage sampling. Journal of the Royal Statistical Society. 1969; A132:426-430.
15. Srivastava SK. A class of estimators using auxiliary information in sample surveys. Canadian Journal of Statistics. 1980; 8: 253-254.
16. Zheng H, Little RJ. Penalized spine nonparametric mixed models for inference about finite population mean from two-stage sampling. Survey Methodology. 2004; 30:209-218.