Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 72(1) 2018 1–6

Calibration Estimation of Regression Coefficient for Two-stage Sampling Design using Single Auxiliary Variable

Pradip Basak, U.C. Sud and Hukum Chandra
ICAR-Indian Agricultural Statistics Research Institute, New Delhi

Received 27 December 2016; Revised 23 January 2017; Accepted 03 February 2017

SUMMARY

Regression analysis is a widely used technique for studying the relationship between variables. In this paper, an attempt has been made to study the estimation of regression coefficient in the context of two-stage survey data using single auxiliary variable. The theory of calibration approach is used to develop the estimators based on assumption that auxiliary information is available at primary stage unit (psu) level, and at both psu and second stage unit level. The expression for variance and variance estimator is obtained. The performance of the developed estimators is evaluated through a real data based simulation study.

Keywords: Regression coefficient, Calibration approach, Two-stage sampling.

1. INTRODUCTION

Survey data are collected to draw inference about the population parameters. Usually, the simple parameters like mean, total or proportion is of main interest. But, many a times when the objective is to establish the pattern of relationship between variables, then regression analysis is an important tool. Regression coefficient is a complex and non-linear population parameter. In the context of estimation of regression parameter, the well-known ordinary least squares (OLS) method is used which is based on the assumptions that sample observations are drawn independently. This assumption of independence only holds if survey data are collected through simple random sampling with replacement (SRSWR) scheme. Now a day, most of the surveys are complex in nature involving clustering, unequal probability of selection, multi-stage, multi-phase and auxiliary information, which violates the independence assumption required for OLS estimation. Modified approaches such as use of sampling design weights in the estimation procedure were suggested by Kish and Frankel (1974). Estimation of regression coefficient based on the method of maximum likelihood estimation was proposed by Holt, Smith and Winter (1980). Scott and Holt (1982) rather than using inclusion probabilities as weights incorporated the effect of clustering in the error structure resulting in heteroscedastic variance-covariance matrix of error terms. They suggested weighted least squares method for the estimation of regression coefficient.

Calibration is a widely used methodology of survey estimation that incorporates auxiliary information into the estimation procedure (Deville and Särndal, 1992). So far, many works has been done using calibration approach for the estimation of parameters like mean, total, variance, covariance etc. in the context of unistage or multi-stage design, see for example Aditya *et al.* (2016), Basak *et al.* (2014a, 2014b, 2016, 2017), Plikusas and Pumputis (2007, 2010), Wu and Luan (2003). Here, an attempt has been made to estimate the finite population regression coefficient using calibration approach in the context of two-stage sampling design where single auxiliary variable related to the study variable is available.

In Section 2, we discuss the general notations used for the development of estimator under two-stage

Corresponding author: Pradip Basak
E-mail address: pradipbasak99@gmail.com

sampling design. Section 3 presents the proposed estimators. In Section 4 variance estimation of the developed estimators is discussed. Section 5 provides empirical evaluation of the developed estimators. Finally, Section 6 presents concluding remarks.

2. THE NOTATIONS

Let *U*=(*1,2,…,k,…,N*) be a finite population of size *N* which is grouped into N_I clusters as $U_1, U_2, \ldots, U_i, \ldots, U_{N_I}$. The size of the *i*th cluster is N_i. Thus, $U = \bigcup_{i=1}^{N_I} U_i$ and $N = \sum_{i=1}^{N_I} N_i$. These clusters are primary stage units (psus) and units within the clusters are second stage units (ssus). The population of clusters is denoted by U_I. At the first stage, a sample of psus s_I of size n_I is drawn from U_I by using any probability sampling scheme. Then at the second stage, a sample of units s_i of size n_i is drawn from the *i*th selected psus, U_i of size N_i by using any probability sampling scheme. Thus, $s = \bigcup_{i=1}^{n_I} s_i$ and $n_s = \sum_{i=1}^{n_I} n_i$, where *s* is the two-stage sample and n_s is the two-stage sample size. Let, π_{Ii} and π_{Iij} be the first and second order inclusion probability at the first stage, whereas for the second stage it is $\pi_{k/i}$ and $\pi_{kl/i}$ respectively.

Let, y be the dependent variable and x be the independent variable under study. Here, it is assumed that auxiliary variable z is associated with dependent variable y. Let, y_{ik}, x_{ik} and z_{ik} be the values of variables y, x, and z corresponding to the *j*th unit of *i*th selected psu. The population total of *y* is given by $t_y = \sum_{i=1}^{N_I}\sum_{k=1}^{N_i} y_{ik} = \sum_{i=1}^{N_I} t_{iy}$, where $t_{iy} = \sum_{k=1}^{N_i} y_{ik}$ is the *i*th psu total of *y*. Similarly, population total of *x* is given by $t_x = \sum_{i=1}^{N_I}\sum_{k=1}^{N_i} x_{ik} = \sum_{i=1}^{N_I} t_{ix}$, where $t_{ix} = \sum_{k=1}^{N_i} x_{ik}$ is the *i*th psu total of *x*. Let Z_i be the *i*th psu total of auxiliary variable *z*. Thus, $Z_i = \sum_{k=1}^{N_i} z_{ik}$ and $t_z = \sum_{i=1}^{N_I} Z_i$.

Here, the parameter of interest is population regression coefficient *B*, defined by

$$B = \frac{\sum_{i=1}^{N_I}\sum_{k=1}^{N_i}\left(x_{ik} - \bar{X}\right)\left(y_{ik} - \bar{Y}\right)}{\sum_{i=1}^{N_I}\sum_{k=1}^{N_i}\left(x_{ik} - \bar{X}\right)^2}$$

where $\bar{X} = \frac{1}{N}\sum_{i=1}^{N_I}\sum_{k=1}^{N_i} x_{ik}$ and $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N_I}\sum_{k=1}^{N_i} y_{ik}$.

The *π*-estimator of population regression coefficient is given by

$$\hat{B}_\pi = \frac{\sum_{i=1}^{n_I} a_{Ii} \sum_{k=1}^{n_i} a_{k/i}\left(x_{ik} - \hat{t}_{x\pi}/N\right)\left(y_{ik} - \hat{t}_{y\pi}/N\right)}{\sum_{i=1}^{n_I} a_{Ii} \sum_{k=1}^{n_i} a_{k/i}\left(x_{ik} - \hat{t}_{x\pi}/N\right)^2} \qquad (1)$$

where, $a_{Ii} = 1/\pi_{Ii}, a_{k/i} = 1/\pi_{k/i}$

$$\hat{t}_{x\pi} = \sum_{i=1}^{n_I} a_{Ii}\hat{t}_{ix}\,,\ \hat{t}_{ix} = \sum_{k=1}^{n_i} a_{k/i} x_{ik}$$

$$\hat{t}_{y\pi} = \sum_{i=1}^{n_I} a_{Ii}\hat{t}_{iy}\,,\ \hat{t}_{iy} \quad \sum a_{k\ i} y_{ik}$$

The *π*-estimator defined in (1) can also be expressed as

$$\hat{B}_\pi = \frac{\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}\left(x_{ik} - \hat{t}_{x\pi}/N\right)\left(y_{ik} - \hat{t}_{y\pi}/N\right)}{\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}\left(x_{ik} - \hat{t}_{x\pi}/N\right)^2} \qquad (2)$$

where a_{ik} is the design weight corresponding to *k*th unit of *i*th selected psu. Thus, $a_{ik} = a_{Ii}a_{k/i}, \forall i = 1,\ldots,n_I$ and $k = 1,\ldots,n_i$.

3. THE PROPOSED ESTIMATORS

Here, we have assumed two different cases of availability of auxiliary information:

Case 1: psu level information is available for single auxiliary variable *z*

Case 2: psu and ssu level information is available for single auxiliary variable *z*

3.1 Calibration Estimation under Case 1

Here, it is assumed that population level auxiliary information is available at psu level, i.e., Z_i is known $\forall i = 1,2,\ldots,N_I$. In this case, the calibration constraint is defined as $\sum_{i=1}^{n_I} w_{Ii} Z_i = \sum_{i=1}^{N_I} Z_i$, where w_{Ii} is the calibrated weight corresponding to the design weight a_{Ii}. The chi-square distance function measuring the distance between w_{Ii} and a_{Ii} is given by

$\sum_{i=1}^{n_I}\left(w_{Ii}-a_{Ii}\right)^2 / a_{Ii}q_{Ii}$.

Thus, the objective function for minimization is given by

$$\phi=\sum_{i=1}^{n_I}\left(w_{Ii}-a_{Ii}\right)^2 / a_{Ii}q_{Ii}-2\lambda_1\left(\sum_{i=1}^{n_I}w_{Ii}Z_i-\sum_{i=1}^{N_I}Z_i\right).$$

The calibrated weight w_{Ii} is obtained by minimizing this objective function using Lagrange multiplier approach. Finally the calibrated weights are obtained as

$$w_{Ii}=a_{Ii}\left\{1+q_{Ii}\lambda Z_i\right\}; i=1,2,...,n_I$$

where, $\lambda=\left(\sum_{i=1}^{N_I}Z_i \text{ - } \sum_{i=1}^{n_I}a_{Ii}Z_i\right) / \left(\sum_{i=1}^{n_I}a_{Ii}q_{Ii}Z_i^2\right)$.

Here, q_{Ii} is a positive constant and for the particular case $q_{Ii}=1$, the calibrated weights are obtained as $w_{Ii}=a_{Ii}\{1+\lambda Z_i\}$, where, $\lambda=\left(\sum_{i=1}^{N_I}Z_i \text{ - } \sum_{i=1}^{n_I}a_{Ii}Z_i\right) / \left(\sum_{i=1}^{n_I}a_{Ii}Z_i^2\right)$.

Now, the calibrated estimators of population total of *y* is obtained as $\hat{t}_{y\pi}^{c(1)}=\sum_{i=1}^{n_I}w_{Ii}\hat{t}_{iy}$. Finally, the calibrated estimator of population regression coefficient is obtained as

$$\hat{B}_{\pi c}^{(1)}=\frac{\sum_{i=1}^{n_I}w_{Ii}\sum_{k=1}^{n_i}a_{k/i}\left(x_{ik}-\hat{t}_{x\pi}/N\right)\left(y_{ik}-\hat{t}_{y\pi}^{c(1)}/N\right)}{\sum_{i=1}^{n_I}w_{Ii}\sum_{k=1}^{n_i}a_{k/i}\left(x_{ik}-\hat{t}_{x\pi}/N\right)^2} \qquad (3)$$

3.2 Calibration Estimation under Case 2

Here, it is assumed that auxiliary information is available at both psu and sssu level. Thus, the calibration constraint is defined as $\sum_{i=1}^{n_I}\sum_{k=1}^{n_i}w_{ik}z_{ik}=\sum_{i=1}^{N_I}\sum_{k=1}^{N_i}z_{ik}$, where w_{ik} is the calibrated weight corresponding to the design weight a_{ik}. In this case, the chi-square distance function is given by $\sum_{i=1}^{n_I}\sum_{k=1}^{n_i}\frac{(w_{ik}-a_{ik})^2}{a_{ik}q_{ik}}$. Now, the objective function

$$\phi=\sum_{i=1}^{n_I}\sum_{k=1}^{n_i}\frac{(w_{ik}-a_{ik})^2}{a_{ik}q_{ik}}-2\lambda_1\left(\sum_{i=1}^{n_I}\sum_{k=1}^{n_i}w_{ik}z_{ik}-\sum_{i=1}^{N_I}\sum_{k=1}^{N_i}z_{ik}\right)$$

is minimized to obtain the calibrated weights, w_{ik}. Lagrangian multiplier approach is used for minimization. Finally, the calibrated weights are obtained as $w_{ik}=a_{ik}\left\{1+q_{ik}\lambda_1 z_{ik}\right\}$

where, $\lambda_1=\left(\sum_{i=1}^{N_I}\sum_{k=1}^{N_i}z_{ik}-\sum_{i=1}^{n_I}\sum_{k=1}^{n_i}a_{ik}z_{ik}\right) / \left(\sum_{i=1}^{n_I}\sum_{k=1}^{n_i}a_{ik}q_{ik}z_{ik}^2\right)$.

Here, we have assumed $q_{ik}=1$ as a particular case. Thus, the calibrated estimator of population regression coefficient under case 2 is given by

$$\hat{B}_{\pi c}^{(2)}=\frac{\sum_{i=1}^{n_I}\sum_{k=1}^{n_i}w_{ik}\left(x_{ik}-\hat{t}_{x\pi}/N\right)\left(y_{ik}-\hat{t}_{y\pi}^{c(2)}/N\right)}{\sum_{i=1}^{n_I}\sum_{k=1}^{n_i}w_{ik}\left(x_{ik}-\hat{t}_{x\pi}/N\right)^2}. \qquad (4)$$

4. VARIANCE ESTIMATION

The developed calibrated estimators are nonlinear in nature. Thus, Taylor series linearization approach is used for variance estimation.

4.1 Variance Estimation under Case 1

The calibrated estimator, $\hat{B}_{\pi c}^{(1)}$ can also be expressed as

$$\hat{B}_{\pi c}^{(1)}=\frac{\sum_{i=1}^{n_I}w_{Ii}\sum_{k=1}^{n_i}a_{k/i}\left(x_{ik}-\sum_{i=1}^{n_I}a_{Ii}\hat{t}_{ix}/N\right)\left(y_{ik}-\sum_{i=1}^{n_I}w_{Ii}\hat{t}_{iy}/N\right)}{\sum_{i=1}^{n_I}w_{Ii}\sum_{k=1}^{n_i}a_{k/i}\left(x_{ik}-\sum_{i=1}^{n_I}a_{Ii}\hat{t}_{ix}/N\right)^2}$$

where, $w_{Ii}=a_{Ii}\left\{1+\left(\sum_{i=1}^{N_I}Z_i-\sum_{i=1}^{n_I}a_{Ii}Z_i\right)\left(\sum_{i=1}^{n_I}a_{Ii}Z_i^2\right)^{-1}Z_i\right\}$.

Let us write, $\hat{t}_z=\sum_{i=1}^{n_I}a_{Ii}Z_i$ and $\hat{A}_z=\sum_{i=1}^{n_I}a_{Ii}Z_i^2$. Now, $\hat{B}_{\pi c}^{(1)}$ can be expressed as functions of several estimated population totals as

$$\hat{B}_{\pi c}^{(1)}=f\left(\hat{t}_{xy},\hat{t}_z,\hat{A}_z,\hat{t}_{xyz},\hat{t}_x,\hat{t}_{xz},\hat{t}_y,\hat{t}_{yz},\hat{N},\hat{t}_{Nz},\hat{t}_{xx},\hat{t}_{xxz}\right)$$

where,

$$\hat{t}_{xy}=\sum_{i=1}^{n_I}a_{Ii}\hat{t}_{ixy}, \hat{t}_{xyz}=\sum_{i=1}^{n_I}a_{Ii}\hat{t}_{ixy}Z_i, \hat{t}_x=\sum_{i=1}^{n_I}a_{Ii}\hat{t}_{ix}, \hat{t}_{xz}=\sum_{i=1}^{n_I}a_{Ii}\hat{t}_{ix}Z_i, \hat{t}_y=\sum_{i=1}^{n_I}a_{Ii}\hat{t}_{iy}, \hat{t}_{yz}=\sum_{i=1}^{n_I}a_{Ii}\hat{t}_{iy}Z_i,$$

$\hat{N}=\sum_{i=1}^{n_I}a_{Ii}\hat{N}_i$, $\hat{t}_{Nz}=\sum_{i=1}^{n_I}a_{Ii}\hat{N}_iZ_i$, $\hat{t}_{xx}$ $\sum a_{Ii}\hat{t}_{ixx}$ and $\hat{t}_{xxz}=\sum_{i=1}^{n_I}a_{Ii}\hat{t}_{ixx}Z_i$. Thus, using Taylor series linearization, the estimator $\hat{B}_{\pi c}^{(1)}$, defined in (3), is approximated by

$$\hat{B}_{\pi c}^{(1)} \approx B + \frac{1}{S_x^2}\sum_{i=1}^{n_I} a_{Ii}\left\{\sum_{k=1}^{n_i} a_{k/i}\left(x_{ik}-\bar{X}\right)E_{ik} - Z_i A_z^{-1}\sum_{i=1}^{N_I}\sum_{k=1}^{N_i}\left(x_{ik}-\bar{X}\right)E_{ik}Z_i\right\}$$

The approximate variance of the estimator is given by

$$V\left(\hat{B}_{\pi c}^{(1)}\right) = \frac{1}{\left(S_x^2\right)^2}\left[\sum_{i=1}^{N_I}\sum_{j=1}^{N_I}\Delta_{Iij}\frac{E_{ic}^{(1)}}{\pi_{Ii}}\frac{E_{jc}^{(1)}}{\pi_{Ij}} + \sum_{i=1}^{N_I}\frac{1}{\pi_{Ii}}\sum_{k=1}^{N_i}\sum_{l=1}^{N_i}\Delta_{kl/i}\frac{\left(x_{ik}-\bar{X}\right)E_{ik}}{\pi_{k/i}}\frac{\left(x_{il}-\bar{X}\right)E_{il}}{\pi_{l/i}}\right]$$

where, $S_x^2 = \sum_{i=1}^{N_I}\sum_{k=1}^{N_i}\left(x_{ik}-\bar{X}\right)^2$

$$E_{ic}^{(1)} = X_{E_i} - B_{ix_E z}Z_i$$

$$X_{E_i} = \sum_{k=1}^{N_i}\left(x_{ik}-\bar{X}\right)E_{ik}.$$

$$B_{ix_E z} = \left(\sum_{i=1}^{N_I} X_{E_i}Z_i\right) / \left(\sum_{i=1}^{N_I} Z_i^2\right).$$

$$E_{ik} = \left(y_{ik}-\bar{Y}\right) - B\left(x_{ik}-\bar{X}\right).$$

The variance estimator of the calibrated estimator $\hat{B}_{\pi c}^{(1)}$ is obtained as

$$\hat{V}\left(\hat{B}_{\pi c}^{(1)}\right) = \frac{1}{\left(\hat{S}_x^2\right)^2}\left[\sum_{i=1}^{n_I}\sum_{j=1}^{n_I}\breve{\Delta}_{Iij}\frac{\hat{E}_{ic}^{(1)}}{\pi_{Ii}}\frac{\hat{E}_{jc}^{(1)}}{\pi_{Ij}} + \sum_{i=1}^{n_I}\frac{1}{\pi_{Ii}}\sum_{k=1}^{n_i}\sum_{l=1}^{n_i}\breve{\Delta}_{kl/i}\frac{\left(x_{ik}-\hat{\hat{X}}\right)\hat{E}_{ik}}{\pi_{k/i}}\frac{\left(x_{il}-\hat{\hat{X}}\right)\hat{E}_{il}}{\pi_{l/i}}\right]$$

where, $\hat{S}_x{}^2 = \sum_{i=1}^{n_I} a_{Ii}\sum_{k=1}^{n_i} a_{k/i}\left(x_{ik}-\hat{\hat{X}}\right)^2$, $\hat{\hat{X}} = \frac{\hat{t}_{x\pi}}{N}$, $\hat{\hat{Y}} = \frac{\hat{t}_{y\pi}}{N}$,

$$\breve{\Delta}_{Iij} = \frac{\pi_{Iij}-\pi_{Ii}\pi_{Ij}}{\pi_{Iij}},\ \breve{\Delta}_{kl/i} = \frac{\pi_{kl/i}-\pi_{k/i}\pi_{l/i}}{\pi_{kl/i}},$$

$$\hat{E}_{ic}^{(1)} = \hat{X}_{E_i} - \hat{B}_{ix_E z}Z_i,\ \hat{X}_{E_i} = \sum_{k=1}^{n_i} a_{k/i}\left(x_{ik}-\hat{\hat{X}}\right)\hat{E}_{ik},$$

$$\hat{E}_{ik} = \left(y_{ik}-\hat{\hat{Y}}\right) - \hat{B}_{\pi}\left(x_{ik}-\hat{\hat{X}}\right) \text{ and } \hat{B}_{ix_E z} = \left(\sum_{i=1}^{n_I} a_{Ii}\hat{X}_{E_i}Z_i\right) / \left(\sum_{i=1}^{n_I} a_{Ii}Z_i^2\right).$$

4.2 Variance Estimation under Case 2

Now, $\hat{B}_{\pi c}^{(2)}$ can also be expressed as

$$\hat{B}_{\pi c}^{(2)} = \frac{\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} w_{ik}\left(x_{ik}-\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}x_{ik}/N\right)\left(y_{ik}-\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} w_{ik}y_{ik}/N\right)}{\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} w_{ik}\left(x_{ik}-\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}x_{ik}/N\right)^2}$$

where,

$$w_{ik} = a_{ik}\left\{1+\left(\sum_{i=1}^{N_I}\sum_{k=1}^{N_i} z_{ik} - \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}z_{ik}\right)\left(\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}z_{ik}^2\right)^{-1} z_{ik}\right\}.$$

Let, $\hat{t}_z = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}z_{ik}$ and $\hat{A}_z = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}z_{ik}^2$. Thus, $\hat{B}_{\pi c}^{(2)}$ can also be expressed as function of several estimated population totals as

$$\hat{B}_{\pi c}^{(2)} = f\left(\hat{t}_{xy},\ \hat{t}_z,\ \hat{A}_z,\ \hat{t}_{xyz},\ \hat{t}_x,\ \hat{t}_{xz},\hat{t}_y,\ \hat{t}_{yz},\hat{N},\hat{t}_{xx},\ \hat{t}_{xxz}\right)$$

where,

$$\hat{t}_{xy} = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}x_{ik}y_{ik},\ \hat{t}_{xyz} = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}x_{ik}y_{ik}z_{ik},\ \hat{t}_x = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}x_{ik},$$

$$\hat{t}_{xz} = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}x_{ik}z_{ik},\ \hat{t}_y = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}y_{ik},\ \hat{t}_{yz} = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}y_{ik}z_{ik},$$

$$\hat{N} = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik},\ \hat{t}_{xx} = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}x_{ik}^2,\ \hat{t}_{xxz} = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}x_{ik}^2z_{ik}.$$

Now, the Taylor linearized estimator of $\hat{B}_{\pi c}^{(2)}$ is obtained as

$$\hat{B}_{\pi c}^{(2)} \approx B + \frac{1}{S_x^2}\left[\sum_{i=1}^{n_I} a_{Ii}\sum_{k=1}^{n_i} a_{k/i}E_{ik}^c\right]$$

where, $E_{ik}^c = x_{E_{ik}} - B_{x_E z}z_{ik}$, $x_{E_{ik}} = \left(x_{ik}-\bar{X}\right)E_{ik}$,

$$B_{x_E z} = \left(\sum_{i=1}^{N_I}\sum_{k=1}^{N_i} x_{E_{ik}}z_{ik}\right) / \left(\sum_{i=1}^{N_I}\sum_{k=1}^{N_i} z_{ik}^2\right).$$

The approximate variance of $\hat{B}_{\pi c}^{(2)}$ is given by

$$V\left(\hat{B}_{\pi c}^{(2)}\right) = \frac{1}{\left(S_x^2\right)^2}\left[\sum_{i=1}^{N_I}\sum_{j=1}^{N_I}\Delta_{Iij}\frac{E_{ic}^{(2)}}{\pi_{Ii}}\frac{E_{jc}^{(2)}}{\pi_{Ij}} + \sum_{i=1}^{N_I}\frac{1}{\pi_{Ii}}\sum_{k=1}^{N_i}\sum_{l=1}^{N_i}\Delta_{kl/i}\frac{E_{ik}^c}{\pi_{k/i}}\frac{E_{il}^c}{\pi_{l/i}}\right]$$

where, $E_{ic}^{(2)} = X_{E_i} - B_{x_E z}Z_i$.

The variance estimator is given as

$$\hat{V}\left(\hat{B}_{\pi c}^{(2)}\right) = \frac{1}{\left(\hat{S}_x^2\right)^2}\left[\sum_{i=1}^{n_I}\sum_{j=1}^{n_I}\breve{\Delta}_{Iij}\frac{\hat{E}_{ic}^{(2)}}{\pi_{Ii}}\frac{\hat{E}_{jc}^{(2)}}{\pi_{Ij}} + \sum_{i=1}^{n_I}\frac{1}{\pi_{Ii}}\sum_{k=1}^{n_i}\sum_{l=1}^{n_i}\breve{\Delta}_{kl/i}\frac{\hat{E}_{ik}^c}{\pi_{k/i}}\frac{\hat{E}_{il}^c}{\pi_{l/i}}\right]$$

where, $\hat{E}_{ic}^{(2)} = \hat{X}_{E_i} - \hat{B}_{x_E z}Z_i$,

$$\hat{B}_{x_E z} = \left(\sum_{i=1}^{n_I} \sum_{k=1}^{n_i} a_{ik} \hat{x}_{E_{ik}} z_{ik} \right) / \left(\sum_{i=1}^{n_I} \sum_{k=1}^{n_i} a_{ik} z_{ik}^2 \right),$$

$$\hat{x}_{E_{ik}} = \left(x_{ik} - \hat{\bar{X}} \right) \hat{E}_{ik} , \ \hat{E}_{ik}^{c} = \hat{x}_{E_{ik}} - \hat{B}_{x_E z} z_{ik} \ .$$

5. EMPIRICAL STUDY

In this study, we considered the following three estimators of population regression coefficient:

(i) π-estimator, $\hat{B}_{\pi}$ given by (1) (denoted as Est-π),

(ii) Calibrated estimator, $\hat{B}_{\pi c}^{(1)}$ given by (3) (denoted as Est-CAL1),

(iii) Calibrated estimator, $\hat{B}_{\pi c}^{(2)}$ given by (4) (denoted as Est-CAL2).

The estimators were evaluated based upon the criteria of percentage absolute relative bias (ARB) and percentage relative root mean squared error (RRMSE), defined by

$$ARB(\hat{B}) = \frac{1}{M} \sum_{i=1}^{M} \left| \frac{\hat{B}_i - B}{B} \right| \times 100 \text{ and}$$

$$RRMSE(\hat{B}) = \sqrt{M^{-1} \sum_{i=1}^{M} \left(\frac{\hat{B}_i - B}{B} \right)^2} \times 100$$

where $\hat{B}_i$ denotes the predicted value of population regression coefficient at simulation run *i*, with true value *B*. Here, *M* denotes the number of simulation run.

The MU284 population dataset (Särndal *et al.*, 1992) was used for simulation. It consists of 284 municipalities of Sweden which are grouped into 50 clusters each containing 5 to 9 municipalities. The aim was to estimate population regression coefficient between variables revenues from the 1985 Municipal taxation (RMT85, measured in millions of kronor) and total number of seats in the municipal council (S82). Here, the number of municipal employees in 1984 (ME84) was used as the auxiliary variable related to the dependent variable, RMT85. From this population, four different combinations of sample: i) $n_I = 20, \ n_i = 4, \ n_s = 80,$ ii) $n_I = 20, \ n_i = 2, \ n_s = 40,$ iii) $n_I = 10, \ n_i = 4, \ n_s = 40,$ iv) $n_I = 10, \ n_i = 2, \ n_s = 20,$ were drawn by using simple random sampling without replacement (SRSWOR) at the both stage. Then the values of different estimators were computed using the sample data. The simulation was repeated to a total number of 5000 times. The values of percentage absolute relative bias and percentage relative root mean square error of different estimators are presented in Table 1.

The results in Table 1 indicate that the gain in efficiency in terms of RRMSE increases with decease in number of sampled psu for both Est-CAL1 and Est-CAL2. However, for a fixed number of sampled psu, the relative gain in efficiency decreases for Est-CAL1 while increases for Est-CAL2 with the increase in the number of sampled ssu. These results clearly show that both Est-CAL1 and Est-CAL2 estimators outperform the existing π-estimator for all the combinations of sample sizes. Further, between two calibration based estimators, the calibration estimator Est-CAL2 performs better than the Est-CAL1 for all the situations in terms of both the criteria.

Table 1. Percentage absolute relative bias (ARB, %), percentage relative root mean square error (RRMSE, %) of different estimators and percentage relative gains in ARB and RRMSE with respect to the π-estimator

Estimators	**Est- π**	**Est-CAL1**	**Est-CAL2**
$n_I = 20, \ n_i = 4, \ n_s = 80$			
ARB, %	36.4495	32.5624	30.1201
RRMSE, %	41.4177	37.4416	34.9051
Relative gain in ARB	-	11.9374	21.0139
Relative gain in RRMSE	-	10.6195	18.6580
$n_I = 20, \ n_i = 2, \ n_s = 40$			
ARB, %	53.8858	49.9855	40.8169
RRMSE, %	59.7347	55.3386	45.1114
Relative gain in ARB	-	7.8029	32.0184
Relative gain in RRMSE	-	7.9440	32.4160
$n_I = 10, \ n_i = 4, \ n_s = 40$			
ARB, %	54.1550	47.2779	43.1806
RRMSE, %	59.6381	52.3364	47.9050
Relative gain in ARB	-	14.5461	25.4151
Relative gain in RRMSE	-	13.9515	24.4924
$n_I = 10, \ n_i = 2, \ n_s = 20$			
ARB, %	64.4190	59.5207	51.8965
RRMSE, %	77.2630	69.2377	57.8493
Relative gain in ARB	-	8.2296	24.1298
Relative gain in RRMSE	-	11.5909	33.5591

6. CONCLUDING REMARKS

This paper describes different estimators of finite population regression coefficient based on the availability of auxiliary information on single variable at psu and at both psu and ssu level. The calibration estimator based on both psu and ssu level information on single auxiliary variable is found to be superior on the basis empirical evaluation through real data based simulation study.

ACKNOWLEDGEMENT

The authors would like to acknowledge the valuable comments and suggestions of the editorial board and the referee. These led to a considerable improvement in the paper. The first author also gratefully acknowledges the INSPIRE Fellowship provided by Department of Science and Technology, Government of India.

REFERENCES

Aditya, K., Sud, U.C., Chandra, H. and Biswas, A. (2016). Calibration based regression type estimator of the population total under two stage sampling design. *J. Ind. Soc. Agric. Statist.*, **70(1)**, 19-24.

Basak, P., Chandra, H. and Sud, U.C. (2014a). Estimation of finite population total for skewed data. *J. Ind. Soc. Agril. Statist.*, **68(3)**, 333-341.

Basak, P., Chandra, H., Sud, U.C. and Lal, S.B. (2014b). Prediction of population total for skewed variable under a log transform. *Int. J. Agril. Statist. Sci.*, **9(2)**, 143-154.

Basak, P., Sud, U.C. and Chandra, H. (2016). Calibration approach based estimator of finite population regression coefficient under two-stage sampling design. *Int. J. Agril. Statist. Sci.*, **12(2)**, 415-422.

Basak, P., Sud, U.C. and Chandra, H. (2017). Calibration estimation of regression coefficient for two-stage sampling design. *J. Ind. Agril. Statist.*, **71(1)**, 1-6.

Deville, J.C. and Sarndal, C.E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.*, **87**, 376-382.

Holt, D., Smith, T. M. F. and Winter, P. D. (1980). Regression analysis of data from complex surveys. *J. Roy. Statist. Soc*., *A (General)*, **143**, 474-487.

Kish, L. and Frankel, M. R. (1974). Inference from complex samples. *J. Roy. Statist. Soc.*, **B-36**, 1-37.

Plikusas, A. and Pumputis, D. (2007). Calibrated estimators of the population covariance. *Acta Applicandae Mathematicae*, **97**, 177-187.

Plikusas, A. and Pumputis, D. (2010). Estimation of finite population covariance using calibration. *Nonlinear Analysis: Modelling and Control*, **15(3)**, 325-340.

Särndal, C.E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer Verlag, New York.

Scott, A.J. and Holt, D. (1982). The effect of two stage sampling on ordinary least squares method. *J. Amer. Statist. Assoc.*, **77**,844-854.

Wu, C. and Luan, Y. (2003). Optimal calibration estimators under two-phase sampling. *J. Off. Statist.*, **19(2)**, 119-131.