Available online at www.isas.org.in/jisas
Journal of the Indian Society of Agricultural Statistics 71(1) 2017 1–6

Calibration Estimation of Regression Coefficient for Two-stage Sampling Design

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

Received 01 July 2016; Revised 17 October 2016; Accepted 20 October 2016

SUMMARY

For studying the relationship between different variables, regression analysis is a widely used technique. The conventional ordinary least square estimator of regression coefficient is not suitable for complex survey data. In this paper, a calibration approach based estimator of finite population regression coefficient has been developed for survey data involving two-stage sampling design. The expression for its variance and variance estimator is also obtained. The improved performance of the proposed estimator is demonstrated through a real data based simulation study.

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

1. INTRODUCTION

Survey statisticians draw sample from the population in order to provide inference about the population parameters. Survey data are generally multivariate in nature and therefore, many a times, the objective of the survey is to establish the pattern of relationship between variables rather than estimation of simple parameters like means or totals. When the variables are quantitative in nature and the interest is to find causal relationship then regression analysis may be an appropriate method. Broadly, in the context of survey data, estimation of parameters of finite population is based on two approaches. One of the approaches is repeated two-step sampling from an infinite population which is known as super population theory for finite population sampling. Another is repeated sampling from a finite fixed population which is known as classical finite population sampling theory. In the context of estimation of parameters of super-population since a linear model is postulated the ordinary least squares approach can be used for estimation of parameters. A key assumption in this approach is that sample elements are independent and identically distributed. This assumption of independence holds good if the data are collected through simple random sampling with replacement. But it does not hold good for other sampling schemes. Now a days, most of the survey designs are complex in nature involving stratification, unequal probabilities of selection, clustering, multi-stages and multi-phases etc. From the regression analysis point of view, any deviation from independence assumption leads to complications in the form of error variance-covariance model. Even cluster sampling which involves only choice of proper sampling units and is relatively a simple sampling scheme is considered somewhat complex from regression analysis point of view. In case of large scale surveys, stratified multistage sampling design is widely used. Here, also the units in a stratum are relatively homogenous which violates the assumption of independence of sample elements required for ordinary least squares estimation of finite population regression coefficient. One of the alternatives may be to use other method of estimation like maximum likelihood estimation as suggested by Nathon and Holt (1980). In the classical finite population sampling theory, regression analysis of survey data requires survey weight of sample

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

elements to be incorporated in the analysis. Modified approaches such as to use sampling design weights in the estimation procedure has been used by Kish and Frankel (1974).

If there is availability of auxiliary information along with the variables under study then the theory of calibration approach proposed by Deville and Sarndal (1992) may be used for estimation of finite population regression coefficient in the case of complex survey data. For example, there is relationship between yield and fertilizer use (FAO 1981). Let yield (y) be dependent variable and independent variable be fertilizer use (x). Suppose an auxiliary variable associated with dependent variable is available example; dependent variable yield (y) is correlated with auxiliary variable minutes of sunshine or date of sowing (Jasemi *et al*. 2013). Similar to this, there may be situation when auxiliary variable associated with independent variable may be available example; fertilizer use (x) applied is correlated with oil price or subsidy in price of fertilizer (Bain 2012).

In Section 2 we discuss calibration estimation of population regression coefficient under two-stage sampling design when population level auxiliary information is available at both psu and ssu level. Section 3 presents variance estimation of the proposed estimators. In Section 4 empirical evaluation of the developed estimators is provided. Finally, Section 5 presents concluding remarks.

2. THE PROPOSED ESTIMATOR

Let us consider a finite population $U = (1, 2,\ldots, k,\ldots, N)$ which is grouped into N_I clusters as $U_1, U_2, \ldots, U_i,\ldots, U_{N_I}$ with sizes of the clusters as $N_1, N_2,\ldots, N_i, \ldots, N_{N_I}$, respectively. Thus, $U = \bigcup_{i=1}^{N_I} U_i$ and $N = \sum_{i=1}^{N_I} N_i$. These clusters are called primary stage units (psus) and the sampling units within the clusters (psus) are called second stage units (ssus). At the first stage, a sample of psus s_I of size n_I is selected from a population of psus U_I of size N_I by using any probability sampling scheme.

Let, the first order and second order inclusion probability at the first stage be π_{Ii} and π_{Iij} respectively.

At the second stage, a sample of units s_i of size n_i is selected from the i^{th} psus, U_i of size N_i, $\forall i \in s_I$ 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, the first order and second order inclusion probability at the second stage be $\pi_{k/i}$ and $\pi_{kl/i}$ respectively.

Let y and x be the study variables. Let us assume that auxiliary variable z is associated with y and auxiliary variable p is associated with x. Let y_{ik}, x_{ik}, z_{ik}, and p_{ik} be values of the variables associated with the k^{th} unit of i^{th} selected psu. The population total of y is given by $t_y = \sum_{k=1}^{N} y_{ik} = \sum_{i=1}^{N_I} t_{iy}$, 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 $t_x = \sum_{k=1}^{N} x_{ik} = \sum_{i=1}^{N_I} t_{ix}$, where $t_{ix} = \sum_{k=1}^{N_i} x_{ik}$ is the i^{th} psu total of x. Let Z_i and P_i be the i^{th} psu total of auxiliary variables z and p respectively. Thus, $Z_i = \sum_{k=1}^{N_i} z_{ik}$ and $P_i = \sum_{k=1}^{N_i} p_{ik}$.

Under this study, 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 $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N_I}\sum_{k=1}^{N_i} y_{ik}$.

Now, the π-estimator of population regression coefficient under two-stage sampling is given by

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

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

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

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

Now, π-estimator of population regression coefficient under two-stage sampling may also be expressed as

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

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

Here, it is assumed that population level auxiliary information is available at both psu and ssu level, i.e. unit level auxiliary information is known. Thus, z_{ik} and p_{ik} is known $\forall i = 1,...,N_I$ and $k = 1,...,N_i$. In this case the calibration constraint are defined as $\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} w_{ik}z_{ik} = \sum_{i=1}^{N_I}\sum_{k=1}^{N_i} z_{ik}$ and $\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} w_{ik}p_{ik} = \sum_{i=1}^{N_I}\sum_{k=1}^{N_i} p_{ik}$, where w_{ik} is the calibrated weight corresponding to the design weight a_{ik}.

Here, the chi-square distance function measuring the distance between w_{ik} and a_{ik} is given by $\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} \frac{(w_{ik} - a_{ik})^2}{a_{ik}q_{ik}}$.

Thus, the objective function for minimization is given by

$$\phi = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} \frac{(w_{ik} - a_{ik})^2}{a_{ik}q_{ik}} - 2\lambda_1\left(\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} w_{ik}z_{ik} - \sum_{i=1}^{N_I}\sum_{k=1}^{N_i} z_{ik}\right) - 2\lambda_2\left(\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} w_{ik}p_{ik} - \sum_{i=1}^{N_I}\sum_{k=1}^{N_i} p_{ik}\right).$$

This objective function is minimized subject to the calibration constraint using Lagrange multiplier approach to obtain the calibrated weight, w_{ik}. Finally the calibrated weights are obtained as

$$w_{ik} = a_{ik}\left\{1 + q_{ik}(\lambda_1 z_{ik} + \lambda_2 p_{ik})\right\}$$

where,

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

and

$$\lambda_2 = \frac{(\sum_{i=1}^{N_I}\sum_{k=1}^{N_i} p_{ik} - \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}p_{ik})\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}q_{ik}z_{ik}^2 - (\sum_{i=1}^{N_I}\sum_{k=1}^{N_i} z_{ik} - \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}z_{ik})\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}q_{ik}p_{ik}z_{ik}}{(\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}q_{ik}z_{ik}^2)(\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}q_{ik}p_{ik}^2) - (\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}q_{ik}p_{ik})^2}.$$

Here, q_{ik} is a positive constant and for the particular case $q_{ik} = 1$, the calibrated weights are given by

$$w_{ik} = a_{ik}\left(1 + \lambda_1 z_{ik} + \lambda_2 p_{ik}\right)$$

where,

$$\lambda_1 = \frac{(\sum_{i=1}^{N_I}\sum_{k=1}^{N_i} z_{ik} - \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}z_{ik})\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}p_{ik}^2 - (\sum_{i=1}^{N_I}\sum_{k=1}^{N_i} p_{ik} - \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}p_{ik})\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}p_{ik}z_{ik}}{(\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}z_{ik}^2)(\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}p_{ik}^2) - (\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}p_{ik})^2}$$

and

$$\lambda_2 = \frac{(\sum_{i=1}^{N_I}\sum_{k=1}^{N_i} p_{ik} - \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}p_{ik})\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}z_{ik}^2 - (\sum_{i=1}^{N_I}\sum_{k=1}^{N_i} z_{ik} - \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}z_{ik})\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}p_{ik}z_{ik}}{(\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}z_{ik}^2)(\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}p_{ik}^2) - (\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}p_{ik})^2}$$

Here, the calibrated estimators of population total of study variables y and x are given by $\hat{t}_{y\pi}^{c(3)} = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} w_{ik}y_{ik}$ and $\hat{t}_{x\pi}^{c(3)} = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} w_{ik}x_{ik}$ respectively.

Thus, the calibrated estimator of population regression coefficient under the availability of both psu and ssu level auxiliary information is given by

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

3. VARIANCE ESTIMATION

The developed calibrated estimator of population regression coefficient is non-linear in nature. There are two approaches for variance estimation of nonlinear estimator: the Taylor series linearization approach and sample reuse approach. In this paper, Taylor series linearization technique is used to derive an approximate variance of the estimator as well as the variance estimator.

Under this case the calibrated estimator, $\hat{B}_{\pi c}^{(3)}$ can also be expressed as

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

where,

$$w_{ik} = a_{ik}\left\{1 + \left(\sum_{i=1}^{N_I}\sum_{k=1}^{N_i}\mathbf{r}'_{ik} - \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}\mathbf{r}'_{ik}\right)\left(\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}\mathbf{r}_{ik}\mathbf{r}'_{ik}\right)^{-1}\mathbf{r}_{ik}\right\},$$

$\mathbf{r}_{ik} = \left(z_{ik}, p_{ik}\right)'$.

Let us assume, $\mathbf{t}_{\mathbf{r}} = \sum_{i=1}^{N_I}\sum_{k=1}^{N_i}\mathbf{r}_{ik}$, $\hat{\mathbf{t}}_{\mathbf{r}} = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}\mathbf{r}_{ik}$ and $\hat{\mathbf{A}}_{\mathbf{r}} = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}\mathbf{r}_{ik}\mathbf{r}'_{ik}$

Then, $\hat{B}_{\pi c}^{(3)} = \frac{U}{V}$ such that

$$U = \hat{t}_{xy} + \left(\mathbf{t}'_{\mathbf{r}} - \hat{\mathbf{t}}'_{\mathbf{r}}\right)\hat{\mathbf{A}}_{\mathbf{r}}^{-1}\hat{\mathbf{t}}_{xy\mathbf{r}} - \frac{2}{N}\left\{\hat{t}_y + \left(\mathbf{t}'_{\mathbf{r}} - \hat{\mathbf{t}}'_{\mathbf{r}}\right)\hat{\mathbf{A}}_{\mathbf{r}}^{-1}\hat{\mathbf{t}}_{y\mathbf{r}}\right\}\left\{\hat{t}_x + \left(\mathbf{t}'_{\mathbf{r}} - \hat{\mathbf{t}}'_{\mathbf{r}}\right)\hat{\mathbf{A}}_{\mathbf{r}}^{-1}\hat{\mathbf{t}}_{x\mathbf{r}}\right\} + \frac{1}{N^2}\left\{\hat{N} + \left(\mathbf{t}'_{\mathbf{r}} - \hat{\mathbf{t}}'_{\mathbf{r}}\right)\hat{\mathbf{A}}_{\mathbf{r}}^{-1}\hat{\mathbf{t}}_{\mathbf{r}}\right\}\left\{\hat{t}_y + \left(\mathbf{t}'_{\mathbf{r}} - \hat{\mathbf{t}}'_{\mathbf{r}}\right)\hat{\mathbf{A}}_{\mathbf{r}}^{-1}\hat{\mathbf{t}}_{y\mathbf{r}}\right\}\left\{\hat{t}_x + \left(\mathbf{t}'_{\mathbf{r}} - \hat{\mathbf{t}}'_{\mathbf{r}}\right)\hat{\mathbf{A}}_{\mathbf{r}}^{-1}\hat{\mathbf{t}}_{x\mathbf{r}}\right\},$$

$$V = \hat{t}_{xx} + \left(\mathbf{t}'_{\mathbf{r}} - \hat{\mathbf{t}}'_{\mathbf{r}}\right)\hat{\mathbf{A}}_{\mathbf{r}}^{-1}\hat{\mathbf{t}}_{xx\mathbf{r}} - \frac{2}{N}\left\{\hat{t}_x + \left(\mathbf{t}'_{\mathbf{r}} - \hat{\mathbf{t}}'_{\mathbf{r}}\right)\hat{\mathbf{A}}_{\mathbf{r}}^{-1}\hat{\mathbf{t}}_{x\mathbf{r}}\right\}^2 + \frac{1}{N^2}\left\{\hat{N} + \left(\mathbf{t}'_{\mathbf{r}} - \hat{\mathbf{t}}'_{\mathbf{r}}\right)\hat{\mathbf{A}}_{\mathbf{r}}^{-1}\hat{\mathbf{t}}_{\mathbf{r}}\right\}\left\{\hat{t}_x + \left(\mathbf{t}'_{\mathbf{r}} - \hat{\mathbf{t}}'_{\mathbf{r}}\right)\hat{\mathbf{A}}_{\mathbf{r}}^{-1}\hat{\mathbf{t}}_{x\mathbf{r}}\right\}^2$$

where,

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

$$\hat{t}_y = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}y_{ik},\ \hat{\mathbf{t}}_{xy\mathbf{r}} = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}x_{ik}y_{ik}\mathbf{r}_{ik},$$

$$\hat{\mathbf{t}}_{x\mathbf{r}} = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}x_{ik}\mathbf{r}_{ik},\ \hat{\mathbf{t}}_{y\mathbf{r}} = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}y_{ik}\mathbf{r}_{ik},$$

$$\hat{N} = \sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}.$$

Here, U and V are functions of several estimators of population totals,

$U = f\left(\hat{t}_{xy}, \hat{\mathbf{t}}_{\mathbf{r}}, \hat{\mathbf{A}}_{\mathbf{r}}, \hat{\mathbf{t}}_{xy\mathbf{r}}, \hat{t}_y, \hat{\mathbf{t}}_{y\mathbf{r}}, \hat{t}_x, \hat{\mathbf{t}}_{x\mathbf{r}}, \hat{N}\right)$ and

$V = f\left(\hat{t}_{xx}, \hat{\mathbf{t}}_{\mathbf{r}}, \hat{\mathbf{A}}_{\mathbf{r}}, \mathbf{t}_{xx\mathbf{r}}, \hat{t}_x, \hat{\mathbf{t}}_{x\mathbf{r}}, \hat{N}\right)$.

Therefore, the estimator $\hat{B}_{\pi c}^{(3)}$ itself is a function of several estimators of population totals

$$\hat{B}_{\pi c}^{(3)} = f\left(\hat{t}_{xy}, \hat{\mathbf{t}}_{\mathbf{r}}, \hat{\mathbf{A}}_{\mathbf{r}}, \hat{\mathbf{t}}_{xy\mathbf{r}}, \hat{t}_y, \hat{\mathbf{t}}_{y\mathbf{r}}, \hat{t}_x, \hat{\mathbf{t}}_{x\mathbf{r}}, \hat{t}_{xx}, \hat{\mathbf{t}}_{xx\mathbf{r}}, \hat{N}\right)$$

These estimators $\hat{t}_{xy}, \hat{\mathbf{t}}_{\mathbf{r}}, \hat{\mathbf{A}}_{\mathbf{r}}, \hat{\mathbf{t}}_{xy\mathbf{r}}, \hat{t}_y, \hat{\mathbf{t}}_{y\mathbf{r}}, \hat{t}_x, \hat{\mathbf{t}}_{x\mathbf{r}}, \hat{t}_{xx}, \hat{\mathbf{t}}_{xx\mathbf{r}}, \hat{N}$ are Horvitz-Thompson estimators or π estimators, and therefore are unbiased estimators of totals

$$t_{xy} = \sum_{i=1}^{N_I}\sum_{k=1}^{N_i} x_{ik}y_{ik},\ \mathbf{t}_{\mathbf{r}} = \sum_{i=1}^{N_I}\sum_{k=1}^{N_i}\mathbf{r}_{ik},$$

$$\mathbf{A}_{\mathbf{r}} = \sum_{i=1}^{N_I}\sum_{k=1}^{N_i}\mathbf{r}_{ik}\mathbf{r}'_{ik},\ \mathbf{t}_{xy\mathbf{r}} = \sum_{i=1}^{N_I}\sum_{k=1}^{N_i} x_{ik}y_{ik}\mathbf{r}_{ik},$$

$$t_y = \sum_{i=1}^{N_I}\sum_{k=1}^{N_i} y_{ik},\ \mathbf{t}_{y\mathbf{r}} = \sum_{i=1}^{N_I}\sum_{k=1}^{N_i} y_{ik}\mathbf{r}_{ik},$$

$$t_x = \sum_{i=1}^{N_I}\sum_{k=1}^{N_i} x_{ik},\ \mathbf{t}_{x\mathbf{r}} = \sum_{i=1}^{N_I}\sum_{k=1}^{N_i} x_{ik}\mathbf{r}_{ik},$$

$$t_{xx} = \sum_{i=1}^{N_I}\sum_{k=1}^{N_i} x_{ik}x_{ik},\ \mathbf{t}_{xx\mathbf{r}} = \sum_{i=1}^{N_I}\sum_{k=1}^{N_i} x_{ik}x_{ik}\mathbf{r}_{ik},$$

$$N = \sum_{i=1}^{N_I} N_i.$$

respectively.

Since $\hat{B}_{\pi c}^{(3)}$ is a function of the unbiased estimators mentioned above. Thus, by using Taylor series linearization method, we approximate the function $\hat{B}_{\pi c}^{(3)}$ by a linear one at the mean point $(\hat{t}_{xy}, \hat{\mathbf{t}}_{\mathbf{r}}, \hat{\mathbf{A}}_{\mathbf{r}}, \hat{\mathbf{t}}_{xy\mathbf{r}}, \hat{t}_y, \hat{\mathbf{t}}_{y\mathbf{r}}, \hat{t}_x, \hat{\mathbf{t}}_{x\mathbf{r}}, \hat{t}_{xx}, \hat{\mathbf{t}}_{xx\mathbf{r}}, \hat{N})$ = $(t_{xy}, \mathbf{t}_{\mathbf{r}}, \mathbf{A}_{\mathbf{r}}, \mathbf{t}_{xy\mathbf{r}}, t_y, \mathbf{t}_{y\mathbf{r}}, t_x, \mathbf{t}_{x\mathbf{r}}, t_{xx}, \mathbf{t}_{xx\mathbf{r}}, N)$. Finally, the Taylor linearized estimator of $\hat{B}_{\pi c}^{(3)}$ is obtained as

$$\hat{B}_{\pi c}^{(3)} \approx B + \frac{1}{S_x^2}\sum_{i=1}^{n_I}\sum_{k=1}^{n_i} a_{ik}\left(x_{ik} - \bar{X}\right)E_{ik} - \frac{1}{S_x^2}\sum_{i=1}^{N_I}\sum_{k=1}^{N_i}\left(x_{ik} - \bar{X}\right)E_{ik}\mathbf{r}'_{ik}\mathbf{A}_{\mathbf{r}}^{-1}\left(\hat{\mathbf{t}}_{\mathbf{r}} - \mathbf{t}_{\mathbf{r}}\right)$$

$$\approx B+\frac{1}{S_x^2}\left[\sum_{i=1}^{n_I}\sum_{k=1}^{n_i}a_{ik}\left\{\left(x_{ik}-\bar{X}\right)E_{ik}-\mathbf{r}_{ik}'\mathbf{A}_{\mathbf{r}}^{-1}\sum_{i=1}^{N_I}\sum_{k=1}^{N_i}\left(x_{ik}-\bar{X}\right)E_{ik}\mathbf{r}_{ik}\right\}\right]$$

$$\approx B+\frac{1}{S_x^2}\left[\sum_{i=1}^{n_I}a_{Ii}\sum_{k=1}^{n_i}a_{k/i}\left\{\left(x_{ik}-\bar{X}\right)E_{ik}-\mathbf{r}_{ik}'\mathbf{A}_{\mathbf{r}}^{-1}\sum_{i=1}^{N_I}\sum_{k=1}^{N_i}\left(x_{ik}-\bar{X}\right)E_{ik}\mathbf{r}_{ik}\right\}\right]$$

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

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

Thus, the approximate variance of the calibrated estimator $\hat{B}_{\pi c}^{(3)}$ using Taylor series linearization method is obtained as

$$V\left(\hat{B}_{\pi c}^{(3)}\right)\approx \frac{1}{\left(S_x^2\right)^2}\left[\sum_{i=1}^{N_I}\sum_{i=1}^{N_I}\breve{\Delta}_{Iij}\frac{\sum_{k=1}^{N_i}\left\{\left(x_{ik}-\bar{X}\right)E_{ik}-\mathbf{r}_{ik}'\mathbf{A}_{\mathbf{r}}^{-1}\sum_{i=1}^{N_I}\sum_{k=1}^{N_i}\left(x_{ik}-\bar{X}\right)E_{ik}\mathbf{r}_{ik}\right\}}{\pi_{Ii}}\frac{\sum_{k=1}^{N_j}\left\{\left(x_{jk}-\bar{X}\right)E_{jk}-\mathbf{r}_{jk}'\mathbf{A}_{\mathbf{r}}^{-1}\sum_{j=1}^{N_I}\sum_{k=1}^{N_i}\left(x_{jk}-\bar{X}\right)E_{jk}\mathbf{r}_{jk}\right\}}{\pi_{Ij}}\right.$$
$$\left.+\sum_{i=1}^{N_I}\frac{1}{\pi_{Ii}}\sum_{k=1}^{N_i}\sum_{l=1}^{N_i}\breve{\Delta}_{kl/i}\frac{\left(x_{ik}-\bar{X}\right)E_{ik}-\mathbf{r}_{ik}'\mathbf{A}_{\mathbf{r}}^{-1}\sum_{i=1}^{N_I}\sum_{k=1}^{N_i}\left(x_{ik}-\bar{X}\right)E_{ik}\mathbf{r}_{ik}}{\pi_{k/i}}\frac{\left(x_{il}-\bar{X}\right)E_{il}-\mathbf{r}_{il}'\mathbf{A}_{\mathbf{r}}^{-1}\sum_{i=1}^{N_I}\sum_{l=1}^{N_i}\left(x_{il}-\bar{X}\right)E_{il}\mathbf{r}_{il}}{\pi_{l/i}}\right]$$

The variance estimator of $\hat{B}_{\pi c}^{(3)}$ is given by

$$\hat{V}\left(\hat{B}_{\pi c}^{(3)}\right)= \frac{1}{\left(\hat{S}_x^2\right)^2}\left[\sum_{i=1}^{n_I}\sum_{i=1}^{n_I}\breve{\Delta}_{Iij}\frac{\sum_{k=1}^{n_i}a_{k/i}\left\{\left(x_{ik}-\hat{\bar{X}}\right)\hat{E}_{ik}-\mathbf{r}_{ik}'\mathbf{A}_{\mathbf{r}}^{-1}\sum_{i=1}^{n_I}\sum_{k=1}^{n_i}a_{ik}\left(x_{ik}-\hat{\bar{X}}\right)\hat{E}_{ik}\mathbf{r}_{ik}\right\}}{\pi_{Ii}}\frac{\sum_{k=1}^{n_j}a_{k/j}\left\{\left(x_{jk}-\hat{\bar{X}}\right)\hat{E}_{jk}-\mathbf{r}_{jk}'\mathbf{A}_{\mathbf{r}}^{-1}\sum_{j=1}^{n_I}\sum_{k=1}^{n_i}a_{jk}\left(x_{jk}-\hat{\bar{X}}\right)\hat{E}_{jk}\mathbf{r}_{jk}\right\}}{\pi_{Ij}}\right.$$
$$\left.+\sum_{i=1}^{n_I}\frac{1}{\pi_{Ii}}\sum_{k=1}^{n_i}\sum_{l=1}^{n_i}\breve{\Delta}_{kl/i}\frac{\left(x_{ik}-\hat{\bar{X}}\right)\hat{E}_{ik}-\mathbf{r}_{ik}'\mathbf{A}_{\mathbf{r}}^{-1}\sum_{i=1}^{n_I}\sum_{k=1}^{n_i}a_{ik}\left(x_{ik}-\hat{\bar{X}}\right)\hat{E}_{ik}\mathbf{r}_{ik}}{\pi_{k/i}}\frac{\left(x_{il}-\hat{\bar{X}}\right)\hat{E}_{il}-\mathbf{r}_{il}'\mathbf{A}_{\mathbf{r}}^{-1}\sum_{i=1}^{n_I}\sum_{l=1}^{n_i}a_{il}\left(x_{il}-\hat{\bar{X}}\right)\hat{E}_{il}\mathbf{r}_{il}}{\pi_{l/i}}\right]$$

where,

$$\hat{S}_x^{\,2}=\sum_{i=1}^{n_I}\sum_{k=1}^{n_i}a_{ik}\left(x_{ik}-\hat{\bar{X}}\right)^2,$$

$$\hat{\bar{X}}=\frac{\hat{t}_{x\pi}}{N},\ \hat{\bar{Y}}=\frac{\hat{t}_{y\pi}}{N},$$

$$\breve{\Delta}_{Iij}=\frac{\Delta_{Iij}}{\pi_{Iij}}=\frac{\pi_{Iij}-\pi_{Ii}\pi_{Ij}}{\pi_{Iij}},$$

$$\hat{E}_{ik}=\left(y_{ik}-\hat{\bar{Y}}\right)-\hat{B}_{\pi}\left(x_{ik}-\hat{\bar{X}}\right),$$

$$\breve{\Delta}_{kl/i}=\frac{\Delta_{kl/i}}{\pi_{kl/i}}=\frac{\pi_{kl/i}-\pi_{k/i}\pi_{l/i}}{\pi_{kl/i}}.$$

4. EMPIRICAL EVALUATIONS

In this Section we report the results from simulation study that illustrate the performance of the proposed estimators. In particular, we consider following two estimators of the population regression coefficient:

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

(ii) Calibrated estimator, $\hat{B}_{\pi c}^{(3)}$ given by (2) (denoted as Est-CAL),

The performance of the estimators was evaluated by percentage absolute relative bias (ARB) and percentage relative root mean squared error (RRMSE), defined by

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

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

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

A real dataset of 284 municipalities of Sweden, referred to as MU284 population was used for simulation. Thus, here population size N = 284. The municipalities are grouped into 50 clusters each containing 5 to 9 municipalities. Here, the aim was to estimate population regression coefficient between variables revenues from the 1985 Municipal taxation (RMT85, measured in millions of kronor) and 1985 population (P85, in thousands) using number of municipal employees in 1984 (ME84) and 1975 population (P75, in thousands) as the auxiliary variables respectively. The correlations between the variables are presented in Table 1.

Table 1. Correlation between different variables in MU284 data

Variables	RMT85	P85	ME84	P75
RMT85	1	0.961	0.999	0.967
P85	0.961	1	0.965	0.998
ME84	0.999	0.965	1	0.971
P75	0.967	0.998	0.971	1

From this population, a two stage sample of size n_s = 80 was selected by drawing 20 psus at the first stage and 4 units from each selected psus at the second stage. The sampling scheme used in both stages is simple random sampling without replacement. Then the various estimators of the population regression coefficient were computed using the sample data. The Monte Carlo simulation was run M=5000 times. Simulation study was done using R software. The values of percentage absolute relative bias and the values of percentage relative root mean square error of different estimators are reported in Table 2.

Table 2. Percentage absolute relative bias (ARB, %) and percentage relative root mean square error (RRMSE, %) of different estimators

Estimator	*ARB, %*	*RRMSE, %*
Est- π	18.9924	21.8049
Est-CAL	15.7059	17.9376

These results in Table 2 show that the value of percentage absolute relative bias for the π-estimator is more than the Est-CAL (calibrated estimator). Therefore, it can be concluded that in terms of absolute relative bias the estimator Est-CAL shows better performance. In the case of percentage relative root mean square error, it is higher for the π-estimator as compared to Est-CAL. Therefore, in terms of criterion of percentage relative root mean square error also the calibrated estimator Est-CAL gives better performance.

5. CONCLUDING REMARKS

This article discusses the calibrated estimator of population regression coefficient under the availability of auxiliary information at both psu and ssu level and its variance estimation using Taylor series linearization approach. The calibration estimator based on both psu and ssu level auxiliary information gives better performance than the simple π-estimator of population regression coefficient.

ACKNOWLEDGEMENTS

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 gratefully acknowledges the INSPIRE Fellowship provided by Department of Science and Technology, Government of India.

REFERENCES

Aditya, K., Sud, U.C., Chandra, H. and Jain, V.K. (2014). A study on calibration estimators of finite population total for two stage sampling design. Project report submitted to ICAR-IASRI.

Bain, B. (2012). Fertilizer trends. Nomura Global Chemistry Leaders Conference.

Devi, M. Memita, Bathla, H.V.L., Sud, U.C. and Sethi, I.C. (2005). On the estimation of finite population regression coefficient. *J. Ind. Soc. Agril. Statist.*, **59(2)**, 118-125.

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

FAO (1981). Crop production levels and fertilizer use. *FAO Fertilizer and Plant Nutrition Bulletin 2*. Food and Agriculture Organization of the United Nations, Rome. **69**.

Jesami, M., Darabi, F., Naseri, R., Naserirad, H., and Bazdar, H. (2013). Effect of planting date and nitrogen fertilizer application on grain yield and yield components in maize (sc 704), *Amer.-Euras. J. Agri. Environ. Sci.*, **13(7)**, 914-919.

Kish, L. and Frankel, M.R. (1974). Inference from complex samples. *J. Roy. Statist. Soc.*, **B36**, 1-37.

Nathan, G. and Holt, D. (1980). Effect of survey design on regression analysis. *J. Roy. Statist. Soc.*, **B42**, 377-386.

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 Control*, **15(3)**, 325-340.

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

Sud, U.C. (1987). Some contributions to regression analysis with survey data. Ph.D. thesis. IARI, New Delhi.