Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 72(3) 2018 193–199

Calibration Estimator of Regression Coefficient using Two Auxiliary Variables

Vandita Kumari, Hukum Chandra and L.M. Bhar
ICAR-Indian Agricultural Statistics Research Institute, New Delhi

Received 21 May 2018; Revised 07 August 2018; Accepted 25 August 2018

SUMMARY

Surveys are often based on complex sample designs where sampling units frequently have different probabilities of being selected. In the survey data analysis, sampling weights must be used to incorporate the sample designs. Regression coefficients are estimated to find the relationship between the study and auxiliary variables. Kish and Frankel (1974) deliberated the use of sampling weights in the estimation of regression coefficients. This paper describes calibration based approach to estimate the regression coefficient using two auxiliary variables. The variance estimation of proposed estimator is also developed. The empirical results based on synthetic and real population show that the proposed estimator, in terms of percent relative bias and percent relative root mean square error, performs better than the existing estimator. The proposed variance estimator shows a satisfactory performance in empirical evaluation.

Keywords: Calibrated weights, Regression coefficients, Auxiliary variable.

1. INTRODUCTION

In sample surveys, sampling weights are being used in the analysis of data. The weights are attached to each unit in the sample such that the sampling unit represents the population from which it is selected. Sampling weights also called as design weights are the reciprocals of the sample inclusion probabilities (Horvitz and Thompson, 1952). In the case of availability of auxiliary variable these design weights are often modified in order to produce an efficient estimator of the population parameter by an approach commonly called as Calibration (Deville and Särndal, 1992). In this approach a new set of weights called as calibrated weights are obtained which are as close as possible to the design weights according to a distance measure subject to a set of constraints. Regression coefficient is estimated to find out the relationship between the study variable and an auxiliary variable. It is generally computed by using ordinary least square (OLS) techniques under the assumption that observations are independently identically distributed. But since the survey data are complex in nature OLS technique is often misleading and thus there need for modification of standard approach. Kish and Frankel (1974) considered design-based inference of finite population parameter. They proposed the use of probability weights for the estimation of finite population regression coefficient. Holt, Smith and Winter (1980) proposed a probability weighted least square method with complex survey data. Devi (2005) developed a double sampling based estimator of finite population regression coefficient. Wu and Fuller (2005) proposed an estimation of regression coefficients with unequal probability samples. Also, see Breidt and Opsomer (2017) who reviewed the design-based, and model-assisted approach for a complex survey data with an application to estimation of regression coefficients. Basak *et al.* (2018) developed calibration approach in the context of two-stage sampling design for the estimation of finite population regression coefficient.

Corresponding author: Vandita Kumari
E-mail address: vandita.kumari@icar.gov.in

There can be situation in survey sampling where the information on more than one auxiliary variable is available. For instance the planting area and the proportion of good seeds in agricultural engineering are two important auxiliary variables when estimating average cotton output (Lu, 2017). So, in a study to find the relationship between average cotton output and planting area, an another auxiliary variable is proportion of good seeds that are highly correlated to the study variable that would not appear explicitly in the regression relationship but can be used to enhance the precision of the estimator. Similarly in an another example, the breed of cow and climate is an important auxiliary attribute when estimating average milk yield (Rhone, 2008) and for a study of relationship between milk yield and climate, the breed of cow may be used as another auxiliary variable that is highly correlated to average milk yield. Motivated with the above arguments, the objective of the present work is to propose a calibrated estimator of the population regression coefficient using two auxiliary variables.

The rest of the article is organised as follows. The next section describes the weighted least square method and theoretical development for the estimation of population regression coefficient with modified weight using calibration approach. The expressions for variance and estimator of variance have also been developed in section 3. The results from model-based simulations have been used to illustrate the performances of the proposed estimator which is presented in section 4. The specific application to real data is presented in section 5. And, finally the concluding remarks with a discussion of potential avenues for further research are given in the section 6.

2. ESTIMATION OF REGRESSION COEFFICIENT

To start with we consider a finite population $U = (U_1, U_2, .., U_N)$ of size N from which a probability sample $s(s \subset U)$ of size n is drawn following a sample design denoted by p(.) The first and second order inclusion probabilities $\pi_i = p_r(i \in s)$ and $\pi_{ij} = p_r(i \text{ and } j \in s)$ are assumed to be strictly positive and known. Let y be the study variables defined on the population U and taking nonnegative values $y_1, y_2, \ldots, y_N$ where the totals of variables are given as $t_y = \sum_{i \in U} y_i$. We denote the two auxiliary variables as x and z having values $x_1, x_2, \ldots, \mathrm{x}_N$ and $z_1, z_2, \ldots, z_N$ that are correlated to the variables y. In this study we are interested in estimation of population regression coefficient (B) of y on x, given as

$$B = \frac{\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)\left(x_i - \bar{X}\right)}{\sum_{i=1}^{N}\left(x_i - \bar{X}\right)^2}, \tag{1}$$

where, $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ and $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$.

The ordinary least square estimator (OLS) of the population regression coefficient is given as

$$\hat{b}_{ols} = \frac{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2},$$

where, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

We can estimate the population regression coefficient using the Hortwitz Thompson estimator of population totals to obtain weighted ordinary least square estimator (WOLS), (see for example Särndal, Swensson and Wretman, 1992 p.195) and is given by

$$\hat{b}_{wols} = \frac{\sum_{i \in s} d_i \left(y_i - \frac{1}{N}\sum_{i \in s} d_i y_i\right)\left(x_i - \frac{1}{N}\sum_{i \in s} d_i x_i\right)}{\sum_{i \in s} d_i \left(x_i - \frac{1}{N}\sum_{i \in s} d_i x_i\right)^2}, \tag{2}$$

here, $d_i = \frac{1}{\pi_i}$ is the sampling design weights of elements i (i=1,2 …, N). The first order Taylor expansion of the $\hat{b}_{wols}$ is written as

$$\hat{b}_{wols} = b + \frac{1}{\sum_{i \in U}\left(x_i - \bar{X}\right)^2}\sum_{i \in s}\frac{1}{\pi_i}\left(x_i - \bar{X}\right)\left[\left(y_i - \bar{Y}\right) - b\left(x_i - \bar{X}\right)\right].$$

The approximate variance of estimator of regression coefficient is

$$AV\left(\hat{b}_{wols}\right) = \frac{1}{\left(\sum_{i \in U}\left(x_i - \bar{X}\right)^2\right)^2}\sum\sum_{i=j \in U}\Delta_{ij}\frac{\left(x_i - \bar{X}\right)E_i}{\pi_i}\frac{\left(x_j - \bar{X}\right)E_j}{\pi_j}, \tag{3}$$

where, $E_i = \left(y_i - \bar{Y}\right) - B\left(x_i - \bar{X}\right)$ and $\Delta_{ij} = \pi_{ij} - \pi_i \pi_j$.

The approximate estimator of variance of (3) is given by

$$A\hat{V}\left(\hat{b}_{wols}\right)=\frac{1}{\left(\sum_{i\in s}d_i\left(x_i-\hat{\bar{x}}\right)^2\right)^2}\sum_{i=j\in s}\sum\Delta'_{ij}\frac{\left(x_i-\hat{\bar{x}}\right)e_i}{\pi_i}\frac{\left(x_j-\hat{\bar{x}}\right)e_j}{\pi_j}, \qquad (4)$$

where, $e_i=\left(y_i-\hat{\bar{y}}\right)-\hat{b}_{wols}\left(x_i-\hat{\bar{x}}\right)$,

$\hat{\bar{y}}=\frac{1}{N}\sum_{i=1}^{n}d_iy_i$, $\hat{\bar{x}}=\frac{1}{N}\sum_{i=1}^{n}d_ix_i$ and $\Delta'_{ij}=\Delta_{ij}/\pi_{ij}$.

We now describe the proposed estimators of population regression coefficient. We assume that $Z=\sum_{i\in U}Z_i$, i.e. the population totals of the auxiliary variable related to study variable y is known. Here, we propose a calibration approach based estimator of finite population regression coefficient. Following the work of Deville and Särndal (1992), we modify the sampling design weights to obtain new calibrated weights w_i. For this purpose, we minimize the chi-square distance between the design weights d_i and calibrated weight w_i by considering distance measure function $\sum_{i\in s}\frac{\left(w_i-d_i\right)^2}{\left(d_iq_i\right)}$, subject to the calibration constraints as $\sum_{i\in s}w_iz_i=Z$. This is an optimization problem where we minimise the following function

$$\phi\left(w_i,\lambda\right)=\sum_{i\in s}\frac{\left(w_i-d_i\right)^2}{\left(d_iq_i\right)}\text{-}\ \lambda\left(\sum_{i\in s}w_iz_i-Z\right),$$

using Lagrange multiplier λ. Minimization of function $\phi\left(w_i,\lambda\right)$ gives new set of calibrated weights as $w_i=d_i+\frac{d_iz_iq_i}{\sum_{i\in s}d_iz_i^2q_i}\left(Z-\sum_{i\in s}d_iz_i\right)$.

Here q_i is suitably chosen constant and we consider $q_i=1$. This value is often used for regression type of estimation. With $q_i=1$, the revised weights are given as

$$w_i=d_i+\frac{d_iz_i}{\sum_{i\in s}d_iz_i^2}\left(Z-\sum_{i\in s}d_iz_i\right). \qquad (5)$$

Using this set of new calibrated weights w_i from equation (5) the calibrated estimator of regression coefficient can be written as

$$\hat{b}_{cal}=\frac{\sum_{i\in s}w_i\left(y_i-\frac{1}{N}\sum_{i\in s}w_iy_i\right)\left(x_i-\frac{1}{N}\sum_{i\in s}w_ix_i\right)}{\sum_{i\in s}w_i\left(x_i-\frac{1}{N}\sum_{i\in s}w_ix_i\right)^2}. \qquad (6)$$

3. ESTIMATION OF VARIANCE

Taylor series linearization is a popular method of variance estimation for complex statistics. Using this method, we approximate the calibrated regression estimator by a linear form.

For the construction of linearization estimator of variance of non-linear function firstly the quantity of the interest i.e. calibrated estimator of regression coefficient given in equation (6) is written as function of the totals of the variables i.e.

$$\hat{b}_{cal}=\frac{\hat{t}_{xy}+\hat{t}_{xyz}A-\frac{2}{N}\left\{\hat{t}_x+\hat{t}_{xz}A\right\}\left\{\hat{t}_y+\hat{t}_{yz}A\right\}+\frac{1}{N^2}\left\{\hat{t}_x+\hat{t}_{xz}A\right\}\left\{\hat{t}_y+\hat{t}_{yz}A\right\}\left\{\widehat{N}+\hat{t}_zA\right\}}{\hat{t}_{xx}+\hat{t}_{xxz}A-\frac{2}{N}\left\{\hat{t}_x+\hat{t}_{xz}A\right\}^2-\frac{1}{N^2}\left\{\hat{t}_x+\hat{t}_{xz}A\right\}^2\left\{\widehat{N}+\hat{t}_zA\right\}}$$

$$=f(\hat{t}_{xy},\hat{t}_{xyz},\hat{t}_x,\hat{t}_{xz},\hat{t}_y,\hat{t}_{yz},\hat{t}_z,\widehat{N},\hat{t}_{xx},\hat{t}_{xxz},\hat{t}_{zz}),$$

where, $\hat{t}_x=\sum_{i\in s}d_ix_i$, $\hat{t}_y=\sum_{i\in s}d_iy_i$, $\hat{t}_z=\sum_{i\in s}d_iz_i$

$\hat{t}_{xy}=\sum_{i\in s}d_ix_iy_i$, $\hat{t}_{xyz}=\sum_{i\in s}d_ix_iy_iz_i$, $\hat{t}_{xz}=\sum_{i\in s}d_ix_iz_i$,

$\hat{t}_{yz}=\sum_{i\in s}d_iy_iz_i$, $\hat{t}_{xx}=\sum_{i\in s}d_ix_i^2$, $\hat{t}_{xxz}=\sum_{i\in s}d_ix_i^2z_i$, $\widehat{N}=\sum_{i\in s}d_i$,

$A=\frac{\left(Z-\hat{t}_z\right)}{\hat{t}_{zz}}$ and $\hat{t}_{zz}=\sum_{i\in s}d_iz_i^2$.

These are the Horvitz Thompson estimators and hence unbiased estimators of their respective totals. Therefore it can be said that $\hat{b}_{cal}$ is a function of unbiased estimators. Then the partial derivative w.r.t. each argument is calculated and is evaluated at the population quantities to form the linearising constants. The linear part of Taylor series expansion of $\hat{b}_{cal}$ at the mean point is given by

$$\hat{b}_{cal}(l)=B+\frac{1}{\left[\sum_U\left(x_i-\bar{X}\right)^2\right]^2}\sum_{i\in s}d_i\left(x_i-\bar{X}\right)E_i-\frac{\sum_UZ_i\left(x_i-\bar{X}\right)E_i}{\sum_UZ_i^2}\left(\sum_{i\in s}d_iZ_i-Z\right)$$

$$=B+\frac{1}{\left[\sum_U\left(x_i-\bar{X}\right)^2\right]^2}\sum_{i\in s}d_iE_{xi}-\frac{\sum_UZ_iE_{xi}}{\sum_UZ_i^2}\left(\sum_{i\in s}d_iZ_i-Z\right)$$

where, $E_{xi}=\left(x_i-\bar{X}\right)E_i$.

The approximate variance can be written as

$$AV\left(\hat{b}_{cal}\right)=V\left(\hat{b}_{cal}(l)\right)=\frac{1}{\left[\sum_U\left(x_i-\bar{X}\right)^2\right]^2}V\left[\sum_{i\in s}d_i\left\{E_{xi}-Z_i\frac{\sum_U Z_iE_{xi}}{\sum_U Z_i^2}\right\}\right]$$

$$=\frac{1}{\left[\sum_U\left(x_i-\overline{X}\right)^2\right]^2}\sum_{i=}\sum_{j\in U}\Delta_{ij}\frac{G_i}{\pi_i}\frac{G_j}{\pi_j}, \qquad (7)$$

where, $G_i=E_{xi}-z_i\dfrac{\sum_U Z_iE_{xi}}{\sum_U Z_i^2}$.

The approximate estimator of variance is given by

$$A\hat{V}\left(\hat{b}_{cal}\right)=\frac{1}{\left[\sum_{i\in s}w_i\left(x_i-\tilde{x}\right)^2\right]^2}\sum_{i=}\sum_{j\in s}\Delta'_{ij}\frac{g_i}{\pi_i}\frac{g_j}{\pi_j}, \qquad (8)$$

where, $g_i=\left(x_i-\tilde{x}\right)e_i-z_i\dfrac{\sum_{i\in s}z_i\left(x_i-\tilde{x}\right)e_i}{\sum_{i\in s}z_i^2}$,

$e_i=\left(x_i-\tilde{x}\right)-\hat{b}_{cal}\left(y_i-\tilde{y}\right)$, $\tilde{x}=\sum_{i\in s}w_ix_i$ and $\tilde{y}=\sum_{i\in s}w_iy_i$.

In the case of SRSWOR, $AV\left(\hat{b}_{cal}\right)$ reduces to

$$AV\left(\hat{b}_{cal}\right)=\frac{(1-f)N^2\sum_{i\in U}G_i^2}{\left(\sum_{i\in U}\left(x_i-\bar{X}\right)^2\right)^2n(N-1)},$$

where, $f=n/N$

In the case of SRSWOR, $A\hat{V}\left(\hat{b}_{cal}\right)$ is given as

$$A\hat{V}\left(\hat{b}_{cal}\right)=\frac{(1-f)N^2\sum_{i\in s}g_i^2}{\left(\sum_{i\in s}w_i\left(x_i-\hat{\bar{x}}\right)^2\right)^2n(n-1)}.$$

4. SIMULATION STUDIES

In this section we illustrate the comparative performance of the existing weighted ordinary least square (WOLS) and the proposed calibrated (CAL) estimator of regression coefficient using the simulation studies based on synthetic population. The simulation studies are the common ways of illustrating and comparing the performance of proposed estimators under the assumed conditions. In our simulation study we consider a finite population of size *N*=5000 units. The variable of interest *y* was generated from the model $y_i=4+\beta_1z_i+\beta_2x_i+e_i; i=1,2,...,N$ where, errors e_i $(i=1,2,...,N)$ are generated from normal distribution with mean 0 and variance σ_e^2 i.e. $e_i\sim N(0,\sigma_e^2)$. Here, the auxiliary variables, $x_i,z_i=(i=1,2,...,N)$ are independently generated from normal distribution with mean 100 and variance σ_x^2 i.e. $x_i\sim N\left(100,\sigma_x^2\right)$ and chi-square i.e. $z_i\sim\chi^2(10)$ respectively. Here, we choose $\beta_1=\beta_2=1$ and fixed the values throughout the simulations. However, we chose different values of error variance σ_x^2 and σ_e^2 to generate different populations having different levels of correlation between the variables. In particular, we considered six different parameter sets. The values of different parameter sets used in the simulation studies are given in Table 1.

Table 1. Parameter sets used in simulation studies.

Parameter set	σ_x^2		(y,x)	$\rho(y,z)$
1A	4	2	0.40	0.88
1B	4	16	0.32	0.71
2A	9	2	0.55	0.80
2B	9	16	0.45	0.67
3A	16	2	0.66	0.73
3B	16	16	0.55	0.61

For each fixed finite population sample of size *n* was taken with simple random sample without replacement. Particularly, M = 5000 samples were drawn for each of the parameter sets to calculate the estimators of regression coefficient and estimate of variance. To examine the sensitivity of population size on the estimator we took three different sample sizes i.e. *n*=100, 200, 500 for our simulation study.

The performance of the developed calibrated estimators has been evaluated on the basis of two measures. These are absolute percentage relative bias (ARB,%) and percentage relative root mean square error (RRMSE,%) that are defined as follows. Let $\hat{b}$ denote the estimator of regression coefficient which is either weighted ordinary least square (WOLS) or calibrated estimator (CAL). Let $\hat{b}_k$ and $\hat{v}_k$ denote the estimator of regression coefficient and its estimate of variance respectively for the sample $k(k=1,...,M)$. The percentage absolute relative bias of an estimator $\hat{b}$ of the population regression coefficient *B* is given by

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

and relative root mean squared errors of these estimators are given as

$$RRMSE(b) = \left(\frac{1}{B} \sqrt{\frac{1}{M} \sum_{k=1}^{M} (\hat{b}_k - B)^2} \right) \times 100.$$

The percentage relative gain (RG,%) in the RRMSE of the proposed CAL estimator over WOLS estimator is given as

$$RG \;= \frac{WOLS - CAL}{CAL} \times 100.$$

The values of percentage relative bias (RB,%) of variance estimates of regression coefficients are also calculated to examine the behavior of estimate of variance with respect to the true variance. The RB is defined as

$$RB(\hat{v}) = \left(\frac{1}{M} \sum_{k=1}^{M} \frac{\hat{v}_k - V}{V} \right) \times 100 \;.$$

Here, the true variance (V) is the empirical simulation based variance calculated as $V = \frac{1}{M} \sum_{k=1}^{M} \left(\hat{b}_k - \frac{1}{M} \sum_{k=1}^{M} \hat{b}_k \right)^2$. The results for the estimators of population regression coefficients obtained from simulation studies are presented in Table 2 and 3. The corresponding results for the variance estimation are reported in Table 4.

The result in Table 2 shows that the relative biases of the estimators decrease as the sample sizes increases. The bias of the proposed CAL estimator is smaller than the existing WOLS estimator for all combination of sample sizes and parameter sets. From Table 3 we find that the relative root mean squared error both the estimators decreases with increase in sample size and there is percentage relative gain in root mean squared error for CAL as compared to the WOLS. As expected, the % relative gain in RRMSE is high for Set 1A and 1B. In these cases, the correlation between y and x is lower and correlation between y and z is higher as compared to other sets. Moreover, when the correlations between $\rho(y,x)$ and $\rho(y,z)$ are at par then the gain is very less. Overall, two point emerged from the results presented in the Table 2 and 3. First, both the values of relative bias and the values of relative root mean squared error decreases as sample size increase for both the existing WOLS estimator as well as the proposed CAL estimator. Second, in term of relative bias and the values of relative root mean squared error, the relative performance of the proposed CAL estimator as compared to the existing WOLS estimator improves with decrease in sample size.

Table 2. Values of percentage Absolute Relative Bias (ARB, %) of WOLS and CAL estimators from simulation studies.

n	**WOLS**	**CAL**	**WOLS**	**CAL**	**WOLS**	**CAL**
	Set1A		**Set2A**		**Set3A**	
100	18.49	11.4	12.43	9.56	9.37	8.06
200	12.86	9.54	8.65	7.56	6.51	6.13
500	7.87	6.98	5.29	5.17	3.98	4.03
	Set1B		Set2B		Set3B	
100	24.00	14.38	16.03	11.73	12.04	9.79
200	16.75	12.13	11.19	9.41	8.4	7.59
500	10.39	8.98	6.94	6.55	5.21	5.09

Table 3. Percentage relative root mean square error (RRMSE, %) and percentage relative gain in RRMSE (RG, %) of WOLS and CAL estimators from simulation studies.

n	**WOLS**	**CAL**	**% RG**	**WOLS**	**CAL**	**% RG**	**WOLS**	**CAL**	**% RG**
	Set1A			**Set2A**			**Set3A**		
100	23.21	14.30	62.31	15.61	11.65	33.99	11.76	9.78	20.25
200	16.14	11.75	37.36	10.85	9.22	17.68	8.17	7.50	8.93
500	9.87	8.53	15.71	6.64	6.34	4.73	5.00	4.97	0.60
	Set1B			Set2B			Set3B		
100	30.14	18.31	64.61	20.13	14.53	38.54	15.11	12.07	25.19
200	21.09	15.10	39.67	14.09	11.60	21.47	10.58	9.37	12.91
500	12.92	10.97	17.78	8.63	8.04	7.34	6.48	6.28	3.18

The empirical performance of estimation of variance obtained from the simulation studies are shown in Table 4. The results in Table 4 show that the relative bias of variance estimate reduces with increase in sample size for both WOLS and CAL estimators and for all combination of parameters sets. Overall, the variance estimates indicate satisfactory performance.

5. APPLICATION TO REAL DATA

In this section we illustrate an application of the proposed estimator with real data. We used the MU284 population given in Appendix C of Särndal, Swensson and Wretman (1992) having 284 units in the population. The variables used for estimation of regression coefficients are real estate values according

Table 4. True variance (V), estimate of variance ($\hat{v}$) and relative bias of the variance (RB, %) of WOLS and CAL estimators from simulation studies.

	V x 10^{-2}		**$\hat{v} \times 10^{-2}$**		**RB ,%**	
n	**WOLS**	**CAL**	**WOLS**	**CAL**	**WOLS**	**CAL**
Set 1A						
100	5.6716	1.7833	5.4121	2.3074	-4.58	29.39
500	1.0258	0.6755	0.9908	0.694	-3.42	2.74
Set1B						
100	9.1862	2.8656	8.8207	3.8718	-3.98	35.11
500	1.6889	1.0833	1.6224	1.153	-3.94	6.44
Set2A						
100	2.5207	1.1531	2.4054	1.334	-4.58	15.69
500	0.4559	0.3762	0.4403	0.3569	-3.41	-5.13
Set2B						
100	4.0828	1.8159	3.9203	2.2264	-3.98	22.61
500	0.7506	0.5974	0.7211	0.5902	-3.93	-1.2
Set3A						
100	1.4179	0.8192	1.3530	0.8772	-4.58	7.08
500	0.2564	0.2342	0.2477	0.2159	-3.40	-7.81
Set3B						
100	2.2965	1.275	2.2052	1.4574	-3.98	14.31
500	0.4222	0.3717	0.4056	0.356	-3.93	-4.22

to 1984 assessment (REV84) as variable of interest (y) and the auxiliary variables x and z are the total number of seats in municipal council in 1982 (S82) and the number of municipal employees in 1984 (ME84) respectively. In this population the correlation between y and x is 0.677 and that between y and z is 0.940. From this population we selected M=5000 samples each of sizes n (n = 25, 50, 75 and 100) by simple random sampling without replacement and population regression coefficients were estimated using two estimators (WOLS and CAL). The result of this application is given in the Table 5.

Table 5. Values of percentage absolute relative bias (ARB, %), percentage relative root mean squared error (RRMSE, %) and percentage relative gain in RRMSE (RG, %) using the real data.

n	**ARB, %**		**RRMSE, %**		
	WOLS	**CAL**	**WOLS**	**CAL**	**% RG**
25	45.93	44.96	56.12	50.06	12.11
50	40.54	37.14	46.36	41.96	10.49
75	34.48	31.71	38.89	36.52	6.49
100	29.05	26.24	32.72	31.18	4.94

From the results reported in Table 5 we can see that in terms the percentage absolute relative bias and percentage relative root mean squared error the proposed estimator CAL has performed consistently better than the WOLS. The percentage relative gain in RRMSE of the proposed estimator increases with decrease in sample size. The results clearly indicate that the proposed CAL estimator shows better performance both in terms of bias and efficiency in real data. The conclusions from real data are identical to the simulation studies based on synthetic population reported in Section 4.

6. CONCLUDING REMARKS

In this paper, a calibration estimator of regression coefficient using auxiliary variables correlated with the study variable has been developed. Our empirical evaluation on the basis of simulation studies and real data show that the proposed estimator is more efficient than the existing estimator. The proposed variance estimation indicates a reasonably good performance. In few simulation setup, the proposed variance estimator shows a biased results. Therefore, some alternative approaches like bootstrap and jacknife based variance estimation can be explored. The proposed estimator of regression coefficient uses two auxiliary variables but it is interesting to explore the use of multiple auxiliary variables. Authors are currently on these issues.

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.

REFERENCES

Breidt, F.J. and Opsomer, J.D. (2017). Model-Assisted Survey Estimation with Modern Prediction Techniques. *Statistical Science,* **32(2)**, 190-205.

Devi, M.M., 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 Särndal, C.E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.*, **87**, 376-382.

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

Rhone, J.A. (2008). Factors affecting milk yield, milk fat, milk quality, and economic performance of dairy farms in the central region of Thailand. Ph.D. thesis. University of Florida.

Horvitz, D.G. and D.J. Thompson. (1952). A generalization of sampling without replacement from afinite universe. *J. Amer. Statist. Assoc.*, **47**, 663-685.

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

Lu, J. (2017). Efficient Estimator of a Finite Population Mean Using Two Auxiliary Variables and Numerical Application in Agricultural, Biomedical, and Power Engineering. *Mathematical Problems in Engineering*, Article ID 8704734, 7 pages.

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

Basak, P., Sud, U.C. and Chandra H.(2018).Calibration estimation of regression coefficient for two-stage sampling design using single auxiliary variable.*J. Ind. Soc. Agril. Statist.***72(1)**, 1-6.