Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 68(3) 2014 333-341

Estimation of Finite Population Total for Skewed Data

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

Received 04 February 2014; Revised 12 February 2014; Accepted 14 February 2014

SUMMARY

In many surveys (for example, agriculture, business enterprises, income and expenditure surveys), data are typically skewed which contain few extreme values and linear model assumptions are questionable. Commonly used survey estimation methods for population total are based on normality assumption, that is, survey data are linear. As a consequence, these methods are both model biased and inefficient for skewed data. We describe estimation of finite population total for skewed data that are linear following a suitable transformation, in particular logarithmic transformation. We demonstrate the comparative performance of different estimators of population total for skewed data using both model based simulations as well as design based simulations. Empirical results clearly reveal that linear model based estimators are inefficient for skewed data.

Keywords: Skewed data, Transformation, Bias correction, Model calibration, Prediction.

1. INTRODUCTION

In analysis of survey data, standard estimation methods for population parameters assume that data are normal and linear model describes the data well. However, in agricultural, business enterprises and income and expenditure surveys data are typically skewed and linear models are questionable for such data. In particular, for such survey data, relationship between study variable and auxiliary variable may not be linear in their original scale, but can be linear in a transformed scale, *e.g*., the logarithmic (log) scale. For example, the total annual farm costs (TCC) in Australian Agricultural and Grazing Industries Survey data is skewed (see Fig. 1) and linear model is not an appropriate model. Under such circumstances, survey estimation based on a linear model may be both model biased and inefficient, and appropriate technique of estimation of finite population parameter is based on a linear model for transformed version of the variable. See for example, Chen and Chen (1996) and Karlberg (2000) and references therein.

Deville and Sarndal (1992) introduced calibration approach for estimation of population parameters. This approach is based on the implicit assumption that study and auxiliary variable is linearly related. When survey data is skewed linearity assumption does not hold good. Wu and Sitter (2001) proposed model calibration approach for estimation of population parameters. This is a general approach of calibration covering both linear and non-linear models. The key idea of model calibration approach is as follows. The relationship between survey variable Y and auxiliary variable X can be either linear or nonlinear but survey variable Y and fitted value of Y (estimated from expected value of Y) is approximately linear. However, in Wu and Sitter (2001) the fitted values used in model calibration approach are biased due to back transformation to the original scale. Under model based framework, Basak

Corresponding author: Hukum Chandra
E-mail address : hchandra@iasri.res.in

et al. (2014) applied the back transformation bias correction in fitted values and considered the model based model calibration approach for estimation of population total. Their empirical results revealed that bias correction leads reduction in bias. Karlberg (2000) described prediction of finite population total under a log normal model. She fitted the log linear model for skewed data and predicted the non-sample part of the population. She also considered the back transformation bias correction in the prediction of nonsample values of survey variable. Several methods have been proposed for dealing with skewed data and they are based on either implicit or explicit underlying model assumptions. Further, sometimes survey variable is highly skewed in nature but there is no auxiliary information to build a good working model. Indeed, we can fit a mean model on log scale for such skewed variable. This article explores these data situation and describes the estimation of population total for skewed data. Throughout in this article we adopt model-based approach of survey estimation. So, all moments are evaluated with respect to a model for the population data.

In the next Section we consider a linear model and briefly introduce the calibration approach to estimation of population quantities from a model-based perspective. In Section 3 we summarize the model based model calibration approach for the estimation of population total and its mean square error estimation. Section 4 illustrates empirical results to compare different estimators of population total. Finally, Section 5 presents concluding remarks.

2. CALIBRATION ESTIMATION UNDER A LINEAR MODEL

To start, we fix our notation. Let U denote a finite population of size N and s denote a sample of size n drawn from this population. Let $\mathbf{y}_U$ denote the N-vector of population values of a characteristic Y of interest and $\mathbf{X}$ denote the p-vector of auxiliary variables that are related, in some sense, to Y. Thus, $\mathbf{x}_U = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N)^T$ denote the corresponding $N \times p$ matrix of population values of auxiliary variables. We further assume that unit level auxiliary information is available for the entire population. Suppose that our primary aim is estimation of population total $t_{Uy} = \sum_U y_i$. The simple estimator of population total t_{Uy}, which does not make use of auxiliary data is given by

$$\hat{t}_y = N\left(\frac{1}{n}\sum_{i\in s} y_i\right) = N\bar{y}_s, \text{ with } \bar{y}_s = \frac{1}{n}\sum_{i\in s} y_i. \quad (1)$$

Deville and Särndal (1992) define an **X**-calibrated linear estimator of t_{Uy} as $\hat{t}_y = \sum_{i\in s} w_i y_i$, where the calibrated weights $\{w_i;\ i \in s\}$ satisfy $\sum_{j\in s} w_i \mathbf{x}_i = \mathbf{t}_{Ux}$. Here $\boldsymbol{t}_{Ux}$ is the vector of population totals of **X**. This approach is based on an implicit assumption that the population values of Y and **X** are linearly related, in which case the calibration constraint is equivalent to ensuring that the estimator $\hat{t}_y$ is an unbiased predictor of t_{Uy} under a linear model for the regression of Y on **X** in the population. A model-based perspective of calibration approach is described as follows. Let us consider that the relationship between Y and **X** in the population can be described by a linear regression model of the form

$$E(\mathbf{y}_U|\mathbf{x}_U) = \mathbf{x}_U\beta \text{ and } Var\ (\mathbf{y}_U|\mathbf{x}_U) = \mathbf{V}_U \quad (2)$$

where β is a $p \times 1$ vector of unknown parameters and $\mathbf{V}_U$ is a positive definite covariance matrix and is known up to a multiplicative constant. Given a sample s of size n from this population, we write $\mathbf{y}_U$ as $\mathbf{y}_U^T = (\mathbf{y}_s^T, \mathbf{y}_r^T)$, where $\mathbf{y}_s^T$ corresponds to n sample units and $\mathbf{y}_r^T$ corresponds to $N - n$ non-sample units. Here $r = U - s$ denotes the population units that are not in sample. Similarly, we can partition $\mathbf{x}_U = \begin{bmatrix}\mathbf{x}_s \\ \mathbf{x}_r\end{bmatrix}$ and $\mathbf{V}_U = \begin{bmatrix}\mathbf{V}_{ss} & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_{rr}\end{bmatrix}$ into their sample and non-sample components. In practice, the variance component parameters in model (2) are unknown and estimated from sample data. Using the estimated value of variance components, vector of weights that defines the Empirical Best Linear Unbiased Predictor (EBLUP) of t_{Uy} is (Royall 1976, Valliant *et al.* 2000, Section 2.4)

$$\mathbf{w}_s^{EBLUP} = (w_i^{EBLUP}; i \in s)$$

$$= \mathbf{1}_s + \hat{\mathbf{H}}_s^T(\mathbf{t}_{Ux} - \mathbf{t}_{sx}) + (\mathbf{I}_s - \hat{\mathbf{H}}_s^T\mathbf{x}_s^T)\hat{\mathbf{V}}_{ss}^{-1}\hat{\mathbf{V}}_{sr}\mathbf{1}_r \quad (3)$$

where $\hat{\mathbf{H}}_s = (\mathbf{x}_s^T\hat{\mathbf{V}}_{ss}^{-1}\mathbf{x}_s)^{-1}\mathbf{x}_s^T\hat{\mathbf{V}}_{ss}^{-1}, \mathbf{1}_s\ (\mathbf{1}_r)$ denotes a vector of 1's of size $n(N-n)$, $\mathbf{t}_{sx}$ is the vector of sample totals of **X** and $\mathbf{I}_s$ is the identity matrix of order n. The EBLUP of population total of Y is then defined as

$$\hat{\mathbf{t}}_y^{EBLUP} = \sum_{i\in s} w_i^{EBLUP} y_i. \quad (4)$$

Note that the EBLUP weights (3) are calibrated to X and the EBLUP (4) is same as the calibration estimator $\hat{t}_y = \sum_{i\in s} w_i y_i$ of Deville and Särndal (1992).

3. MODEL BASED MODEL CALIBRATION WEIGHTING FOR POPULATION ESTIMATION

If the underlying population model is non-linear, the calibration estimator can be model-biased, and hence inefficient. Model calibration was introduced by Wu and Sitter (2001) as a model-assisted method of calibrated weighting when the underlying regression relationship is non-linear. A model-based perspective of Wu and Sitter's model calibration approach can be stated as follows. Suppose the relationship between Y and $\mathbf{X}$ in the population can be described as

$$E(\mathbf{y}_U|\mathbf{x}_U) = h(\mathbf{x}_U;\eta) \text{ and}$$

$$Var(\mathbf{y}_U|\mathbf{x}_U) = \Sigma = diag(\sigma_i^2; i = 1,...,N) \quad (5)$$

where $\eta = (\eta_0, ..., \eta_p)^T$ and σ_i^2 are unknown model parameters. Here $h(\mathbf{x}_U;\eta)$ denotes the N-vector of mean function, $h(\mathbf{x}_i;\eta)$, which is a known function of $\mathbf{x}_i$ and η. Further, it is also assumed that population units are mutually uncorrelated. Let $\hat{\eta}$ denote a 'model-efficient' estimator of η with associated fitted values $h(\mathbf{x}_i;\hat{\eta})$. Usually, there is a linear relationship between the actual values y_i of Y and their corresponding fitted values $\hat{y}_i = h(\mathbf{x}_i;\hat{\eta})$. Thus, we replace (2) by a linear model of the form

$$E(y_i \mid \hat{y}_i) = \alpha_0 + \alpha_1 \hat{y}_i \text{ and } Cov(y_i, y_j \mid \hat{y}_i, \hat{y}_j) = \omega_{ij}. \quad (6)$$

The model (6) is referred to the 'fitted value' linear model defined by (2). Let $\mathbf{J}_U$ denote the population 'design matrix' defined by (6), *i.e.*, $\mathbf{J}_U = (\mathbf{1}_U \ \hat{\mathbf{y}}_U)$. Following sample and non-sample partitioning of $\mathbf{J}_U$ and $\Omega_U = [\omega_{ij}]$ as above, and using the estimated variance components of Ω_U, EBLUP weights for population total of Y under the general linear 'fitted value' model (6) are given by

$$\mathbf{w}^{MC\text{-}EBLUP} = (w_i^{MC\text{-}EBLUP})$$

$$= \mathbf{1}_s + \hat{\mathbf{H}}_{MC}^T(\mathbf{J}_U^T\mathbf{1}_U - \mathbf{J}_s^T\mathbf{1}_s) + (\mathbf{I}_s - \mathbf{H}_{MC}^T\mathbf{J}_s^T)\hat{\Omega}_{ss}^{-1}\hat{\Omega}_{sr}\mathbf{1}_r \quad (7)$$

where $\hat{\mathbf{H}}_{MC} = (\mathbf{J}_s^T\hat{\Omega}_{ss}^{-1}\mathbf{J}_s)^{-1}\mathbf{J}_s^T\hat{\Omega}_{ss}^{-1}$ and $\mathbf{1}_U$ denotes vector of 1's of size N and $\mathbf{I}_U$ and $\mathbf{I}_r$ denote identity matrices of order N and $N - n$ respectively. These weights are model-calibrated under (6) since $\mathbf{J}_s^T\mathbf{w}^{MC\text{-}EBLUP} = \mathbf{J}_U^T\mathbf{1}_U$. The model-based model calibration estimator of population total of Y is given by

$$\hat{t}_y^{MC\text{-}EBLUP} = \sum_{i\in s} w_i^{MC\text{-}EBLUP} y_i. \quad (8)$$

3.1 A Log Transformation Model

We now consider a special case of log transformed model. Let us assume that the relationship between survey variable Y and auxiliary variable $\mathbf{X}$ is not linear in the raw scale but it is linear in the log scale. That is, log(Y) and log($\mathbf{X}$) (or sometimes $\mathbf{X}$) is linear. Then the log scale model is

$$l_i = \log(y_i) = \mathbf{z}_i^T\boldsymbol{\beta} + \varepsilon_i; i = 1,...,N, \quad (9)$$

where $\mathbf{z}_i^T = (1, \log(x_{i1}), ..., \log(x_{ip}))$, $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$ and. $\varepsilon_i \sim N(0, \sigma^2)$. Under model (9), the predicted values of Y are

$$\hat{y}_i = \hat{E}(y_i \mid \mathbf{x}_i) = \exp(\mathbf{z}_i^T\hat{\beta} + \hat{\sigma}^2/2) = h(\mathbf{x}_i;\hat{\eta}). \quad (10)$$

We can specify the fitted value model like (6) and then obtain the EBLUP weights (7) and the model calibration estimator (8). Here, we see that

$$E(\hat{y}_i \mid \mathbf{x}_i) = E\{\exp(\mathbf{z}_i^T\hat{\beta} + \hat{\sigma}^2/2)\} \neq E(y_i \mid \mathbf{x}_i)$$
$$= \exp(\mathbf{z}_i^T\beta + \sigma^2/2).$$

That is, the predicted values $\hat{y}_i = \exp(\mathbf{z}_i^T\hat{\beta} + \hat{\sigma}^2/2) = h(\mathbf{x}_i;\hat{\eta})$ are biased. This bias arises due to back transformation. As a consequence, the model calibration estimator (8) is also model biased. A Taylor series approximation has been used to correct this bias. See Basak *et al.* (2014) for details. Then the bias corrected fitted values are defined as

$$\hat{y}_i^{BC} = \hat{k}_i^{-1}\hat{y}_i = h^{BC}(\mathbf{x}_i;\hat{\eta}), \quad (11)$$

where $\hat{k}_i = 1 + \frac{1}{2}\left(\mathbf{z}_i^T\hat{V}(\hat{\boldsymbol{\beta}})\mathbf{z}_i + \frac{1}{4}\hat{V}(\hat{\sigma}^2)\right) = 1 + \frac{s^2 a_{ii}}{2} + \frac{s^4}{4n}$,

with $a_{ii} = \mathbf{z}_i^T(\mathbf{Z}_s^T\mathbf{Z}_s)^{-1}\mathbf{z}_i$. Here superscript of "BC" denotes "bias corrected" version. Using bias corrected version of "fitted values" (11) we define the fitted value model (6) and the EBLUP weights (7), referred as bias corrected version of model calibration EBLUP weight (BCMC-EBLUP), denote them by $\mathbf{w}^{BCMC\text{-}EBLUP} = (w_i^{BCMC\text{-}EBLUP})$. We then obtain the bias corrected version model-based model calibration estimator of population total as

$$\hat{t}_y^{BCMC-EBLUP} = \sum_{i\in s} w_i^{BCMC-EBLUP} y_i. \qquad (12)$$

Under the log normal model (9), using prediction approach Karlberg (2000) also described the predictor of population total given by

$$\hat{t}_y^{KB} = \sum_{i\in s} y_i + \sum_{i\in r}\left\{\hat{l}_i^{-1}\exp(\mathbf{z}_i^T\hat{\boldsymbol{\beta}} + \frac{s^2}{2})\right\}. \qquad (13)$$

Here $\hat{l}_i = \exp(\frac{s^2 a_{ii}}{2} + \frac{s^4}{4n}) \approx \hat{k}_i$.

3.2 Mean Square Error Estimation

The mean square error estimation of prediction based estimator of population total $\hat{t}_y^{KB}$ can be followed from Karlberg (2000). It is noteworthy that unlike the model calibration estimators, the prediction approach based Karlberg estimator (13) is non-linear in nature thereby causes difficulty in mean square error estimation. The model calibration estimator has the advantage of mean square error estimation since it has a weighted linear form, see Basak *et al.* (2014). For example, suppose that the estimator of population mean $m_U = N^{-1}\sum_U Y_i$ of variable Y is $\hat{m}_U = \sum_s w_i^* y_i$, expressed as a weighted linear estimator such that $w_i^* = O(n^{-1})$ and $\sum_s w_i^* = 1$. An estimator of the mean square error (MSE) of population total of Y is then

$$mse(\hat{t}_y) = N^2[v(\hat{m}_U) + \{b(\hat{m}_U)\}^2]$$

$$= \sum_{i\in s}\left\{1 + \sum_{k\in s}\phi_{ki}^2 - 2\phi_{ii}\right\}^{-1}\{(a_i^2 + n^{-1}(N-n)\}(y_i - \hat{h}_i)^2$$

$$+\left\{\sum_s w_i\hat{h}_i - \sum_U \hat{h}_i\right\}^2, \qquad (14)$$

where $a_i = Nw_i^* - 1 = w_i - 1$ and $w_i; i\in s$ are weights for prediction of population total of Y. Here $\hat{h}_i = h(\mathbf{x}_i;\hat{\boldsymbol{\eta}})$ is an estimate of $E(y_i|\mathbf{x}_i)$. In particular, $h(\mathbf{x}_i;\hat{\boldsymbol{\eta}}) = \sum_{k\in s}\phi_{ki}y_k$, where $\phi_{ki} = O(n^{-1})$ and $\sum_s \phi_{ki} = 1$. See Basak *et al.* (2014) for detail theoretical development of MSE estimate (14) for various calibration weighting based estimators of population total of Y.

3.3 Estimation of Population Total with No Auxiliary Information

In many situations, although survey variable is skewed but there is no availability of auxiliary information. In such cases, $E(\mathbf{y}_U|\mathbf{x}_U) = \mu\mathbf{1}_U$ and $Var(\mathbf{y}_U|\mathbf{x}_U) = \mathbf{V}_U$. That is, underlying becomes a mean model. Then the EBLUP weights are $\mathbf{w}_s^{EBLUP} = \mathbf{1}_s + \frac{1}{n}\mathbf{1}_s(N-n) = \frac{N}{n}\mathbf{1}_s$ and the EBLUP of population total of Y is, $\hat{t}_y^{EBLUP} = N\bar{y}_s$, same as the estimator (1). Further, the fitted value model (6) leads to

$$E(y_i \mid \hat{y}_i) = \mathrm{e}\exp(\mu + \frac{\hat{\sigma}^2}{2}) = \lambda \text{ and}$$

$$Cov(y_i, y_j \mid \hat{y}_i, \hat{y}_j) = \omega_{ij} = \exp(2\mu + \sigma^2)\{\exp(\sigma^2) - 1\} = \delta,$$

where λ and δ are constant independent of sample units. In this case also the EBLUP weights are $\mathbf{w}^{MC\text{-}EBLUP} = \frac{N}{n}\mathbf{1}_s$. As a results, the model-based model calibration estimator of population total is $\hat{t}_y^{MC-EBLUP} = N\bar{y}_s$. We further see that bias corrected version of fitted values are $\hat{y}_i = \hat{k}_i^{-1}\exp(\mu + \frac{\hat{\sigma}^2}{2}) = \lambda^*$ with $\hat{k}_i = 1 + \frac{s^2}{2n} + \frac{s^4}{4n}$, for all i. So the bias corrected version of model based model calibrated estimator also reduces to $\hat{t}_y^{BCMC-EBLUP} = N\bar{y}_s$. In the absence of auxiliary information calibration estimators (both linear and model calibration versions of estimators) reduce to simple sample mean based estimator of population total. Indeed, calibration based approaches are applicable if auxiliary information are available in the survey.

We see that Karlberg's prediction based estimator for population total (13) reduces to

$$\hat{t}_y^{KB} = \sum_{i\in s} y_i + \sum_{i\in r}\left\{\hat{l}_i^{-1}\exp(\hat{\mu} + \frac{s^2}{2})\right\}$$

$$= n\bar{y}_s + \sum_{i\in r}\left\{\hat{l}_i^{-1}\exp(\hat{\mu} + \frac{s^2}{2})\right\}, \qquad (15)$$

where $\hat{l}_i = \exp\left(\frac{s^2}{2n} + \frac{s^4}{4n}\right)$. This estimator is not the same as the simple sample mean based estimator (1).

It is evident that in case no auxiliary information is available calibration based estimators reduce to $\hat{t}_y = N\bar{y}_s$ and this estimator is based on implicit assumption of normality. However, the estimator (15) seems suitable for skewed survey variable.

4. EMPIRICAL EVALUATIONS

In this Section we report the results from model-based and design-based simulation studies that illustrate the performance of the different estimators of the population total defined in the preceding Sections. In particular, we consider following five estimators of the population total:

(i) Simple sample mean based estimator $\hat{t}_y$ given by (1) (denoted as SRS),

(ii) Under linear model (2), calibration estimator (or EBLUP) $\hat{t}_y^{EBLUP}$ (denoted as LC),

(iii) Under transform model (9), model calibration estimator (or MC-EBLUP) $\hat{t}_y^{MC-EBLUP}$ given in (8) (denoted as MC),

(iv) Under transform model (9), model calibration estimator with bias correction (or BCMC-EBLUP) $\hat{t}_y^{BCMC-EBLUP}$ given in (12) (denoted as BCMC), and

(v) Under transform model (9), prediction approach based Karlberg estimator $\hat{t}_y^{KB}$ given in (13) (denoted by KB).

The performance of the various estimators was measured by the simulated relative bias (RB, in percentage) and relative root mean square error (RRMSE, in percentage), defined by

$$RB(\hat{T}) = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{\hat{T}_i - T_i}{T_i}\right)\times 100$$

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

where T_i denotes the actual value of population total at simulation run i, with predicted value $\hat{T}_i$ and M denotes the number of simulation run. In the case of design based simulations $T_i = T$ since population is fixed.

4.1 Model-based Simulation Study

In the model-based simulations, a finite population of size $N = 2000$ units was generated from a model $\log(y) = 1 + x + \varepsilon$ where $x \sim Gamma(1, 1)$ and $\varepsilon \sim N(0, \sigma^2)$. From this population, a sample of size $n = 50, 100, 150, 200$ was taken by simple random sampling without replacement. Then various estimators were computed using the sample data and all the fitted values. The Monte Carlo simulation was run $M = 5000$

Table 1. Percentage relative bias (RB, %) and percentage relative root mean square error (RRMSE, %) of different estimators in model based simulation Set A.

Rho	*n*	SRS	LC	MC	BCMC	KB
RB, %						
0.80	200	1.754	–5.690	3.776	0.850	2.231
	150	–2.143	–9.639	4.579	0.421	2.357
	100	2.842	–10.046	6.451	0.135	2.741
	50	8.350	–16.093	11.071	–2.243	2.530
0.60	200	1.898	–5.417	10.857	2.645	6.073
	150	–1.741	–8.920	13.407	1.889	6.439
	100	3.027	–9.470	18.586	1.394	6.857
	50	8.378	–15.071	36.159	–1.330	5.491
0.40	200	1.999	–4.950	27.670	7.372	14.026
	150	–1.202	–7.891	35.301	6.694	14.579
	100	3.322	–8.302	50.986	7.421	14.539
	50	8.105	–13.544	131.295	10.952	9.556
RRMSE, %						
0.80	200	99.737	87.905	26.122	23.355	24.438
	150	110.569	100.327	28.768	24.880	26.294
	100	148.056	123.307	34.729	28.915	30.624
	50	256.892	158.905	55.744	37.789	41.352
0.60	200	101.874	89.865	59.244	44.021	49.564
	150	112.909	102.102	65.802	45.100	50.848
	100	150.929	127.846	83.570	53.026	58.808
	50	257.399	160.144	185.747	74.816	78.052
0.40	200	105.153	93.320	160.098	86.871	107.647
	150	116.491	104.784	194.324	89.197	102.028
	100	156.600	136.651	283.627	115.966	118.905
	50	257.245	162.319	1040.879	216.700	140.983

times. Simulations based on this model are referred to as Set A simulations. Here, we considered three different values of σ^2 such that the correlation coefficient between log(y) and x are 0.80, 0.60 and 0.40 respectively. This leads three different finite populations. Simulation studies were carried out in R software. The values of percentage relative bias and the values of percentage relative root mean square error of different estimators for three different values of correlation coefficients are reported in Table 1.

These results in Table 1 show that the value of percentage relative bias is highest for linear calibration based estimator (LR) when the correlation coefficient between transformed variables is relatively high, *i.e*., Rho = 0.80 while at the moderate values of correlation coefficient, *i.e*. Rho = 0.60 and 0.40 the value of percentage relative bias is high for the model calibration based estimator without bias correction (MC). However, the value of percentage relative bias is smaller for the BCMC predictor for all values of correlation coefficient and sample size. Further, the relative bias of the BCMC predictor increases as the correlation coefficient between transformed variables decreases for all sample sizes. One significant result is that the relative bias of the BCMC predictor decreases as sample size decreases, in contrast to the other estimators, for all the given values of correlation coefficient indicating that the BCMC predictor gives good result even for small sample where problem of bias is more. For all the given values of correlation coefficient, relative bias becomes negative for sample size n = 50 due to over bias correction. This is obvious since bias correction is based on large sample approximation.

In the case of percentage relative root mean square error, it is highest for simple sample mean (SRS) followed by linear calibration based estimator (LR) and least for the BCMC estimator for all values of correlation coefficient and sample size. It was also observed that as the correlation coefficient between transformed variables decreases, the relative root mean square error of the BCMC predictor increases for all sample sizes. Similarly, relative root mean square error also increases as sample size decreases for all the given values of correlation coefficient. However, in all the cases the gain in relative bias and relative root mean square error of model calibration with bias correction as compared to usual model calibration is substantial.

In Section 3.2 we noticed that the calibration approach based estimators (*i.e*., LC, MC and BCMC) reduces to the SRS when there is no auxiliary information. However, in this case the KB estimator is still different from the SRS. Consequently, if variable of interest Y for survey estimation is skewed and no auxiliary information available to describe the relationship with Y, it is interesting to examine the performance of Karlberg estimator KB. We now describe a model based simulations to examine the performance of two estimators, SRS and KB when there is no auxiliary information available. These simulations are referred as Set B of model based simulations. We referred here, a finite population consisting of N = 2000 units was generated from a model $\log(y) = \mu + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$ and $\mu = 1$. Three different finite populations were used by choosing different values of. σ^2. From each population, a sample of size n = 50, 100, 150, 200 was taken by simple random sampling without replacement. The Monte Carlo simulation was run M = 5000 times. The values of percentage relative bias and percentage relative root mean square error of two estimators for three different populations are reported in Table 2. The results in Table 2 clearly show that relative biases and relative root mean square errors of the KB estimator are smaller than the SRS. In case of skewed data, even if there is no auxiliary information available then also KB estimator can be used and it gives better performance as compared to simple sample mean (SRS).

Table 2. Percentage relative bias (RB, %) and percentage relative root mean square error (RRMSE, %) of two estimators in model based simulations Set B.

n	σ = 0.50		σ = 0.65		σ = 0.80	
	SRS	KB	SRS	KB	SRS	KB
RB, %						
200	0.014	0.005	0.067	0.049	0.153	0.124
150	–0.135	–0.115	–0.218	–0.170	–0.350	–0.243
100	–0.105	–0.068	–0.187	–0.107	–0.320	–0.163
50	–0.073	–0.054	–0.121	–0.078	–0.203	–0.114
RRMSE, %						
200	3.692	3.674	5.036	4.975	6.807	6.623
150	4.126	4.119	5.598	5.560	7.523	7.380
100	5.240	5.228	7.108	7.053	9.549	9.351
50	7.506	7.503	10.153	10.132	13.587	13.430

4.2 Design-based Simulation Study

The design-based simulations are based on real survey data set. The survey data set we used the same data that was reported in Chandra and Chambers (2011). That is, a sample of 1652 farms that participated in the Australian Agricultural and Grazing Industries Survey (AAGIS) conducted by the Australian Bureau of Agricultural and Resource Economics. We considered this original sample data as a target population of 1652 farms. From this fixed population, we draw samples of different sizes n = 50, 100, 150, 200 by simple random sampling without replacement sampling scheme. In particular, for each sample sizes we draw M = 5000 samples. Here, the aim was to estimate total annual farm costs (TCC, measured in A$) using farm size (hectares) as the auxiliary variable.

We also did an exploratory data analysis with original sample (or our target population) of 1652 farms to examine the behavior of variable of interest (TCC) and auxiliary variable (farm size or FS). The histogram in Fig. 1 shows that the survey variable TCC is skewed. However, the histogram is reasonably normal when the variable TCC was plotted on log scale. This is also true for the auxiliary variable (farm size or FS) plotted in

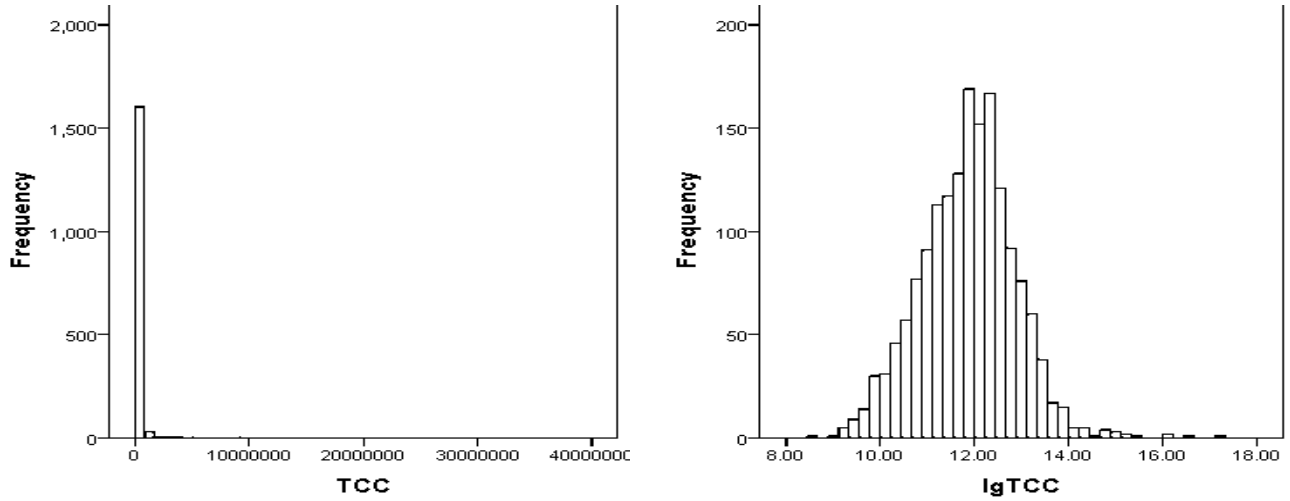

Fig. 1. Histogram of total annual farm costs (TCC) (left) and log (TCC) (right)

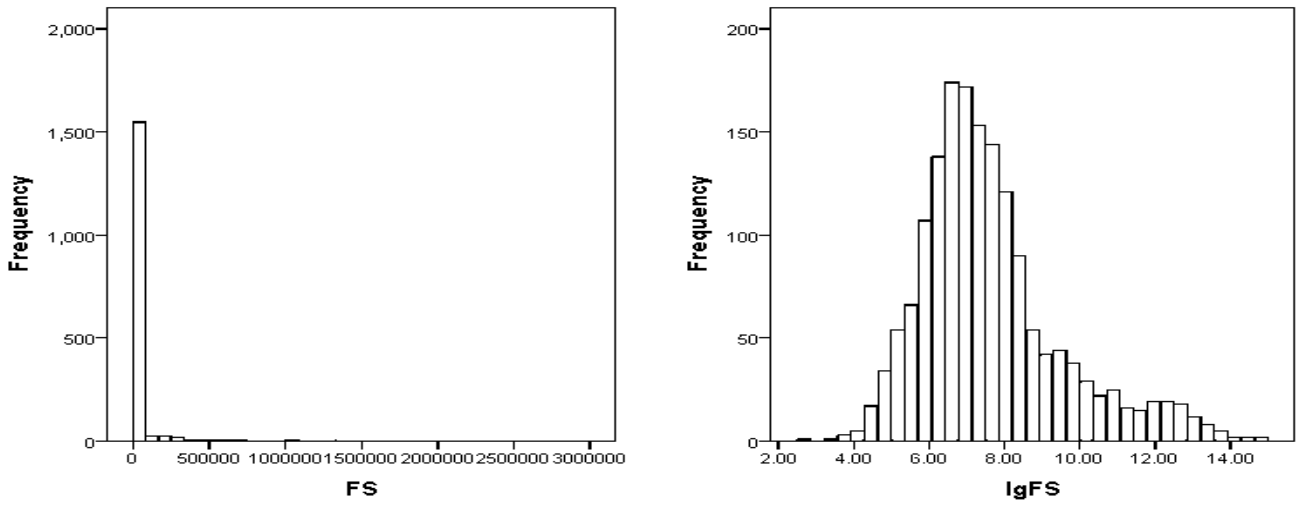

Fig. 2. Histogram of Farm size (left) and log (Farm size) (right)

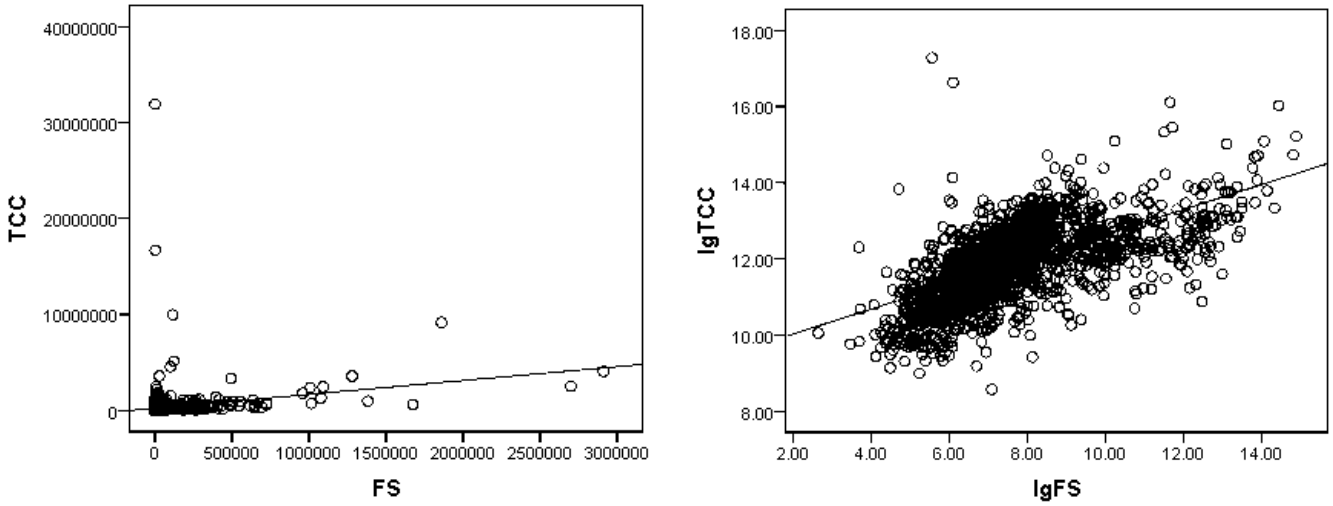

Fig. 3. Scatterplot of TCC and Farm size on raw scale (left) and log scale (right)

Fig. 2. We further examined the linear relationship between TCC and FS. Fig. 3 shows that the linear relationship between TCC and FS is very weak on raw scale. However, this relationship improved on log scale. This analysis clearly reveals that the data is skewed and hence the model calibration approach can be explored. It is noteworthy that the linear relationship on log transform scale is reasonable but not very strong and hence we expect an indicative result of the model calibration estimator with back transformation bias correction.

The results were generated for five different estimator using $M = 5000$ samples each of sizes $n = 50, 100, 150, 200$ are reported in Table 3. Design-based simulations serve to complement model-based simulations, providing evidence of comparative performance and robustness in realistic data scenarios. Table 3 shows the results for the design-based simulations using the AAGIS data.

Table 3. Percentage relative bias (RB, %) and percentage relative root mean square error (RRMSE, %) of different estimators in AAGIS population.

n	SRS	LC	MC	BCMC	KB
RB, %					
200	–0.177	–0.330	2.040	1.922	–6.876
150	–0.121	–0.184	2.207	2.041	–7.079
100	–0.232	–0.086	2.860	2.597	–7.106
50	0.118	3.739	4.604	3.958	–6.606
RRMSE, %					
200	24.444	23.790	16.597	16.555	11.966
150	28.213	27.369	19.401	19.334	13.027
100	34.998	34.805	25.778	25.589	15.327
50	52.827	61.765	44.525	43.214	21.408

These results in Table 3 showed that relative bias of all the estimators are within a reasonable range of 5 per cent except for KB. Relative bias is highest for KB estimator for all the sample sizes whereas it is least for SRS for almost all sample sizes. However, relative root mean square error is least for KB estimator in all the cases. Simple sample mean based SRS and linear calibration estimator (LC) shows higher relative root mean square error as compared to other estimators of population total. It was also observed that relative root mean square error decreases as the sample size increases for all the estimators. Overall, the bias corrected model calibration estimator shows a satisfactory performance as compared to other estimators of population total.

5. CONCLUSIONS

This paper discusses different estimators of population total for skewed data. The bias corrected model calibration estimator uses the bias corrected fitted value for calibration which is particularly effective when the population is skewed. In the absence of auxiliary information, linear calibration estimator, model calibration estimator and bias corrected model calibration estimator reduces to simple sample mean

and in this case Karlberg estimator is an alternative which shows improved performance as compared to simple sample mean. So, it is better to use Karlberg estimator for skewed data when auxiliary information is not available. But if auxiliary information is available it is better to use bias corrected model calibration estimator for estimation of population total of skewed data. However, before using this estimator it is recommended that users should check the working model for their data. Further, this paper considered a special case of log transformation for skewed data and estimators are applicable for skewed variable taking strictly positive values only. However, skewed data often take zero values of observation, referred as the semicontinuous variable so model calibration based approach needs to be extended for this case. Authors are currently working in estimation of population total for semicontinuous data.

ACKNOWLEDGEMENTS

Valuable comments and suggestions provided by a referee are gratefully acknowledged. The first author also gratefully acknowledges the Junior Research Fellowship provided by Indian Council of Agricultural Research, New Delhi.

REFERENCES

Basak, P., Chandra, H., Sud, U.C. and Lal, S.B. (2014). Prediction of population total for skewed variable under a log transform model. *Inter. J. Agric. Statist. Sci.*, **9(2)**, 143-153.

Chandra, H. and Chambers, R.L. (2011). Small area estimation under transformation to linearity. *Survey Methodology*, **37**, 39-51.

Chen, G. and Chen, J. (1996). A transformation method for finite population sampling calibrated with empirical likelihood. *Survey Methodology*, **22**, 139-146.

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

Karlberg, F. (2000). Population total prediction under a lognormal superpopulation model. *Metron*, **LVIII**, 53-80.

Royall, R.M. (1970). On finite population sampling under certain linear regression models. *Biometrika*, **57**, 377-387.

Royall, R.M. (1976). The linear least-squares prediction approach to two-stage sampling. *J. Amer. Statist. Assoc.,* **71**, 657-664.

Valliant, R., Dorfman, A. H. and Royall, R. M. (2000). *Finite Population Sampling and Inference*. Wiley, New York.

Wu, C. and Sitter, R.R. (2001). A model calibration approach to using complete auxiliary information from survey data. *J. Amer. Statist. Assoc.,* **96**, 185-193.