Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 70(1) 2016 19-24

Calibration Based Regression Type Estimator of the Population Total under Two Stage Sampling Design

Kaustav Aditya, U.C. Sud, Hukum Chandra and Ankur Biswas
ICAR-Indian Agricultural Statistics Research Institute, New Delhi

Received 25 August 2015; Revised 21 January 2016; Accepted 03 March 2016

SUMMARY

Regression type estimators of the population total were developed using the calibration approach under the assumption that the population level auxiliary information is available at primary stage unit level under two stage sampling design. The variance and the estimator of the variance of the proposed estimators were also developed. Theoretical results obtained are demonstrated through simulation studies. Empirical results show that the proposed estimators outperforms the usual regression estimators under two stage sampling design in terms of the criteria of relative bias and relative root mean square error.

Keywords: Auxiliary information, Calibration approach, Regression type estimator, Primary stage unit, Two stage sampling.

1. INTRODUCTION

Estimation in sample surveys is conducted mainly by attaching weights to sample data and then computing weighted averages. Sometimes, auxiliary information may be employed to improve survey estimates. In this context, a set of sample weights is said to have the calibration property if it reproduces exactly known population quantities when applied to the sample values of the corresponding auxiliary variables. It is based on the argument that "weights that perform well for the auxiliary variables also should perform well for the study variable" (Deville and Särndal, 1992). This auxiliary information is often used by survey statisticians to increase the precision of estimators of commonly used population parameters i.e. population mean or population total. The most common form of estimators which use auxiliary information, are ratio and regression estimator. In fact, the regression estimator (GREG) is a special case of the calibration estimator when the chosen distance function is the Chi-square distance (Deville and Särndal, 1992). The main difference between the GREG approach and the calibration approach is in GREG approach the predicted values are generated using an assisting model whereas in calibration approach it does not depend on any assumption about the assisting model. Assisting model is an imagined relationship between study variable and auxiliary variable and can have many forms: linear, nonlinear, generalized linear, mixed (model with some fixed, some random effects), and many more. In the past twenty years or so, calibration itself became an important topic in survey research and a large amount of literature has been devoted to it, so much so that it gained significant attention not only in the field of survey methodology, but also of survey practice.

Corresponding author: Kaustav Aditya
E-mail: kaustav.aditya@.icar.gov.in

Calibrated weights, mainly derived using the techniques in Deville and Särndal (1992), are currently employed by several national statistical agencies to produce official estimates from real life surveys. Following Deville and Särndal (1992) a lot of work has been carried out in the context of calibration estimation i.e. Singh *et al.* (1998, 1999), Folsom and Singh (2000), Farrell and Singh (2002), Wu and Sitter (2001), Sitter and Wu (2002), Kott (2006), Estevao and Särndal (2002, 2006), Sud *et al.* (2014) but most of the studies in this context is only restricted to single stage or two phase sampling designs whereas in large scale surveys two stage or multistage sampling designs are generally used. Hence, there is a need to develop methodologies for calibration estimators for multistage sampling designs. In this paper, we have proposed calibration based regression type estimators under two stage sampling design by modifying the sampling design weight with the help of auxiliary information. There are several cases of availability of auxiliary information for two-stage sampling design, depending on whether the information is available at the cluster level or element level. We consider the case of availability of population level auxiliary information at the cluster level.

In what follows, a regression type estimator has been proposed using the calibration approach under two stage sampling design in the presence of complex auxiliary information and the regression line does not pass through the origin. Proposed calibration approach based estimators along with expressions for variance and variance estimator have been developed in section 2. The improved performance of the proposed estimator over the usual estimator under two stage sampling design is demonstrated through a simulation study in section 3. In section 4 the concluding remarks were made.

2. THE PROPOSED ESTIMATOR

We consider a simple case where information on only one auxiliary variable is available. Let, the population of elements $U=\{1,\ldots,k,\ldots,N_I\}$ is partitioned into clusters, $U_1, U_2,\ldots, U_i,\ldots, U_{N_I}$. They are also called the primary stage units (psus) when there are two stages of selection. The size of U_i is denoted as N_i. We have

$$U=\bigcup_{i=1}^{N_I} U_i \text{ and } N=\sum_{i=1}^{N_I} N_i .$$

At stage one, a sample of psus, s_I, is selected from U_I according to the design $p_I(.)$ with the inclusion probabilities π_{Ii} and π_{Iij} at the psu level. The size of s_I is n_I psus. The sampling units at the second stage (ssu) are population elements, labeled $k = 1,\ldots, N$. Given that the psu U_i selected at the first stage a sample s_i of size n_i units is drawn from U_i according to some specified design $p_i(.)$ with inclusion probabilities $\pi_{k/i}$ and $\pi_{kl/i}$. For the second stage sampling we are assuming the invariance and independence property. The whole sample of elements and its size is defined as,

$$s=\bigcup_{i=1}^{S_I} s_i \text{ and } n_s=\sum_{i=1}^{n_I} n_i .$$

The inclusion probabilities at the first stage is given as,

$$\pi_{Ii}=\Pr(i\in s_I),$$

$$\pi_{Iij}=\begin{cases}\Pr(i\&j\in s_I), i \text{ and } j \text{ belongs to different psus}\\ \pi_{Ii}, i \text{ and } j \text{ belongs to same psus}\end{cases}$$

The inclusion probabilities for the second stage is given as,

$$\pi_{k/i}=\Pr(k\in s_i \mid i\in s_I) \text{ and}$$

$$\pi_{kl/i}=\begin{cases}\Pr(k\&l\in s_i \mid i\in s_I), k \text{ and } l \text{ are different}\\ \pi_{k/i}, k \text{ and } l \text{ are same}\end{cases}$$

Let the study variable be y_k which is observed for $k\in s$. The parameter to estimate is the population total $t_y=\sum_{i=1}^{N} y_k=\sum_{i=1}^{N_I} t_{yi}$ where

$t_{yi}=\sum_{i=1}^{N_i} y_k$ =i-th psu total.

Let, the population level auxiliary information (z_i) is available at the psu level i.e. for national surveys for certain establishments say hospitals, with numerous but fairly large clusters, say at tehsil or sub-districts level. Let y_k be the study variable value for *k*-th hospital. Due to administrative status of the tehsils or sub district, much information is ordinarily available to create useful tehsils or sub district level auxiliary values z_i. These auxiliary values are usually obtained from demographic sources, a census, or a current population survey and can be used for explaining cluster totals. Likewise, the proposed estimator can be used in any practical situation like Household surveys involved in estimation of total population, men in the work force, women in the work force, or the number of children five years old or younger or average income of men/women in the work force or proportion of households with total income below the poverty level or yield estimates for geo-political regions or for other cross-sectional domains with availability of suitable auxiliary information from last census or other demographic sources.

Let the population level auxiliary information (z_i) is available at the psu level and the value of z_i is observed for all the sampled clusters and a correct value of $\sum_{i=1}^{N_I} z_i$ is available, the simple Horvitz-Thompson (1952) estimator under two stage sampling design is given as,

$$\hat{t}_{HT} = \sum_{i=1}^{n_I} \frac{\hat{t}_{yi\pi}}{\pi_{Ii}} = \sum_{i=1}^{n_I} a_{Ii}\hat{t}_{yi\pi} = \sum_{i=1}^{n_I} a_{Ii}\left(\sum_{k=1}^{n_i} \frac{y_k}{\pi_{k/i}}\right)$$

where, $a_{Ii} = \dfrac{1}{\pi_{Ii}}$ is the design weight and $\hat{t}_{yi\pi}$ be the estimator of the cluster total. Using the well known calibration approach we modify the design weight of the estimator. For this purpose the proposed estimator will be,

$$\hat{t}^{c}_{y\pi} = \sum_{i=1}^{n_I} w_{Ii}\hat{t}_{yi\pi}$$

For this purpose, we minimize the chi-square type distance function given by

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

subject to the constraints

$$\sum_{i=1}^{n_I} w_{Ii}z_i = \sum_{i=1}^{N_I} z_i \text{ and } \sum_{i=1}^{n_I} w_{Ii} = \sum_{i=1}^{n_I} a_{Ii}\,.$$

Essentially, the reason for choosing the chi-square distance function is that it minimizes the conditional value of the distance between w_{Ii} and a_{Ii} given the realized sample (Deville and Särndal 1992). This is an optimization problem where we wish to minimize the following function

$$\varphi(w_{Ii},\lambda) = \sum_{i=1}^{n_I} \frac{\left(w_{Ii} - a_{Ii}\right)^2}{a_{Ii}q_{Ii}} + \lambda_1\left[\sum_{i=1}^{n_I} w_{Ii}z_i - \sum_{i=1}^{N_I} z_i\right] + \lambda_2\left[\sum_{i=1}^{n_I} w^{p}{}_{Ii} - \sum_{i=1}^{n_I} a_{Ii}\right]$$

using the method of Lagrange multiplier. Minimization of function $\varphi(w_{Ii},\lambda)$ gives new set of weights

$$w_{Ii} = a_{Ii} + \frac{a_{Ii}q_{Ii}\left(\sum_{i=1}^{N_I} z_i - \sum_{i=1}^{n_I} a_{Ii}z_i\right)}{\sum_{i=1}^{n_I} a_{Ii}q_{Ii}z^2{}_i - \frac{\left(\sum_{i=1}^{n_I} a_{Ii}q_{Ii}z_i\right)^2}{\sum_{i=1}^{n_I} a_{Ii}q_{Ii}}}\left[z_i - \frac{\sum_{i=1}^{n_I} a_{Ii}q_{Ii}z_i}{\sum_{i=1}^{n_I} a_{Ii}q_{Ii}}\right];i=1,2,\ldots,n_I$$

It is noteworthy that these new weights are calibrated to the population total of *z*'s. Here, we considered $q_{Ii} = 1$.

The estimator based on the revised weights is given by

$$\hat{t}^{c}_{y\pi} = \sum_{i=1}^{n_I} a_{Ii}\hat{t}_{yi\pi} + \sum_{i=1}^{n_I} a_{Ii}\left\{\frac{\left(\sum_{i=1}^{N_I} z_i - \sum_{i=1}^{n_I} a_{Ii}z_i\right)}{\sum_{i=1}^{n_I} a_{Ii}z^2{}_i - \frac{\left(\sum_{i=1}^{n_I} a_{Ii}z_i\right)^2}{\sum_{i=1}^{n_I} a_{Ii}}}\right\}\left[z_i - \frac{\sum_{i=1}^{n_I} a_{Ii}z_i}{\sum_{i=1}^{n_I} a_{Ii}}\right]\hat{t}_{yi\pi}$$

$$= \hat{t}_{HT} + \hat{b}\left[\sum_{i=1}^{N_I} z_i - \sum_{i=1}^{n_I} a_{Ii}z_i\right]$$

where, $\hat{b} = \dfrac{\sum_{i=1}^{n_I} a_{Ii}\hat{t}_{yi\pi}\left(z_i - \dfrac{\sum_{i=1}^{n_I} a_{Ii}z_i}{\sum_{i=1}^{n_I} a_{Ii}}\right)}{\left(\sum_{i=1}^{n_I} a_{Ii}z^2{}_i - \dfrac{\left(\sum_{i=1}^{n_I} a_{Ii}z_i\right)^2}{\sum_{i=1}^{n_I} a_{Ii}}\right)}$

Under an equal probability without replacement design (SRSWOR) the estimator is given by

$$\hat{t}^c_{y\pi} = \hat{t}_{HT} + \frac{\sum_{i=1}^{n_I}\hat{t}_{yi\pi}\left[z_i - \frac{1}{n_I}\sum_{i=1}^{n_I} z_i\right]}{\sum_{i=1}^{n_I} z^2{}_i - \frac{1}{n_I}\left(\sum_{i=1}^{n_I} z_i\right)^2}\left[\sum_{i=1}^{N_I} z_i - \frac{N_I}{n_I}\sum_{i=1}^{n_I} z_i\right]$$

Following Särndal *et al.* (1992) this estimator can also be written as,

$$\hat{t}^c_{y\pi} = \sum_{i=1}^{n_I} w_{Ii}\hat{t}_{yi\pi} = \sum_{i=1}^{n_I} a_{Ii}g_{\ is_I}\hat{t}_{yi\pi}$$

The Approximate variance of the proposed estimator under Case 1 was obtained by first order Taylor series linearization technique and was given by

$$V(\hat{t}^c_{y\pi}) = \sum_{i=1}^{N_I}\sum_{j=1}^{N_I}\Delta_{Iij}\frac{U_{Ii}}{\pi_{Ii}}\frac{U_{Ij}}{\pi_{Ij}} + \sum_{i=1}^{N_I}\frac{1}{\pi_{Ii}}\sum_{k=1}^{N_i}\sum_{l=1}^{N_i}\Delta_{kl/i}\frac{y_k}{\pi_{k/i}}\frac{y_l}{\pi_{l/i}},$$

where,

$U_{Ii} = y_{(c)i} - \beta z_i$, $\Delta_{Iij} = (\pi_{Iij} - \pi_{Ii}\pi_{Ij})$,

$\Delta_{kl/i} = \pi_{kl/i} - \pi_{k/i}\pi_{l/i}$, $y_{(c)i} = \sum_{k=1}^{N_i} y_k$, and

$$\beta = \frac{\sum_{i=1}^{N_I} y_{(c)i}z_i - \frac{1}{N_I}\sum_{i=1}^{N_I} y_{(c)i}\sum_{i=1}^{N_I} z_i}{\sum_{i=1}^{N_I} z^2{}_i - N_I\left(\sum_{i=1}^{N_I} z_{\ i}\right)^2}..$$

Following, Särndal *et al.* (1992), the Yates-Grundy form of estimator of variance of the calibration estimator was given by,

$$\hat{V}_{YG}(\hat{t}^c_{y\pi}) = \frac{1}{2}\sum_{i=1}^{n_I}\sum_{j=1}^{n_I} d_{Iij}(w_{Ii}u_{Ii} - w_{Ij}u_{Ij})^2$$

$$+\frac{1}{2}\sum_{i=1}^{n_I}\frac{g^2{}_{is_I}}{\pi^2{}_{Ii}}\sum_{k=1}^{n_i}\sum_{l=1}^{n_i} d_{kl/i}(\frac{y_k}{\pi_{k/i}} - \frac{y_l}{\pi_{l/i}})^2,$$

where,

$u_{Ii} = y_{(c)i} - \hat{\beta}z_i$, $d_{Iij} = \dfrac{(\pi_{Ii}\pi_{Ij} - \pi_{Iij})}{\pi_{Iij}}$,

$d_{kl/i} = \dfrac{(\pi_{k/i}\pi_{l/i} - \pi_{kl/i})}{\pi_{kl/i}}$ and

$$\hat{\beta} = \frac{\sum_{i=1}^{n_I} a_{Ii}y_{(c)i}z_i - \frac{\sum_{i=1}^{n_I} a_{Ii}y_{(c)i}\sum_{i=1}^{n_I} a_{Ii}z_i}{\sum_{i=1}^{n_I} a_{Ii}}}{\sum_{i=1}^{n_I} a_{Ii}z^2{}_i - \frac{\left(\sum_{i=1}^{n_I} a_{Ii}z_{\ i}\right)^2}{\sum_{i=1}^{n_I} a_{Ii}}}.$$

It is acceptable to use the design weights in the variance estimation but Deville and Särndal (1992) suggested that using the calibration weight (w_{Ii}) in the variance estimator makes it both design consistent and nearly model-unbiased. In calibration estimation, increase in number of constraints while calibrating the design weight increases the precision of the estimators. With single constraint $\sum_{i=1}^{n_I} w_{Ii}z_i = \sum_{i=1}^{N_I} z_i$, the proposed estimator becomes the simple regression estimator (GREG) when distance function under consideration is Chi-square distance function (Deville and Särndal, 1992) where as in our case we have used two constraints $\sum_{i=1}^{n_I} w_{Ii}z_i = \sum_{i=1}^{N_I} z_i$ and $\sum_{i=1}^{n_I} w_{Ii} = \sum_{i=1}^{n_I} a_{Ii}$ which is we are adding more information while calibrating the design weight. Adding more information while estimation improves the precision of the estimator and thus our proposed estimator is better than the simple GREG estimator.

3. EMPIRICAL EVALUATION

In this section, we report the results from simulation studies that aim at assessing the performance of the developed calibration

estimators under two stage sampling design with respect to the simple regression estimator Särndal *et al.* (1992, p. 308). In this study we have considered the case of two stage sampling where sample selection at each stage is governed by equal probability without replacement sampling design (SRSWOR). Here, we also have considered the situation that the size of the psu and the corresponding ssus were fixed. For empirical evaluation, a bi-variate normal population is generated and used for the study where BVN (22, 25, 2, 5, *r)*. For the case of simplicity we have assumed that, N_I = 50, N_i = 100 whereas n_I = 15, n_i = 30 and n_I = 20, n_i = 40 and there is availability of auxiliary information for both PSU and SSU level. For the study we have selected a total of 1000 samples from the population using two stage SRSWOR and also considered different levels of correlation between the study variable and the auxiliary variable. We have considered the value of correlation coefficient as r = 0.5, 0.7 and 0.9 for simulation study. We have compared the proposed calibration type regression estimator of the population total ($\hat{t}^{c}_{y\pi}$) with the usual regression estimator ($\hat{t}_{yAr}$) given in Särndal *et al.* (1992, p. 308) under two stage sampling design when auxiliary information is available at cluster level. For the empirical evaluation a SAS macro was developed for selection of the samples using two stage SRSWOR sampling design.

The performance measures used for empirical evaluation were percentage Relative Bias (%RB) and percentage Relative Root Mean Squared Error (%RRMSE). The formula of Relative Bias and Relative Root Mean Squared Error of any estimator of the population parameter θ are given by

$$RB(\hat{\theta}) = \frac{1}{S}\sum_{i=1}^{S}\left(\frac{\hat{\theta}_i - \theta}{\theta}\right)\times 100,$$

$$RRMSE(\hat{\theta}) = \frac{1}{\theta}\sqrt{\frac{1}{S}\sum_{i=1}^{S}(\hat{\theta}_i - \theta)^2}\times 100$$

where, $\hat{\theta}_i$ are the value of the estimator generated through simulation study and θ is the overall population total for the character under study. The results corresponding to %RB of the proposed calibration type regression estimator ($\hat{t}^{c}_{y\pi}$) with respect to the simple regression estimator ($\hat{t}_{yAr}$) were reported in Table 1 whereas the results corresponding to %RRMSE were given in Table 2.

Table 1. %RB of proposed calibration type regression estimator ($\hat{t}^{c}_{y\pi}$) with respect to the simple regression estimator ($\hat{t}_{yAr}$)

Sample Size and Correlation	$\hat{t}^{c}_{y\pi}$	$\hat{t}_{yAr}$
n_I = 15,n_i = 30, r = 0.5	0.154	0.215
n_I = 15,n_i = 30, r = 0.7	0.126	0.128
n_I = 15,n_i = 30, r = 0.9	0.106	0.105
n_I = 20,n_i = 40, r = 0.5	0.114	0.210
n_I = 20,n_i = 40, r = 0.7	0.115	0.185
n_I = 20,n_i = 40, r = 0.9	0.108	0.109

Table 2. %RRMSE of proposed calibration type regression estimator ($\hat{t}^{c}_{y\pi}$) with respect to the simple regression estimator ($\hat{t}_{yAr}$)

Sample Size and Correlation	$\hat{t}^{c}_{y\pi}$	$\hat{t}_{yAr}$
n_I = 15,n_i = 30, r = 0.5	0.061	0.066
n_I = 15,n_i = 30, r = 0.7	0.012	0.024
n_I = 15,n_i = 30, r = 0.9	0.005	0.010
n_I = 20,n_i = 40, r = 0.5	0.014	0.016
n_I = 20,n_i = 40, r = 0.7	0.024	0.028
n_I = 20,n_i = 40, r = 0.9	0.105	0.179

From Table 1 it can be seen that, with respect to %RB the calibration type regression estimator for the situation of availability of auxiliary information at the cluster level was performing better than the simple regression estimator ($\hat{t}_{yAr}$) under two stage sampling design for most of the cases except the situation n_I = 15, n_i = 30, r = 0.9 when the simple regression estimator has less %RB than the calibration type regression estimator ($\hat{t}^{c}_{y\pi}$). Further, it can be seen that for the situations n_I = 15, n_i = 30, r = 0.7 and n_I = 20, n_i = 40, r = 0.9 the proposed estimator and the usual regression estimator have almost the same %RB but after observing all the cases of selection of the sample and correlation between the study variable and the auxiliary variable it can be seen that the proposed calibration type regression

estimator ($\hat{t}_{y\pi}^{c}$) is performing better than the simple regression estimator ($\hat{t}_{yAr}$). Table 2 reveals that, with respect to %RRMSE, the proposed calibration type regression estimator for the situation of availability of auxiliary information at cluster level was performing better than the simple regression estimator ($\hat{t}_{yAr}$) under two stage sampling design. It was evident from the result that the proposed estimator performs better in case of all the different sample sizes drawn from the population. It also performs better than the simple regression estimator ($\hat{t}_{yAr}$) under two stage sampling design for all different levels of correlation between the study variable and the auxiliary variable.

4. CONCLUDING REMARKS

Using the calibration approach proposed by Deville and Särndal (1992) we have been able to develop a regression type estimator of population total when the study and the complex auxiliary variables are linearly related. The proposed calibration type regression estimator of population total performs better than the simple regression estimators given in Särndal *et al.* (1992, p. 308) under two stage sampling design when auxiliary information is available of at cluster level with respect to % relative bias when selection of sample out of the population is done using equal probability without replacement sampling design. Three different levels of correlation between the study variable and the auxiliary variable were considered. Further, it can also be seen that the proposed calibration based regression estimators of population total out performs the simple regression estimator under two stage sampling design with respect to % relative root mean square error. Hence, based on the simulation study, it can also be concluded that the proposed estimator is better than simple regression estimators under two stage sampling design when auxiliary information is available of at cluster level and use of an extra constraint during optimization of the calibration weight increases the precision of the estimator.

REFERENCES

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

Duchesne, P. (1999). Robust calibration estimators. *Survey Methodology*, **25**, 43-56.

Estevao, V.M. and Särndal, C.E. (2006). Survey estimates by calibration on complex auxiliary information. *Inter. Statist. Rev.*, **74**, 127-147.

Estevao, V.M. and Särndal, C.E. (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *J. Official Statist.*, **18(2)**, 233-255.

Folsom, R.E. and Singh, A.C. (2000). The generalized exponential model for design weight calibration for extreme values, nonresponse and post stratification. *Proceedings, Section on Survey Research Methods, American Statistical Association*, 598-603.

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

Kott, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, **32**, 133-142.

Plikusas, A. and pumputis, D. (2010). Estimation of the finite population covariance using calibration. *Lithuanian Maths. J.*, **15**, 325-340.

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

Singh, S., Horn, S., Choudhury, S. and Yu, F. (1999). Calibration of the estimators of variance. *Austr. and New Zealand J. Statist.*, **41(2)**, 199-212.

Singh, Sarjinder, Horn, S. and Yu, F. (1998). Estimation of variance of general regression estimator: Higher level calibration approach. *Survey Methodology*, **24(1)**, 41-50.

Sitter, R.R. and Wu, C. (2002). Efficient estimation of quadratic finite population functions. *J. Amer. Statist. Assoc.*, **97**, 535-543.

Sud, U.C., Chandra, H. and Gupta, V.K. (2014). Calibration approach based regression type estimator for inverse relationship between study and auxiliary variable. *J. Statist. Theo. Practice*, **8(4)**, 707-721.

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.