RASHI 2 (2) : 01-06 (2017)

Calibration Estimators under Two Stage Sampling Design when Population Level Auxiliary Information was not available

[1]Kaustav Aditya, [1]Arpan Bhowmik, [1]Ankur Biswas and [2]Shrila Das

[1]ICAR-Indian Agricultural Statistics Research Institute
[2] ICAR-Indian Agricultural Research Institute

ABSTRACT

Auxiliary information is often used to improve the precision of estimators of finite population total. Calibration approach is widely used for making efficient use of auxiliary information in survey estimation. We proposed the regression type estimators of the population total using the calibration approach under the assumption that the population level auxiliary information is available at secondary stage unit level under two stage sampling design. Through Simulation it was found that all the proposed estimators are performing better than the usual Horvitz-Thompson estimators under two stage sampling design.

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

Introduction

Survey statisticians are always concerned with improvement of methods for estimation of the finite population total, mean, proportion and other parameters. The estimators which use auxiliary variables are often more accurate than the standard ones. Calibration is commonly used in survey sampling to include auxiliary information to increase the precision of the estimators of population parameter. A calibration estimator uses calibrated weights, which are as close as possible, according to a given distance measure, to the original sampling design weights while also respecting a set of constraints, the calibration equations. For every distance measure there is a corresponding set of calibrated weights and a calibration estimator (Deville and Särndal, 1992). Calibration has an intimate link to practice. The fixation on weighting methods on the part of the leading national statistical agencies is a powerful driving force behind calibration. To assign an appropriate weight to an observed variable value, and to sum the weighted variable values to form appropriate aggregates, is firmly rooted procedure. It is used in statistical agencies for estimating various descriptive finite population parameters: totals, means, and functions of totals. Weighting is easy to explain to users and other stakeholders of the statistical agencies. Weighting of units by the inverse of their inclusion probability found firm scientific backing long ago in papers such as Hansen and Hurwitz (1943), Horwitz and Thompson (1952). Weighting became widely accepted. Later, post stratification weighting achieved the same status. Calibration weighting extends both of these ideas. Calibration weighting is outcome dependent~ the weights depend on the observed sample. Calibration is often described as "a way to get consistent estimates". (Here "consistent" refers not to "randomization consistent" but to "consistent with known aggregates".) The calibration equations impose consistency on the weight system, so that, when applied to the auxiliary variables, it will confirm (be consistent with) known aggregates for those same auxiliary variables. Consistency through calibration has a broader implication than just agreement with known population auxiliary totals. Consistency can, for example, be sought with appropriately estimated totals, arising in the current survey or in other surveys.

There are three major advantages of calibration approach in survey sampling.

I. The calibration approach leads to consistent estimates.
II. It provides an important class of technique for the efficient combination of data sources.
III. Calibration approach has computational advantage to calculate estimates.

The calibration approach focuses on the weights given to the units for the purpose of estimation. Calibration implies that a set of starting weights (usually the sampling design weights) are transformed into a set of new weights, called calibrated weights. The calibrated weight of a unit is the product of its initial weight and a calibration factor. The calibration factors are obtained by minimizing a function measuring the distance between the initial weights and the calibrated weights, subject to the constraint that the calibrated weights yield exact estimates of the known auxiliary population totals. The population total is estimated by a linear estimator whose weights are as close as possible to some benchmark weights and which at the same time satisfy some calibration constraints with respect to some suitable auxiliary variables.

E-mail: arpan.stat@gmail.com

Calibration Estimators under Two Stage Sampling Design

Consider a finite population $U=\{1,\ldots, k,\ldots, N\}$ consisting of N units. A sample s of size n is drawn without replacement according to a probabilistic sampling plan with 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. The study variable y is observed for each unit in the sample hence is known for all $i \in s$, and the values $x_1, x_2, ..., x_N$ are known. Let y_i be the value of the variable of interest, y, for the i^{th} population element, with which is also associated an auxiliary variable x_i. For the elements $i \in s$, observe (y_i , x_i). The population total of auxiliary variable x, $X = \sum_{i=1}^{N} x_i$ is assumed to be accurately known. The objective is to estimate the population total $Y = \sum_{i=1}^{N} y_i$. Deville and Sarndal (1992) used calibration on known population total X to modify the basic sampling design weights. Let the Horvitz-Thompson estimator of the population total be $\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \sum_{i=1}^{n} d_i y_i$, where $d_i = \frac{1}{\pi_i}$ is the sampling weight, defined as the inverse of the inclusion probability for unit i. An attractive property of the HT estimator is that it is guaranteed to be unbiased regardless of the sampling design. Its variance under the sampling design is given as

$$V(\hat{Y}_{HT}) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_{ij} - \pi_i\pi_j\right)\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}$$

Now let us suppose that $\{x_i, i = 1, \ldots, N\}$ is available and $X = \sum_{i=1}^{N} x_i$, the population total for x is known. Ideally we would like, $\sum_{i=1}^{n} d_i x_i = X$. But sometimes this is not true. The idea behind calibration estimators is to find weights, $i = 1, ..., N$ close to d_i, based on a distance function, such that, $\sum_{i=1}^{n} w_i x_i = X$. We wish to find weights w_i similar to d_i so as to preserve the unbiased property of the HT estimator. Once w_i is found the calibration estimator for $Y = \sum_{i=1}^{N} y_i$ would be $\hat{Y}_c = \sum_{i=1}^{n} w_i y_i$.

Given a sample s, we want to find , $i = 1, ..., N$ close to d_i based on a distance function $D(w, d)$ subject to the constraint equation $\sum_{i=1}^{n} w_i x_i = X$. The optimization problem where we want to minimize

$$Q(w_1,...,w_n,\lambda) = \sum_{i=1}^{n} D(w_i,\ d_i) - \lambda\left(\sum_{i=1}^{n} w_i x_i - X\right) \quad \ldots (1)$$

using the method of Lagrangian multipliers.

Here q is the tuning parameter that can be manipulated to achieve the optimum minimal of the Eq. (1). A simple case considered by Deville and Sarndal (1992) is the minimization of chi-square type distance function given by $\sum_{i=1}^{n} \frac{(w_i - d_i)^2}{d_i q_i}$. Where q_i are suitably chosen weights. In most of the situations, the value of $q_i = 1$. By minimizing the $\sum_{i=1}^{n} \frac{(w_i - d_i)^2}{d_i q_i}$ subject to constraint equation $\sum_{i=1}^{n} w_i x_i = X$ the weights w_i was obtained

$$w_i = d_i + \frac{d_i q_i x_i}{\sum_{i=1}^{n} d_i q_i x_i^2}\left(X - \sum_{i=1}^{n} d_i x_i\right).$$

Substitution of the value of w_i in $\hat{Y}_c = \sum_{i=1}^{n} w_i y_i$ gives

$$\hat{Y}_c = \sum_{i=1}^{n} d_i y_i + \frac{\sum_{i=1}^{n} d_i q_i x_i y_i}{\sum_{i=1}^{n} d_i q_i {x_i}^2} \left(X - \sum_{i=1}^{n} d_i x_i \right)$$

$$= \hat{Y}_{HT} + \hat{B}\left(X - \hat{X}_{HT}\right)$$

where, $\hat{B} = \frac{\sum_{i=1}^{n} d_i q_i x_i y_i}{\sum_{i=1}^{n} d_i q_i {x_i}^2}$. Written in this form, we see that $\hat{Y}_c$ is the same as the linear GREG estimator (Cassel *et al.*, 1976). In fact, the GREG estimator is a special case of the calibration estimator when the chosen distance function is the Chi-square distance (Deville and Sarndal, 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, an imagined relationship between study variable and auxiliary variable, can have many forms: linear, nonlinear, generalized linear, mixed (model with some fixed, some random effects), and so on. In terms of efficiency, Deville and Sarndal showed that for medium to large samples, the choice of $D(w, d)$ does not make a large impact on the variance of $\hat{Y}_c$. The variance of the calibration estimator was given as,

$$V(\hat{Y}_c) = V\left(\hat{Y}_{HT} + B\left(X - \hat{X}_{HT}\right)\right)$$

$$= V\left(\hat{Y}_{HT} - B\hat{X}_{HT}\right)$$

$$= V\left(\sum_{i=1}^{n} d_i (y_i - Bx_i)\right)$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{N} \Delta_{ij} \left(d_i (y_i - Bx_i)\right)\left(d_j (y_j - Bx_j)\right)$$

As $E(\hat{B}) = B$ then BX is the true population parameter and its variance will become zero. The estimator of variance of the estimator was given as, $E(\hat{B}) = B$. Where $e_i = y_i - \hat{B}x_i$ and $\Delta_{ij} = \left(\pi_{ij} - \pi_i \pi_j\right)$. This technique of calibration is called as the lower level calibration approach. Deville and Sarndal (1992) have also shown that the use of in the variance estimator makes it both design consistent and nearly model unbiased.

Proposed Calibration Estimators

In what follows, regression type estimators have 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. Aditya *et al*. (2016, 2017) developed calibration estimators under two stage sampling design when population level auxiliary information is available. In this paper, we have developed calibration estimators for the situation when there was unavailability of auxiliary information under two stage sampling design using two phase two stage sampling design. In this case we have also considered two cases of unavailability of auxiliary information which are given as,

Case 1. The population level auxiliary information (z_i) is unavailable at the psu level.

Case 2. The population level complete auxiliary information (x_k) is unavailable at the ssu level.

We consider a simple case where information on only one auxiliary variable is available. Let, the population of elements $U=\{1,..., k,..., N_I\}$ is partitioned into clusters, $U_1, U_2,..., U_i,..., U_{Ni}$. 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 $k = 1,..., N$. 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.

Case 1. For this situation, we consider a two phase sampling based approach to estimate the population total of the psu level auxiliary variable z_i. In order to estimate the population total of the auxiliary variable, a large first phase sample s_I' of size n_I' is selected from N_I psus in the population following a sampling design $p_I'(.)$ such that sampling weight for cluster *i* is given by $a_{Ii}' = 1/\pi_{Ii}'$, where $\pi_{Ii}' = P(i \in s_I')$ is first phase inclusion probability of cluster *i*. Then, a smaller second phase sample of s_I of size n_I is selected from s_I' by a sampling design $p_I(.)$ such that the sampling weight for the cluster *i* is $a_{i/s_I'} = 1/\pi_{i/s_I'}$, where $\pi_{i/s_I'} = P(i \in s_I / s_I')$ is the inclusion probability of psu *i*, given s_I' and the study variable is measured on it. We assume that, all the n_I' cluster provided information on the auxiliary variable z_i at the first phase while n_i cluster out of n_I' provide information on study variable. In the second stage the sampling is same as the earlier cases. Hence, the total sampling design weight for the *i*-th cluster is given as $a_{Ii} = a_{Ii}' a_{i/s_I'}$. Here the calibration estimator of the population total will be given as,

$$\hat{t}_{y\pi d}^{c} = \sum_{i=1}^{n_I} w_{i/s_I'} \hat{t}_{yi\pi}.$$

Where, is the calibrated weight. To obtain $w_{i/s_I'}$ we have minimized the chi-square type distance function $\sum_{i=1}^{n_I} \frac{\left(w_{i/s_I'} - a_{Ii}\right)^2}{a_{Ii} q_{Ii}}$ subject to the constraint $\sum_{i=1}^{n_I} w_{i/s_I'} z_i = \sum_{i=1}^{n_I'} a_{Ii}' z_i$ using the Lagrangian multiplier technique. The value of $w_{i/s_I'}$ is given as,

$$w_{i/s_I'} = a_{Ii} + a_{Ii} q_{Ii} z_i \left[\frac{\sum_{i=1}^{n_I'} a'_{Ii} z_i - \sum_{i=1}^{n_I} a_{i/s_I'} z_i}{\sum_{i=1}^{n_I} a_{Ii} q_{Ii} z^2{}_i} \right]$$

Depending on the value of q_{Ii} the calibration estimator can take the form of different estimators as the earlier cases of lower level calibration.

Case 2. For this situation, we consider a two phase sampling based approach to estimate the population total of the ssu level auxiliary variable x_k. In order to estimate the population total of the auxiliary variable, a large first phase sample s_I' of size n_I' is selected from N_i ssus in the population following a sampling design $p_{k/i}(.)$ such that sampling weight for ssu *k* is given by $a_{k/s_i'} = 1/\pi_{i/s_I'}$, where $\pi_{k/s_i'} = P(k \in s_i / s_i')$ is first phase inclusion probability of *k*-th ssu. Then, a smaller second phase sample of s_i of size n_i is selected from s_I' by a sampling design $p_{k/i}(.)$ such that the sampling weight for the ssu *k* is $a_{k/s_i'} = 1/\pi_{i/s_i'}$, where $\pi_{k/s_i'} = P(k \in s_i / s_i')$ is the inclusion probability of ssu *k*, given s_I' and the study variable is measured on it. Hence, the total sampling design weight for the *k*-th ssu is given as $a_k = a'_{k/i} a_{k/s_i'}$. Here the calibration estimator of the population total will be given as,

$$\hat{t}^*_{y\pi ud} = \sum_{i=1}^{N_I} \sum_{k=1}^{N_i} \hat{y}_k + \sum_{k=1}^{n_s} w_{dk} e_{ks}$$

where, $\hat{y}_k = \hat{\beta} x_k$, $e_{ks} = y_k - \hat{y}_k$, $\hat{\beta} = \dfrac{\sum_{k=1}^{n_s} a_k y_k x_k}{\sum_{k=1}^{n_s} a_k x^2{}_k}$ and rest of the terms are defined earlier. Here w_{dk} is the calibration weight obtained by minimizing the chi-square type distance function $\sum_{k=1}^{n_s} \dfrac{(w_{dk} - a_k)^2}{a_k q_k}$ subject to the constraint $\sum_{k=1}^{n_s} w_{dk} x_k = \sum_{k=1}^{n_I'} a'_{k/i} x_k$ using the Lagrangian multiplier technique.

The value of w_{dk} is given as, $w_{dk} = a_k + \dfrac{a_k q_k x_k}{\sum_{k=1}^{n_s} a_k q_k x^2{}_k} \left(\sum_{k=1}^{n_I'} a'_{k/i} x_k - \sum_{k=1}^{n_s} a_k x_k \right)$.

Simulation Study and Conclusion

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/ssu is 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 and N_i =100 whereas the selected samples are of size n_I =15, n_i =30 and n_I = 20, n_i = 40 and there is availability of auxiliary information for 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. Through Simulation it was found that all the proposed estimators are performing better than the usual Horvitz-Thompson estimators under two stage sampling design.

REFERENCES

Aditya Kaustav, Sud UC, Chandra Hukum and Biswas Ankur (2016). Calibration Based Regression Type Estimator of the Population Total under Two Stage Sampling Design. *Journal of the Indian Society of Agricultural Statistics*, **70(1)**: 19-24.

Aditya Kaustav, Biswas Ankur, Gupta K Ashok and Chandra Hukum (2017). District Level Crop Yield Estimation Using Calibration Approach. *Current Science*. **112 (9)**: 1927-31.

Aditya Kaustav, Sud, U.C. and Chandra Hukum (2016). Calibration Approach Based Estimation of Finite Population Total Under Two Stage Sampling. *J. Indian Society of Agricultural Statistics*. **70(3)**: 219-26.

Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling. *J. American Statistical Association*, **87**: 376-82.

Estevao, V. M. and Särndal, C. E. (2006). Survey Estimates by Calibration on Complex Auxiliary Information, *International Statistical Review*, **74**:127-47.

Estevao, V. M. and Särndal, C.E. (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *J. Official Statistics*, **18(2)**: 233-55.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. American Statistical Association*, **47**: 663-85.

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.Statistical theory and Practice.* In press.

Sud, U. C., Chandra, H. and Gupta, V. K. (2014). Calibration based product estimator in single and two phase sampling. *Journal of Statistical theory and Practice.* In press.

Wu, C., and Sitter, R.R. (2001). A model calibration approach to using complete auxiliary information from survey data. *J. American Statistical Association*, **96**:185-93.