Application of Calibration Approach for Regression Coefficient Estimation under Two-stage Sampling Design

Pradip Basak, Kaustav Aditya, Hukum Chandra and U.C. Sud
ICAR-Indian Agricultural Statistics Research Institute, New Delhi
Email: pradipbasak99@gmail.com, ucsud@iasri.res.in, hchandra@iasri.res.in and kaustav@iasri.res.in

1. Introduction

Now a days, survey data are complex and multivariate in nature which involves clustering, stratification, unequal probability of selection, multi-stages and multi-phases. The traditional method of estimation of regression coefficient is ordinary lest squares (OLS) estimation which is based on the assumption that sample observations are drawn independently. This assumption of independence holds only if the sample observations are drawn using simple random sampling with replacement (SRSWR) but for other sampling designs it does not hold. One such complex design is two-stage sampling design which is widely used in large scale surveys. In two-stage sampling, sample is selected in two stages. In the first stage, clusters are selected and in the second stage, a specified number of elements are investigated from the selected clusters. The clusters which form the units of sampling at the first stage are known as primary stage units (PSU) and the elements within the clusters are known as second stage units (SSU). As for example, in case of crop, surveying fields can be taken as first stage units and plots within the fields can be taken as second stage units.

Kish and Frankel (1974) suggested use of sampling design weights in the estimation procedure as an alternative to the OLS. Estimation of regression coefficient based on maximum likelihood estimation was suggested by Holt, Smith and Winter (1980).In the presence of auxiliary information, calibration approach was suggested by Deville and Särndal, 1992for the improvement of the estimator of population parameters. Work on calibration approach based estimation of population parameters like mean, total, proportion, covariance has already been done under uni-stage or multi-stage designs, see for example Aditya *et al.* (2016), Plikusas and Pumputis (2007, 2010).Thus, under the availability of auxiliary information in the two-stage design, the theory of calibration approach is used here for the improvement of the estimator of population regression coefficient.

2. Methodology

Let *U*=(*1,2,…,k,…,N*) be a finite population of size *N* comprising of N_I clusters as $U_1, U_2, \dots, U_i, \dots, U_{N_I}$ with size of the clusters $N_1, N_2, \dots, N_3, \dots, N_I$ respectively. These clusters are nothing but primary stage units (psus) and units within the clusters are second stage units (ssus). At the first stage, a sample of clusters s_I of size n_I is drawn from the population of clusters U_I and at the second stage, a sample of units s_i of size n_i is drawn from the *i*th selected cluster, U_i of size N_i by using any probability sampling scheme. Let, π_{Ii} and π_{Iij} be the first and second order inclusion probability at the first stage and at the second stage it is $\pi_{k/i}$ and $\pi_{kl/i}$ respectively. Let, $a_{Ii} = 1/\pi_{Ii}$, $a_{k/i} = 1/\pi_{k/i}$ and $a_{ik} = a_{Ii}a_{k/i}$.

Let, *y* and *x* be the variables under study. Here, *y* is dependent variable and *x* is explanatory variable. Let us assume, auxiliary variable *z* is associated with dependent variable *y* and information on auxiliary variable *z* is available at psu level. Let, the sample observations corresponding to the jth unit of ith cluster are denoted by y_{ik}, x_{ik} and z_{ik}. Now, the population total of variables *y*, *x* and *z* are given by $t_y = \sum_{i=1}^{N_I}\sum_{k=1}^{N_i} y_{ik} = \sum_{i=1}^{N_I} t_{iy}$, $t_x = \sum_{i=1}^{N_I}\sum_{k=1}^{N_i} x_{ik} = \sum_{i=1}^{N_I} t_{ix}$ and $t_z = \sum_{i=1}^{N_I}\sum_{k=1}^{N_i} z_{ik} = \sum_{i=1}^{N_I} Z_i$ respectively, where t_{iy}, t_{ix} and Z_i is the *i*thcluster total of *y*, *x* and *z* respectively. We have assumed that Z_i is known for all psu's.

Population regression coefficient *B* under two-stage sampling design is given by

$$B = \frac{\sum_{i=1}^{N_I}\sum_{k=1}^{N_i}\left(x_{ik} - \bar{X}\right)\left(y_{ik} - \bar{Y}\right)}{\sum_{i=1}^{N_I}\sum_{k=1}^{N_i}\left(x_{ik} - \bar{X}\right)^2}$$

where $\bar{X} = \frac{1}{N}\sum_{i=1}^{N_I}\sum_{k=1}^{N_i} x_{ik}$ and $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N_I}\sum_{k=1}^{N_i} y_{ik}$.

The usual *π*-estimator of this population regression coefficient, *B* is given by

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

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

Thus, using calibration approach the estimator of population total of variable *y* is obtained as $\hat{t}_{y\pi}^c = \sum_{i=1}^{n_I} w_{Ii}\hat{t}_{iy}$ .Finally, the estimator of population regression coefficient under two-stage design is obtained as

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

3. Empirical Evaluation

A population of 284 municipalities of Sweden containing information on several variables was used for empirical evaluation. The population was grouped into 50 clusters each containing 5 to 9 municipalities. At the first stage, some clusters were selected from the 50 clusters using simple random sampling without replacement and at the second stage some municipalities were selected from each selected clusters using same sample design. From the selected municipalities observations were recorded on the variables 1985 Municipal taxation (RMT85, measured in millions of kronor), total number of seats in the municipal council (S82) and number of municipal employees in 1984 (ME84). The objective was to study the pattern of relationship between variables RMT85 and S82 using ME84 as the auxiliary variable. From this population, three different combinations of sample: i) $n_I = 20,\ n_i = 4,\ n_s = 80,$ ii) $n_I = 20,\ n_i = 2,\ n_s = 40,$ and iii) $n_I = 10,\ n_i = 2,\ n_s = 20,$ were drawn. In the empirical evaluation, two estimators of finite population regression coefficient were considered for comparison purpose:

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

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

The performance of the estimators were evaluated by the criteria of percentage absolute relative bias (ARB) and percentage relative root mean square error (RRMSE)

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

where $\hat{B}_i$ denotes the estimated value of population regression coefficient at simulation run *i*, with true value *B*.

The result of the empirical study indicates that the calibrated estimator (EST-CAL1) has an lower ARB as compared to the π-estimator (EST-π). Similarly, in terms of RRMSE the estimator EST-CAL1 has an advantage as compared to the existing π-estimator. The results are displayed through a graphical representation in Figure 1 and Figure2.

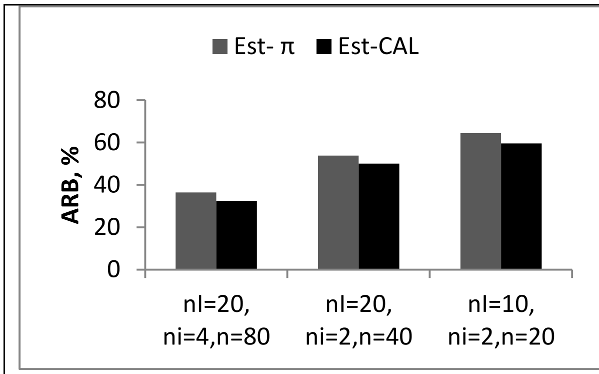

Figure 1

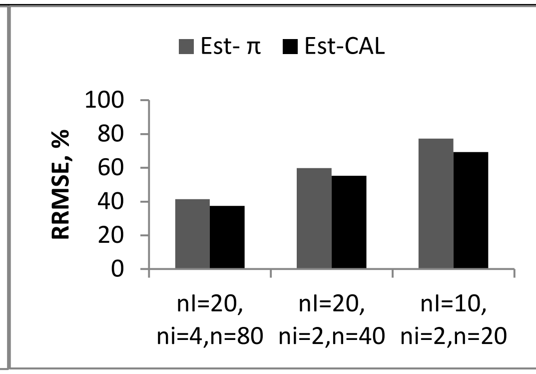

Figure 2

4. Concluding Remarks

This study discusses about the calibrated estimator of population regression coefficient in the presence of psu level auxiliary information. The calibrated estimator found to be satisfactory as compared to the existing OLS estimator which violates the independence of observations assumption under two-stage sampling design.

References (if any)

1. Aditya, K., Sud, U. C., Chandra, H. and Biswas, A. (2016).Calibration based regression type estimator of the population total under two stage sampling design. *Journal of Indian Society of Agricultural Statistics*, **70(1)**, 19-24.
2. Deville, J. C. and Sarndal, C. E. (1992).Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**, 376-382.
3. Holt, D., Smith, T. M. F. and Winter, P. D. (1980).Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society*. *Series A (General)*,**143**, 474-487.
4. Kish, L. and Frankel, M.R. (1974).Inference from complex samples. *Journal of the Royal Statistical Society*, **B36**, 1-37.
5. Plikusas, A. and Pumputis, D. (2007). Calibrated estimators of the population covariance. *Acta Applicandae Mathematicae*, **97**, 177-187.
6. Plikusas, A. and Pumputis, D. (2010). Estimation of finite population covariance using calibration. *Nonlinear Analysis: Modelling and Control*, **15(3)**, 325-340.
7. Särndal, C. E., Swensson, B. and Wretman, J. (1992).*Model Assisted Survey Sampling*. Springer Verlag, New York.

Terms - Do not remove or change this section (It should be emailed back to us as is)

- This form is for genuine submissions related to biotechnology topics only.
- You should be the legal owner and author of this article and all its contents.
- If we find that your article is already present online or even containing sections of copied content then we treat as duplicate content - such submissions are quietly rejected.
- If your article is not published within 3-4 days of emailing, then we have not accepted your submission. Our decision is final therefore do not email us enquiring why your article was not published. We will not reply. We reserve all rights on this website.
- Do not violate copyright of others, you will be solely responsible if anyone raises a dispute regarding it.
- Similar to paper based magazines, we do not allow editing of articles once they are published. Therefore please revise and re-revise your article before sending it to us.
- Too short and too long articles are not accepted. Your article must be between 500 and 5000 words.
- We do not charge or pay for any submissions. We do not publish marketing only articles or inappropriate submissions.
- Full submission guidelines are located here: http://www.biotecharticles.com/submitguide.php
- Full Website terms of service are located here: http://www.biotecharticles.com/privacy.php

As I send my article to be published on BiotechArticles.com, I fully agree to all these terms and conditions.