Calibrated and model-calibrated estimators of the finite population covariance

Dalius Pumputis[1]

[1] Vilnius Pedagogical University, Institute of Mathematics and Informatics, Lithuania
e-mail: dpumputis@vpu.lt, dpumputis@yahoo.co.uk

Abstract

Two types of calibrated estimators of the finite population covariance are considered. The estimators of the first type are defined by some different calibration equations and loss functions. They may use one or several weighting systems. The estimators of the second type are constructed using linear regression model, some calibration equations and loss functions. The estimators are compared by simulation.

1 Introduction

The calibration method introduced by Deville and Särndal (1992) was used to improve the estimators of the finite population total using known auxiliary variables. The estimation of more complicated finite population parameters is also important, but the literature on this topic is not wide. The calibrated estimator of the ratio of two totals was introduced by Krapavickaitė and Plikusas (2005). Harms and Duchesne (2006), Rueda, Martínez, Martínez and Arcos (2007) considered calibrated estimators of quantiles and estimation of the distribution function with calibration methods. In the paper Plikusas and Pumputis (2007) some calibrated estimators of the finite population covariance were introduced. These estimators have one weighting system and are defined by using different calibration equations and different loss functions. The definition given in this paper may be extended to the case of multiple weighting systems. Sitter and Wu (2002) proposed model-calibrated method to estimate the finite population covariance. The estimators are constructed under the assumption that the relationship between study variables y, z and auxiliary variable $\mathbf{x} = (x_1, x_2, \ldots, x_J)$ can be described by a linear regression model and using some calibration equations and loss function. In the following sections we recall these estimators and present some simulation study.

2 Calibrated estimators of the finite population covariance

Consider a finite population $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ of N elements. Let y and z be two study variables defined on the population $\mathcal{U}$ and taking real nonnegative values. The values of the variables y and z are not known. Suppose the known auxiliary variables a and b are available.

Let the covariance

$$Cov(y,z) = \frac{1}{N-1}\sum_{k=1}^{N}\left(y_k - \frac{1}{N}\sum_{k=1}^{N} y_k\right)\left(z_k - \frac{1}{N}\sum_{k=1}^{N} z_k\right)$$

be parameter of interest.

2.1 Estimators with one weighting system

In the book (Särndal, Swensson, Wretman 1992, p. 187) one can find well-known only design based estimator of the covariance

$$\widehat{Cov}(y,z)=\frac{1}{N-1}\sum_{k\in\mathbf{s}}d_k\Big(y_k-\frac{1}{N}\sum_{k\in\mathbf{s}}d_ky_k\Big)\Big(z_k-\frac{1}{N}\sum_{k\in\mathbf{s}}d_kz_k\Big), \tag{1}$$

here d_k- sample design weights.

We modify the design weights d_k and consider the calibrated estimator of the covariance of the following shape

$$\widehat{Cov}_w^{(1)}(y,z)=\frac{1}{N-1}\sum_{k\in\mathbf{s}}w_k\Big(y_k-\frac{1}{N}\sum_{k\in\mathbf{s}}w_ky_k\Big)\Big(z_k-\frac{1}{N}\sum_{k\in\mathbf{s}}w_kz_k\Big).$$

The new weights w_k are defined under the following conditions:

a) the weights w_k satisfy some calibration equation;

b) the distance between the weights d_k and w_k is minimal according to some loss function L.

The conditions a) and b) can be specified in different ways. Let us introduce three different calibration equations:

N)

$$\frac{1}{N-1}\sum_{k\in\mathbf{s}}w_k(a_k-\widehat{\mu}_{aw})(b_k-\widehat{\mu}_{bw})=Cov(a,b),$$

here

$$\widehat{\mu}_{aw}=\frac{1}{N}\sum_{k\in\mathbf{s}}w_ka_k,\quad \widehat{\mu}_{bw}=\frac{1}{N}\sum_{k\in\mathbf{s}}w_kb_k.$$

L)

$$\frac{1}{N-1}\sum_{k\in\mathbf{s}}w_k(a_k-\mu_a)(b_k-\mu_b)=Cov(a,b), \tag{2}$$

$$\mu_a=\frac{1}{N}\sum_{k=1}^{N}a_k,\mu_b=\frac{1}{N}\sum_{k=1}^{N}b_k,$$

T)

$$\sum_{k\in\mathbf{s}}w_ka_k=\sum_{k=1}^{N}a_k,\quad \sum_{k\in\mathbf{s}}w_kb_k=\sum_{k=1}^{N}b_k.$$

We will call three types of calibration, respectively to the calibration equations listed above, as *nonlinear calibration*, *linear calibration* and *calibration of the totals.*

Below we present the list of loss functions that can be used for the final specification of calibrated weights w_k:

$$L_1 = \sum_{k\in\mathbf{s}} \frac{(w_k - d_k)^2}{d_k q_k}, \qquad L_2 = \sum_{k\in\mathbf{s}} \frac{w_k}{q_k}\log\frac{w_k}{d_k} - \frac{1}{q_k}(w_k - d_k),$$

$$L_3 = \sum_{k\in\mathbf{s}} 2\frac{\left(\sqrt{w_k} - \sqrt{d_k}\right)^2}{q_k}, \qquad L_4 = \sum_{k\in\mathbf{s}} -\frac{d_k}{q_k}\log\frac{w_k}{d_k} + \frac{1}{q_k}(w_k - d_k),$$

$$L_5 = \sum_{k\in\mathbf{s}} \frac{(w_k - d_k)^2}{w_k q_k}, \qquad L_6 = \sum_{k\in\mathbf{s}} \frac{1}{q_k}\left(\frac{w_k}{d_k} - 1\right)^2,$$

$$L_7 = \sum_{k\in\mathbf{s}} \frac{1}{q_k}\left(\frac{\sqrt{w_k}}{\sqrt{d_k}} - 1\right)^2.$$

2.2 Estimators with two weighting systems

Let us consider some other, more general estimators of the finite population covariance, which are constructed using two weighting systems and having the following shape:

$$\widehat{Cov}_w^{(2)}(y,z) = \frac{1}{N-1}\sum_{k\in\mathbf{s}} w_k^{(1)}\left(y_k - \frac{1}{N}\sum_{k\in\mathbf{s}} w_k^{(2)} y_k\right)\left(z_k - \frac{1}{N}\sum_{k\in\mathbf{s}} w_k^{(2)} z_k\right), \tag{3}$$

here the weights $w_k^{(1)}$ and $w_k^{(2)}$ satisfy one of the calibration equations proposed and minimize the distance between the design weights and calibrated weights according to some loss function. For example, the weights $w_k^{(1)}$ can be defined using the linear calibration and loss function L_1, whereas the weights $w_k^{(2)}$ can be calculated using the calibration of totals and the same loss function.

Many experiments showed, that estimators, which use two systems of weights, are more precise. They have smaller variance, mean square error and coefficient of variation.

3 Model-calibrated estimators of the finite population covariance

Consider a finite population $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ consisted of N elements. Let's associate unit u_i with the vectors $\mathbf{y}_i$ and $\mathbf{x}_i$ of study and auxiliary variables values. In the paper of Sitter and Wu (2002) a quadratic finite population function is taken as a parameter of interest. Every quadratic function can be defined as

$$T = \sum_{i=1}^{N}\sum_{j=i+1}^{N} \phi(\mathbf{y}_i, \mathbf{y}_j),$$

here $\phi(\cdot,\cdot)$ is a symmetric function. The function T may be expressed as

$$T = \sum_{\alpha=1}^{N^*} t_\alpha,$$

here α is the number prescribed to the pair (ij) of indexes in the sequence of all possible pairs satisfying condition $i < j$; the corresponding set of indexes is denoted by $\mathbf{s}^*$; $t_\alpha = \phi(\mathbf{y}_i, \mathbf{y}_j)$, $N^* = N(N-1)/2$. So T is a population total defined on the population of N^* elements and may be estimated by model-calibrated estimators, proposed in the paper of Wu and Sitter (2001).

The model-calibrated approach is extended for the quadratic functions and a model-calibrated estimator for such a type of functions is defined as follows:

$$\hat{T}_{MC} = \sum_{i\in\mathbf{s}}\sum_{j>i} d_{ij}\phi(\mathbf{y}_i,\mathbf{y}_j) + \left\{\sum_{i=1}^{N}\sum_{j=i+1}^{N}\phi(\hat{\mathbf{y}}_i,\hat{\mathbf{y}}_j) - \sum_{i\in\mathbf{s}}\sum_{j>i} d_{ij}\phi(\hat{\mathbf{y}}_i,\hat{\mathbf{y}}_j)\right\}\hat{B}, \tag{4}$$

here $d_{ij} = 1/\pi_{ij}$, π_{ij} - the second-order inclusion probabilities, $\hat{B} = C(u,v)/C(u,u)$,
$C(u,v) = \sum\sum_{(ij)\in\mathbf{s}^*} d_{ij}q_{ij}(u_{ij}-\bar{u})(v_{ij}-\bar{v})$, $u_{ij} = \phi(\hat{\mathbf{y}}_i,\hat{\mathbf{y}}_j)$, $v_{ij} = \phi(\mathbf{y}_i,\mathbf{y}_j)$,
$\bar{u} = \sum\sum_{(ij)\in\mathbf{s}^*} d_{ij}q_{ij}u_{ij}/\sum\sum_{(ij)\in\mathbf{s}^*} d_{ij}q_{ij}$, and $\bar{v}$, $C(u,u)$ are defined similarly; $\hat{\mathbf{y}}_i$ are fitted values, which we get using a certain semi-parametric model ξ (Sitter, Wu 2002, p. 535-543); q_{ij} are known positive weights.
In the special case when we take $\phi(\mathbf{y}_i,\mathbf{y}_j) = \frac{1}{N(N-1)}(y_i-y_j)(z_i-z_j)$, $\mathbf{y}_i = (y_i,z_i)'$, quadratic function T would be the population covariance

$$Cov(y,z) = \frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j=i+1}^{N}(y_i-y_j)(z_i-z_j).$$

Suppose the the relationship between y_i and $\mathbf{x}_i$ (z_i and $\mathbf{x}_i$) can be described by the linear regression model $E_\xi(y_i) = \mathbf{x}_i^{'}\beta$ ($E_\xi(z_i) = \mathbf{x}_i^{'}\gamma$). Then using the expression (4) we get the model-calibrated estimator of the population covariance

$$\widehat{Cov}_{MC}(y,z) = \widehat{Cov}_{HT} + \hat{\beta}^{'}(S_{\mathbf{x}}^2 - s_{\mathbf{x}}^2)\hat{\gamma}\hat{B} \tag{5}$$

here $\hat{\beta} = \left\{\sum_{i\in\mathbf{s}} d_i\mathbf{x}_i\mathbf{x}_i^{'}\right\}^{-1}\sum_{i\in\mathbf{s}} d_i\mathbf{x}_i y_i$ is the design-based estimator of the regression coefficients β. Estimator of γ is defined similarly;

$$\widehat{Cov}_{HT} = \frac{1}{N(N-1)}\sum_{i\in\mathbf{s}}\sum_{j>i} d_{ij}(y_i-y_j)(z_i-z_j),$$

$$S_{\mathbf{x}}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i-\bar{\mathbf{X}})(\mathbf{x}_i-\bar{\mathbf{X}})^{'}, \quad s_{\mathbf{x}}^2 = \frac{1}{N(N-1)}\sum_{i\in\mathbf{s}}\sum_{j>i} d_{ij}(\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^{'}.$$

In the next section we will compare by simulation the calibrated estimators, which use two systems of weights, with model-calibrated estimator of the covariance. It should be noted that theoretical comparison possess some difficulties.

4 Simulation study

The simulation study is carried out to compare the two calibrated estimators of the form (3) with model-calibrated estimator of the finite population covariance (5). For the definition of the first estimator the loss functions L_1, and L_6 are used to specify the calibrated weights $w_k^{(1)}$ and $w_k^{(2)}$. For both cases the weights $w_k^{(1)}$ are required to satisfy the calibration equation (2), whereas the system of weights $w_k^{(2)}$ is defined using calibration of total. The respective calibrated estimators are denoted by $\widehat{Cov}_{w1}^{(2)}(y,z)$, $\widehat{Cov}_{w6}^{(2)}(y,z)$ referring to the loss functions L_1 and L_6. To show the advantages of use of auxiliary information, the simple only design based estimator of the covariance (1) was also included in to simulation.

Table 1. The main estimated characteristics of accuracy for the estimators of the finite population covariance.
True value of covariance: $Cov(y,z) = 65862789$

Estimator	Estimate	Variance	Bias	MSE	cv
$\rho(y,a) = 0.81,\ \rho(z,b) = 0.90,\ \rho(y,b) = 0.63,\ \rho(z,a) = 0.60$					
$\widehat{Cov}^{(2)}_{w1}(y,z)$	66187636	2.21E+13	104570	2.21E+13	0.0710
$\widehat{Cov}^{(2)}_{w6}(y,z)$	66064504	2.18E+13	-18562	2.18E+13	0.0706
$\widehat{Cov}_{MC}(y,z)$	75642103	8.90E+13	9559036	1.80E+14	0.1247
$\widehat{Cov}(y,z)$	61292369	9.92E+13	-4790697	1.22E+14	0.1625
$\rho(y,a) = 0.21,\ \rho(z,b) = 0.90,\ \rho(y,b) = 0.63,\ \rho(z,a) = 0.15$					
$\widehat{Cov}^{(2)}_{w1}(y,z)$	60549920	1.00E+14	-5533146	1.31E+14	0.1655
$\widehat{Cov}^{(2)}_{w6}(y,z)$	60510963	1.00E+14	-5572103	1.31E+14	0.1653
$\widehat{Cov}_{MC}(y,z)$	83499670	2.44E+14	17416604	5.48E+14	0.1872
$\widehat{Cov}(y,z)$	61029215	1.04E+14	-5053851	1.30E+14	0.1674
$\rho(y,a) = 0.23,\ \rho(z,b) = 0.31,\ \rho(y,b) = 0.19,\ \rho(z,a) = 0.16$					
$\widehat{Cov}^{(2)}_{w1}(y,z)$	60959385	1.04E+14	-5123681	1.31E+14	0.1676
$\widehat{Cov}^{(2)}_{w6}(y,z)$	60953712	1.04E+14	-5129355	1.31E+14	0.1675
$\widehat{Cov}_{MC}(y,z)$	60990137	9.53E+13	-5092930	1.21E+14	0.1601
$\widehat{Cov}(y,z)$	61173915	1.04E+14	-4909151	1.28E+14	0.1664

We consider the real population of size 300 from the Lithuanian Enterprise Survey. This population is stratified into two strata by the size of the survey variable y. The stratified simple random sample is used as a sample design. The sample size $n = 100$ is allocated to strata, using Neymans optimal allocation.

1000 samples were drawn and for each of them two calibrated estimators $\widehat{Cov}^{(2)}_{w1}(y,z)$, $\widehat{Cov}^{(2)}_{w6}(y,z)$, model-calibrated estimator $\widehat{Cov}_{MC}(y,z)$ and design based estimator $\widehat{Cov}(y,z)$ were computed. The estimated variance, bias, mean square error (MSE) and coefficient of variation (cv) for each estimator and for some different sets of auxiliaries, having different correlation ρ with the study variables, are presented in Table 1. The auxiliary variables $\mathbf{x} = (a,b)$ and equal weights $q_{ij} = 1$ were used for the model-calibrated estimator as well as the weights $q_k = 1$ are imputed for the calibrated estimators $\widehat{Cov}^{(2)}_{w1}(y,z)$ and $\widehat{Cov}^{(2)}_{w6}(y,z)$.

For the fixed correlation between auxiliary and study variables the calibrated estimators $\widehat{Cov}^{(2)}_{w1}(y,z)$, $\widehat{Cov}^{(2)}_{w6}(y,z)$ produce similar results. The reason for this is that they belong to the same class of estimators and both are constructed using very similar loss functions.

In the case of a highly correlated auxiliary variables the estimators, which use two systems of weights, are the best, despite our preliminary expectation that accuracy of model-calibrated estimator be similar to $\widehat{Cov}^{(2)}_{w1}(y,z)$, $\widehat{Cov}^{(2)}_{w6}(y,z)$ or even higher.

When only one well correlated auxiliary variable is taken ($\rho(y,a) = 0.21$ and $\rho(z,b) = 0.90$), the model-calibrated estimator has the biggest variance and MSE. The estimators $\widehat{Cov}^{(2)}_{w1}(y,z)$, $\widehat{Cov}^{(2)}_{w6}(y,z)$ perform slightly better than the design based estimator of the covariance $\widehat{Cov}(y,z)$.

In the case of low correlated auxiliary variables all the estimators are of similar quality. The model-

calibrated estimator has a little bit smaller variance and MSE.

To conclude, it is difficult to say if pattern of the Table 1 would remain the same in the case of other population when data would be described ideally by a linear regression model. Further simulation study is needed.

References

Deville, J. C., Särndal, C. E. (1992) Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**, 376-382.

Harms, T., Duchesne, P. (2006) On calibration estimation for quantiles. *Survey Methodology*, **52**, 37-52.

Krapavickaitė, D., Plikusas, A. (2005) Estimation of a Ratio in the Finite Population. *Informatica*, **16**, 347-364.

Plikusas, A., Pumputis, D. (2007) Calibrated estimators of the population covariance. *Acta Applicandae Mathematicae*, **97**, 177-187.

Rueda, M., Martínez, S., Martínez, H., Arcos, A. (2007) Calibration methods for estimating quantiles. *Metrika*, **66**, 355-371.

Rueda, M., Martínez, S., Martínez, H., Arcos, A. (2007) Estimation of the distribution function with calibration methods. *Journal of Statistical Planning and Inference*, **137**, 435-448.

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

Sitter, R. R., Wu, C. (2002) Efficient Estimation of Quadratic Finite Population Functions in the Presence of Auxiliary Information. *Journal of the American Statistical Association*, **97**, 535-543.

Wu, C., Sitter, R. R. (2001) A model-calibration Approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, **96**, 185-193.