ESTIMATION OF THE FINITE POPULATION COVARIANCE

A. Plikusas, D. Pumputis
Institute of Mathematics and Informatics, Vilnius, LITHUANIA
e-mail: plikusas@ktl.mii.lt, dpumputis@yahoo.co.uk

Abstract

Some calibrated estimators of the finite population covariance are presented. The estimators are constructed using different calibration equations and different loss functions. In the most cases the explicit solution of the calibration problem does not exist. The approximate iterative equations for the calibrated weights can be derived.

1 Introduction

There are various statistical methods for improving estimators, using auxiliary information. The calibration method is one of them. The growing power of calculation abilities stimulates the consideration of estimators, based on the extensive use of auxiliary information. The calibrated estimators of the finite population totals are widely used in the practice of the official statistics. These estimators enables to reduce the variance of estimates, provided the auxiliary variables are well correlated with the study variable. In the case of social surveys, the calibrated estimators are attractive because of their property to estimate the known totals of the auxiliary variables without error. We present some estimators of the finite population covariance which are also constructed using auxiliary variables.

2 Formulation of the problem

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 values \( \{y_1, y_2, \ldots, y_N\} \) and \( \{z_1, z_2, \ldots, z_N\} \) respectively. The values of the variables \( y \) and \( z \) are not known. We are interested in the estimation of the covariance

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

One of it’s standard estimators (see Särndal at all (1992)) is,

\[
\hat{Cov}(y, z) = \frac{1}{N-1} \sum_{k \in s} d_k \left( y_k - \frac{1}{N} \sum_{k \in s} d_k y_k \right) \left( z_k - \frac{1}{N} \sum_{k \in s} d_k z_k \right).
\]

Here \( s \) denotes a probability sample set, \( d_k = 1/\pi_k \) are sample design weights, \( \pi_k \) is a probability of inclusion of the element \( k \) into the sample set \( s \).
Suppose, that some auxiliary information is available. Let the variable $a$ with the population values $\{a_1, a_2, \ldots, a_N\}$ and the variable $b$ with the values $\{b_1, b_2, \ldots, b_N\}$ be known auxiliary variables. Denote their known covariance by $Cov(a, b)$. We will construct a new calibrated estimator of the $Cov(y, z)$ using known auxiliary variables. If the auxiliary variables are well correlated with the study variables, the variance of the calibrated estimator can be smaller compare to the variance of only design based estimator $\widehat{Cov}(y, z)$.

In general the calibrated estimator is defined by the calibrated weights $w_k$, that are used instead of the design weights $d_k$ and satisfy the following conditions:

- a) the weights $w_k$ of the calibrated estimator satisfy some calibration equation;
- b) the distance between the design weights $d_k$ and calibrated weights $w_k$ is minimal under the some loss function $L$.

Several loss functions may be used to measure the distance between the design weights $d_k$ and calibrated weights $w_k$. Below there are some examples

\[
L_1 = \sum_{k \in s} \frac{(w_k - d_k)^2}{d_k q_k}, \quad L_2 = \sum_{k \in s} \frac{w_k}{q_k} \log \frac{w_k}{d_k} - \frac{1}{q_k}(w_k - d_k),
\]

\[
L_6 = \sum_{k \in s} \frac{1}{q_k} \left( \frac{w_k}{d_k} - 1 \right)^2, \quad L_7 = \sum_{k \in s} \frac{1}{q_k} \left( \sqrt{\frac{w_k}{d_k}} - 1 \right)^2.
\]

Here the weights $q_k > 0$, $k \in s$, are free additional weights. We can modify the calibrated estimator by choosing these weights. Otherwise one can take $q_k = 1$ for all $k$. In practice the loss function $L_1$ is beeing used.

Simulation results show, that the efficiency (mean square error) of the calibrated estimator does not depend significantly on the kind of loss function used, and we can say opposite concerning the calibration equation.

Let us specify the definition of a calibrated estimator by considering some different cases.

### 3 Estimators with one weighting system

Construct a calibrated estimator using one system of calibrated weights. In this case the estimator we looking for is of the form

\[
\widehat{Cov}_{w}^{(1)}(y, z) = \frac{1}{N-1} \sum_{k \in s} w_k \left( y_k - \frac{1}{N} \sum_{k \in s} w_k y_k \right) \left( z_k - \frac{1}{N} \sum_{k \in s} w_k z_k \right).
\]

Let us consider three types of calibration equations in order to define the weights $w_k$, $k \in s$. One can suggest that the loss function $L_1$ is used in all three cases.

1. **Nonlinear calibration.** Let us take the calibration equation

\[
\widehat{Cov}_{w}^{(1)}(y, z) = Cov(a, b).
\]

171
We call this case a **nonlinear calibration** because the calibration equation (1) is nonlinear with respect calibrated weights \( w_k \). The explicit solution of the minimization problem does not exist even in the case of loss function \( L_1 \). Only approximate iterative solution can be found.

2. **Linear calibration.** Consider the calibration equation

\[
\frac{1}{N-1} \sum_{k \in s} w_k (a_k - \mu_a) (b_k - \mu_b) = Cov(a, b), \quad \mu_a = \frac{1}{N} \sum_{k=1}^{N} a_k, \quad \mu_b = \frac{1}{N} \sum_{k=1}^{N} b_k. \tag{2}
\]

This case can be called **linear calibration**, because here we are calibrating the total of the variable \((a - \mu_a)(b - \mu_b)\). So the weights \( w_k \) can be found as calibrated weights when estimating totals (see Deville & Särndal 1992).

3. **Calibration of totals.** The weights \( w_k \) are defined using the calibration of totals:

\[
\sum_{k \in s} w_k^{(a)} a_k = \sum_{k=1}^{N} a_k = t_a, \quad \sum_{k \in s} w_k^{(b)} b_k = \sum_{k=1}^{N} b_k = t_b.
\]

This case can be motivated by the current survey practice when one weight system is being used for the estimation of all parameters needed. Simulation results show that the quality of estimates is the lowest compare to the case 1 and case 2. The equations for the calibrated weights in the case of one weighting system are presented in Plikusas and Pumputis (2007). The simulation results are also provided in this paper.

4 **Estimators of the covariance with multiple weighting system**

Consider a calibrated estimator of the following shape:

\[
\widehat{Cov}_w^{(2)}(y, z) = \frac{1}{N-1} \sum_{k \in s} w_k^{(a,b)} \left( y_k - \frac{1}{N} \sum_{k \in s} w_k^{(a)} y_k \right) \left( z_k - \frac{1}{N} \sum_{k \in s} w_k^{(b)} z_k \right).
\]

Several calibration equations may be also used in this case. Let us consider some of them.

1. One can take a nonlinear equation

\[
\widehat{Cov}_w^{(2)}(a, b) = Cov(a, b). \tag{3}
\]

This case is the most complicated analytically, the expressions for the approximate iterative solutions of the calibration equation (3) are complicated.

2. We can calibrate separately the totals \( t_a, t_b \), and the total of the variable \((a - \mu_a)(b - \mu_b)\):

\[
\sum_{k \in s} w_k^{(a)} a_k = t_a, \quad \sum_{k \in s} w_k^{(b)} b_k = t_b, \tag{4}
\]
\[
\frac{1}{N-1} \sum_{k=1}^{s} w_k^{(a,b)} (a_k - \mu_a)(b_k - \mu_b) = Cov(a, b).
\]  

(5)

The reasonable choice of the loss function in case 1 and 2 may be as follows:

\[
L(w, d) = \alpha_1 \sum_{k=1}^{s} \frac{(w_k^{(a)} - d_k)^2}{d_k q_k} + \alpha_2 \sum_{k=1}^{s} \frac{(w_k^{(b)} - d_k)^2}{d_k q_k} + \alpha_3 \sum_{k=1}^{s} \frac{(w_k^{(a,b)} - d_k)^2}{d_k q_k},
\]

here \(\alpha_1 + \alpha_2 + \alpha_3 = 1\), \(0 < \alpha_i < 1\), \(i = 1, 2, 3\).

3. We can take a double weighting system by calibrating the totals \(t_a\) and \(t_b\) using one weighting system

\[
\sum_{k=1}^{s} w_k^{(tot)} a_k = t_a, \quad \sum_{k=1}^{s} w_k^{(tot)} b_k = t_b,
\]

and adding the second weighting system by solving the equation (5).

Loss function in this case be:

\[
L' = \alpha \sum_{k=1}^{s} \frac{(w_k^{(tot)} - d_k)^2}{d_k q_k} + (1 - \alpha) \sum_{k=1}^{s} \frac{(w_k^{(a,b)} - d_k)^2}{d_k q_k}, \quad 0 < \alpha < 1.
\]

5 Concluding remarks

These remarks are based on the simulation study.

1. The nonlinear calibration leads to stable estimates in a case of one weight system. In the case of well correlated auxiliaries \(a\) and \(b\), the lowest \(MSE\) is observed in the case of a linear calibration. The case of calibration of totals leads to a highest \(MSE\) in most cases.

2. The use of several weighting systems is more efficient compare to the case of one weight system.

3. The solution of calibration equations does not exist for some samples from a skewed population. These samples are ignored in the simulation procedure.

References


