NONLINEAR CALIBRATION

[1]Aleksandras Plikusas

[1]Statistics Lithuania, Institute of Mathematics and Informatics, Lithuania
e-mail: Plikusas@ktl.mii.lt

Abstract

The definition of a calibrated estimator of the finite population parameter which may be not population total is discussed. Some estimators of the ratio of two population totals and population covariance is presented.

1 Introduction

Regression and calibrated estimators of the finite population totals are often met in the finite population statistics. These estimators are based on the use of auxiliary variables. The values of the auxiliary variables are known for all population elements. The definition and main properties of the calibrated estimator of the population total is given in the paper of Deville and Särndal (1992). The important subclass of the calibrated estimators are generalized regression estimators (GREG), which can be defined as calibrated estimator by choosing special loss function. The properties of GREG estimators of totals are considered in (Särndall, Swensson and Wretman, 1992). The estimation of the ratio of two population totals, population variance, population covariance as well as other population parameters is also topical. We will construct the calibrated (we may also call regression) estimators of the ratio of totals and the population covariance and provide the possible definition of a calibrated estimator of a more complicated parameters.

2 Calibrated estimator of total

Let us consider the finite population $\mathcal{U} = \{u_1, u_2, \dots, u_N\}$. We can also assume $\mathcal{U} = \{1, 2, \dots, N\}$. Denote the unknown population total of the variable y by

$$t_y = \sum_{k=1}^{N} y_k,$$

and Horvitz-Thompson estimator

$$\widehat{t}_y = \sum_{k\in s} \frac{y_k}{\pi_k} = \sum_{k\in s} d_k y_k.$$

Here $\pi_k = \mathbf{P}(k \in s)$, $k = 1, \dots, N$ – inclusion probability of the element $k \in \mathcal{U}$ into the sample s, $d_k = 1/\pi_k$, $k \in \mathcal{U}$ – sample design weights.

Let us suppose that for every population element k the vector of auxiliary values $\mathbf{a}_k = (a_{k1}, \dots, a_{kJ})'$ is known. It means we have J known auxiliary variables $a^{(1)}, \dots, a^{(J)}$. In official statistics the auxiliary variables may be known from the previous census, administrative data, other sources. Denote the known total

$$\mathbf{t_a} = \sum_{k=1}^{N} \mathbf{a}_k.$$

Calibrated estimator of the total t_y (Deville and Särndal 1992)

$$\widehat{t}_w = \sum_{k \in s} w_k \, y_k$$

is defined by the following conditions

a) using weights w_k the known total $\mathbf{t_a}$ is estimated without error:

$$\hat{\mathbf{t}}_\mathbf{a} = \sum_{k \in s} w_k \, \mathbf{a}_k = \mathbf{t_a};$$

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

In most practical cases the loss function

$$L = L_1(w, d) = \sum_{k \in s} \frac{(w_k - d_k)^2}{d_k q_k}$$

is being used. Here q_k, $k = 1, \dots, N$, – are free additional weights. We can also modify estimator by choosing weights q_k

Usually in survey practice we have many, say q, study variables $y^{(1)}, \dots, y^{(q)}$. The notation can be summarized in the table below

Population element	study variables	auxiliary variable(s)
$u_1 \rightarrow$	$y_1^{(1)}, \dots, y_1^{(q)}$	$\mathbf{a}_1 = (a_{11}, \dots, a_{1J})'$
$u_2 \rightarrow$	$y_2^{(1)}, \dots, y_2^{(q)}$	$\mathbf{a}_2 = (a_{21}, \dots, a_{2J})'$
$\dots$	$\dots$	$\dots$
$u_N \rightarrow$	$y_N^{(1)}, \dots, y_N^{(q)}$	$\mathbf{a}_N = (a_{N1}, \dots, a_{NJ})'$
totals	$t_y^{(i)} = \sum_{k=1}^N y_k^{(i)}$	$\mathbf{t_a} = \sum_{k=1}^N \mathbf{a}_k$

It is known that in case auxiliary variables are well correlated with study variable, the mean square error of the calibrated estimator is lower compare to the Horvitz-Thompson estimator. It can be mentioned, that the problem of the selection of auxiliary variables is not well studied. If J auxiliary variables are available one can choose from 2^J possible collections of auxiliary variables for the construction of calibrated estimators. In many practical applications the same auxiliary variables (it means the same weights) are being used for all study variables. Simulation study on the data of the Lithuanian Survey on Wages and Salaries show, that using different auxiliaries for different study variables we can reduce sampling error.

3 Calibrated estimator of the ratio

Let variables y and z be defined on $\mathcal{U}$ and take values $\{y_1, y_2, \ldots, y_N\}$ and $\{z_1, z_2, \ldots, z_N\}$, respectively. Let t_y and t_z be unknown population totals of y and z:

$$t_y = \sum_{k=1}^{N} y_k, \quad t_z = \sum_{k=1}^{N} z_k,$$

We are interested in the estimation of the ratio of two totals $R = t_y/t_z$. Suppose, the auxiliary variables a and b, having known population values $\{a_1, a_2, \ldots, a_N\}$ and $\{b_1, b_2, \ldots, b_N\}$ are available. We assign auxiliary variable a to the study variable y and and b to z. It means that a serves as auxiliary information for the study variable y and b – for the study variable z. So, we assume that the population totals

$$t_a = \sum_{k=1}^{N} a_k, \quad t_b = \sum_{k=1}^{N} b_k$$

and the ratio $R_0 = t_a/t_b$ are known.

One can take a straight estimator of the ratio R by taking the Horvitz-Thompson estimators of the totals t_y and t_z: $\widehat{R} = \widehat{t}_y/\widehat{t}_z$. Here

$$\widehat{t}_y = \sum_{k\in s} d_k y_k, \quad \widehat{t}_z = \sum_{k\in s} d_k z_k.$$

We shall construct a new estimator of the ratio R having the form

$$\widehat{R}_w = \frac{\sum_{k\in s} w_k^{(1)} y_k}{\sum_{k\in s} w_k^{(2)} z_k}. \tag{1}$$

Here the weights $w_k^{(i)}$, $i = 1, 2$, are defined under the two following conditions:

a) the weights $w_k^{(i)}$ satisfy the calibration equation

$$R_0 = \frac{\sum_{k\in s} w_k^{(1)} a_k}{\sum_{k\in s} w_k^{(2)} b_k}; \tag{2}$$

b) the weights $w_k^{(i)}$ are as close as possible to the initial design weights d_k according to the distance measure

$$L^2(w,d) = \alpha \sum_{k \in s} \frac{(w_k^{(1)} - d_k)^2}{d_k\, q_k} + (1-\alpha) \sum_{k \in s} \frac{(w_k^{(2)} - d_k)^2}{d_k\, q_k}\,. \tag{3}$$

Here q_k, $q_k > 0$, are free additional weights.

One can modify the calibrated estimator $\widehat{R}_w$ by choosing q_k or simply put $q_k = 1$ for all k.

The weights $w_k^{(i)}$, defining the estimator of the ratio $\widehat{R}_w$ can be found explicitly. Preliminary simulation results show that in some cases calibrated estimator of the ratio have lower variance than the ratio of two calibrated estimators of totals. It is not easy to compare the variances of these estimators analytically. Some special cases of the calibrated estimator of the ratio were considered by Plikusas (2003), and Krapavickaitė & Plikusas (2005).

4 Estimation of the population covariance

Suppose we are interested in the estimation of the population covariance

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

Consider the one of the standard estimators of the covariance

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

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 covariance by $Cov(a,b)$. We will construct a new calibrated estimator of the $Cov(y,z)$ using known auxiliary variables a and b. If the auxiliary variables are well correlated with the study variables, we can expect the variance of the calibrated estimator be smaller compare to the variance of estimator $\widehat{Cov}(y,z)$. The calibrated estimator

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

of the covariance $Cov(y,z)$ is defined under the following conditions:

a) the estimator $\widehat{Cov}_w$ estimates the known covariance $Cov(a,b)$ without error:

$$\widehat{Cov}_w(a,b) = \frac{1}{N-1} \sum_{k \in s} w_k \Big(a_k - \frac{1}{N}\sum_{k \in s} w_k a_k\Big)\Big(b_k - \frac{1}{N}\sum_{k \in s} w_k b_k\Big) = Cov(a,b); \tag{4}$$

b) the distance between the design weights d_k and calibrated weights w_k is minimal under the some loss function L.

It should be noted that in this case the explicit solution of the minimization problem does not exist even in the case of loss function (5) The iterative equations can be used to find the calibrated weights.

We can also use some other calibration equation instead of (4), for example,

$$\widehat{Cov}_w(a,b) = \frac{1}{N-1}\sum_{k\in s} w_k\big(a_k-\mu_a\big)\big(b_k-\mu_b\big) = Cov(a,b); \qquad (5)$$

Here

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

The case when calibration equation (5) is used can be called linear calibration, because here we are calibrating the total of the variable $(a-\mu_a)(b-\mu_b)$.

5 Some general definition of nonlinear calibration

Taking into account the examples above we will define the (nonlinear) calibrated estimator, in case the parameter of interest θ is some function of the population totals: $\theta = f(t_y^{(1)},\dots,t_y^{(q)})$. Suppose we have selected q different collections of auxiliary variables $\mathbf{a}^{(1)},\dots,\mathbf{a}^{(q)}$ that are assigned to study variables $y^{(1)},\dots,y^{(q)}$. Denote the auxiliary totals by

$$\mathbf{t}_a^{(j)} = \sum_{k=1}^{N} \mathbf{a}_k^{(j)}, \quad j=1,\dots,q$$

and write formally

$$\hat{t}_{wy}^{(j)} = \sum_{k\in s} w_k^{(j)} y_k^{(j)}, \quad \hat{\mathbf{t}}_{wa}^{(j)} = \sum_{k\in s} w_k^{(j)} \mathbf{a}_k^{(j)}, \quad j=1,\dots,q.$$

The calibrated weights $w_k^{(j)}$ can be defined by the conditions

a) for some (it may be vector valued) functions g_1 and g_2

$$g_1(\hat{\mathbf{t}}_{wa}^{(1)},\dots,\hat{\mathbf{t}}_{wa}^{(q)}) = g_2(\mathbf{t}_a^{(1)},\dots,\mathbf{t}_a^{(q)})$$

b) the weight systems $w_k^{(j)}$ are as close as possible to the design weights d_k according to some loss function L.

The calibrated estimator of $\theta = f(t_y^{(1)},\dots,t_y^{(q)})$ be $\widehat{\theta} = f(\hat{t}_{wy}^{(1)},\dots,\hat{t}_{wy}^{(q)})$. Here we can take the loss function

$$L = \sum_{j=1}^{q} \alpha_j \sum_{k\in s} \frac{(w_k^{(j)}-d_k)^2}{d_k q_k}$$

with $\alpha_j \geq 0$ and $\sum_{j=1}^{q} \alpha_j = 1$. The loss function is minimized also by α_j, $j = 1, \ldots, q$. Of course, the existence of the solution of such calibration problem is under the question. The simulation examples of calibration of covariance show that for properly chosen iterative equations and loss functions the calibrated weights exist for almost all samples.

References

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

[2] D. Krapavickaitė, A. Plikusas. Estimation of a Ratio in the Finite Population. *Informatica*, 2005, **16**(3), p. 347-364.

[3] A. Plikusas, Calibrated weights for the estimators of the ratio, *Lith. Math. J.*,**43**, 543-547 (2003).

[4] C.-E. Särndal, B. Swensson, J.Wretman, *Model Assisted Survey Sampling*, Springer-Verlag, NewYork (1992).