Nonlinear calibration

Aleksandras Plikusas

Institute of Mathematics and Informatics

Statistics Lithuania

Vilnius, Lithuania

Finite population $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$.

Unknown population **total** of y:

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

Horvitz-Thompson estimator

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

$\pi_k = \mathbf{P}(k \in s)$, $k = 1, \ldots, N$ – inclusion probability of the element $k \in \mathcal{U}$,

$d_k = 1/\pi_k$, $k \in \mathcal{U}$ – design weights.

There are many survey variables in a real survey

$y^{(1)}, \ldots, y^{(q)}$.

In practice:

the same auxiliaries are being used for all variables,

the same weights for all variables.

Example.

Lithuanian survey on wages and salaries. The estimates of different totals can be more accurate, when using different auxiliaries for different variables compare to the case all for all.

q study variables $y^{(1)}, \ldots, y^{(q)}$

J auxiliary variables

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

Problems

1. How to chose the auxiliary variables? (2^J different choices for each $y^{(i)}$)

2. How to define a regression estimator of

the ratio?

the population variance?

the population covariance?

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

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

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

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

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

Examples of distance functions

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

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

$$L_3 = \sum_{k\in s} 2\frac{(\sqrt{w_k} - \sqrt{d_k})^2}{q_k},$$

$$L_4 = \sum_{k\in s} -\frac{d_k}{q_k} \log \frac{w_k}{d_k} + \frac{1}{q_k}(w_k - d_k),$$

$$L_5 = \sum_{k\in s} \frac{(w_k - d_k)^2}{w_k q_k},$$

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

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

$q_k, \quad k = 1, \ldots, N,$ – free additional weights

$GREG$ estimator

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

$GREG$ **(calibrated) estimator of the total:**

$$\widehat{t}_w = \sum_{k \in s} w_k \, y_k = \widehat{t}_y + (\mathbf{t_a} - \widehat{\mathbf{t}}_\mathbf{a})' \left(\sum_{k \in s} d_k q_k \mathbf{a}_k \mathbf{a}_k' \right)^{-1} \sum_{k \in s} \mathbf{a}_k q_k d_k y_k$$

Example 1. If $J = 1$, $(\mathbf{a}_k = a_k)$ and $q_k = 1/a_k$, then

$\widehat{t}_w = \frac{\widehat{t}_y}{\widehat{t}_a} t_a, \quad \widehat{t}_a = \sum_{k \in s} d_k a_k, \quad t_a = \sum_{k=1}^{N} a_k,$

$\widehat{t}_w$ – ratio estimator.

Example 2. If $J = 2$, $\mathbf{a}_k = (1, a_k)'$ and $q_k = 1$, we have

$\widehat{t}_w$ – standard regression estimator.

$$\widehat{t}_w = \widehat{t}_{y\ reg} = \widehat{t}_y + \widehat{B}(t_a - \widehat{t}_a),$$

It the case of simple random sample ($\pi_k = \frac{n}{N}$):

$$\widehat{B} = \frac{\sum_{i \in s}(y_i - \bar{y})(a_i - \bar{a})}{\sum_{i \in s}(a_i - \bar{a})^2} = \frac{\widehat{s}_{ay}}{\widehat{s}_a^2}$$

Example. Estimation of the ratio of two totals.

Two study variables

$y \to \{y_1, y_2, \ldots, y_N\}$

$z \to \{z_1, z_2, \ldots, z_N\}$

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

Two auxiliaries with known values

$a \to \{a_1, a_2, \ldots, a_N\} \quad t_a = \sum_{k=1}^{N} a_k$

$b \to \{b_1, b_2, \ldots, b_N\} \quad t_b = \sum_{k=1}^{N} b_k$

Parameter

$$\theta = R = \frac{\sum_{k=1}^{N} y_k}{\sum_{k=1}^{N} z_k}$$

Estimator

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

Calibration equation (nonlinear):

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

Loss function

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

q_k – free additional weights

Proposition. *The weights* $w_k^{(1)}$, $w_k^{(2)}$, $k \in s$, *of the calibrated estimator* $\widehat{R}_w$ *which satisfy (nonlinear) calibration equation and minimize* L *are given by*

$$w_k^{(1)} = d_k\left(1 - \frac{(1-\alpha)\sum_{k\in s} d_k(a_k - R_0 b_k)}{(1-\alpha)\sum_{k\in s} d_k q_k a_k^2 + \alpha R_0^2 \sum_{k\in s} d_k q_k b_k^2}\, q_k a_k\right), \quad k \in s.$$

$$w_k^{(2)} = d_k\left(1 + \frac{\alpha\sum_{k\in s} d_k(a_k - R_0 b_k)}{(1-\alpha)\sum_{k\in s} d_k q_k a_k^2 + R_0^2\alpha \sum_{k\in s} d_k q_k b_k^2} R_0 q_k b_k\right), \quad k \in s.$$

Here α *is*

$$\alpha = \frac{\sqrt{\sum_{k\in s} d_k q_k a_k^2}}{\sqrt{\sum_{k\in s} d_k q_k R_0^2 b_k^2} + \sqrt{\sum_{k\in s} d_k q_k a_k^2}}$$

Comments.

1. The approximate variance is calculated and estimator of the variance of $\widehat{R}_w$ is constructed.

2. It is not easy to compare analytically $\widehat{R}_w$ with the estimator when regression (calibrated) estimators are used in the nominator and denominator, even in the case of simple random sample.

3. The explicit solution also exists in the case loss function

$$L_6 = \alpha \sum_{k \in s} \frac{1}{q_k} \left(\frac{w_k^{(1)}}{d_k} - 1 \right)^2 + (1 - \alpha) \sum_{k \in s} \frac{1}{q_k} \left(\frac{w_k^{(2)}}{d_k} - 1 \right)^2$$

Comparison of approximate variances

(Krapavickaitė & Plikusas (2005))

The approximate variance of the calibrated estimator of the ratio is not larger than approximate variance of the straight estimator of the ratio for SRS and $\alpha = 1/2$:

$$AVar(\widehat{R}_w) \leq AVar(\widehat{R})\,.$$

$$\widehat{R} = \frac{\sum_{k \in s}\, d_k\, y_k}{\sum_{k \in s}\, d_k\, z_k}$$

Ratio estimator of the ratio

$$\widehat{R}^{(rat)} = \frac{(\widehat{t}_y/\widehat{t}_a)\, t_a}{(\widehat{t}_z/\widehat{t}_b)\, t_b}$$

Approximate variance of the calibrated estimator of the ratio is not larger than approximate variance of the ratio estimator of the ratio for SRS and $\alpha = 1/2$:

$$AVar(\widehat{R}_w) \leq AVar(\widehat{R}^{(rat)})\,.$$

Preliminary simulation results show, that the calibrated estimator of the ratio have smaller MSE compare to the estimator when regression estimators are taken in nominator and denominator. The best gain is in the case

$$\rho(y,z) = 0.1 \quad \rho(y,a) = 0.8 \quad \rho(z,b) = 0.8 \quad \rho(a,b) = 0.1$$

Estimation of the population covariance

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

Standard estimator

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

Auxiliary variables a and b

$\{a_1, a_2, \ldots, a_N\}$

$\{b_1, b_2, \ldots, b_N\}$

Covariance between a and b: $Cov(a, b)$

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 $Cov(y, z)$ can be defined under the following conditions:

a) the weights w_k 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.

Calibration equations

Nonlinear

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

Linear

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

Here

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

It should be noted that in the case of the calibration equation (1) the explicit solution of the minimization problem does not exist even in the case of loss function L_1 The iterative equations can be used to find the calibrated weights.

The case when calibration equation (2) is used can be called linear calibration, because here we are calibrating the total of the variable $(a - \mu_a)(b - \mu_b)$.

Another possible calibrated estimator

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

Calibration equation

1)

$$\sum_{k\in s} w_k^{(a)} a_k = t_a, \quad \sum_{k\in s} w_k^{(b)} b_k = t_b$$

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

2)

$$\sum_{k\in s} w_k^{(tot)} a_k = t_a, \quad \sum_{k\in s} w_k^{(tot)} b_k = t_b$$

Loss function

$$L = \alpha_1 \sum_{k \in s} \frac{(w_k^{(a)} - d_k)^2}{d_k\, q_k} + \alpha_2 \sum_{k \in s} \frac{(w_k^{(2)} - d_k)^2}{d_k\, q_k} + \alpha_3 \sum_{k \in s} \frac{(w_k^{(a,b)} - d_k)^2}{d_k\, q_k}\,,$$

$$\alpha_1 + \alpha_2 + \alpha_3 = 1$$

$$L' = \alpha \sum_{k \in s} \frac{(w_k^{(tot)} - d_k)^2}{d_k\, q_k} + (1 - \alpha) \sum_{k \in s} \frac{(w_k^{(a,b)} - d_k)^2}{d_k\, q_k}\,,$$

Estimation of functions of totals.

q variables $y^{(1)}, \ldots, y^{(q)}$.

$$t_y^{(1)} = \sum_{k=1}^{N} y_k^{(1)}, \quad \ldots, \quad t_y^{(q)} = \sum_{k=1}^{N} y_k^{(q)}$$

$\theta = f(t_y^{(1)}, \ldots, t_y^{(q)})$ – parameter we are interested.

Simplest procedure:

$\widehat{\theta} = f(\widehat{t}_y^{(1)}, \ldots, \widehat{t}_y^{(q)})$ – estimator of θ (no auxiliaries).

Suppose there are auxiliaries $\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(q)}$ assigned to the variables $y^{(1)}, \ldots, y^{(q)}$. Calibrated estimators $\widehat{t}_{wy}^{(1)}, \ldots, \widehat{t}_{wy}^{(q)}$ of $t_y^{(1)}, \ldots, t_y^{(q)}$, and

$\widehat{\theta}_w = f(\widehat{t}_{wy}^{(1)}, \ldots, \widehat{t}_{wy}^{(q)})$ – calibrated estimator of θ

1. Use the same weights (the same collection of auxiliary variables) for all totals:

$$\widehat{t}_{wy}^{(1)} = \sum_{k \in s} w_k \, y_k^{(1)}, \quad \ldots, \quad \widehat{t}_{wy}^{(q)} = \sum_{k \in s} w_k y_k^{(q)}$$

2. Use different weights (different collections of auxiliary variables)

$$\widehat{t}_{wy}^{(1)} = \sum_{k \in s} w_k^{(1)} \, y_k^{(1)}, \quad \ldots, \quad \widehat{t}_{wy}^{(q)} = \sum_{k \in s} w_k^{(q)} y_k^{(q)}$$

Denote

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

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

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

e.g. $$\frac{\sum_{k \in s} w_k^{(1)} a_k}{\sum_{k \in s} w_k^{(2)} b_k} = \frac{\sum_{k=1}^N a_k}{\sum_{k=1}^N b_k}$$

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 $\theta = f(t_y^{(1)}, \ldots, t_y^{(q)})$ be $\widehat{\theta} = f(\widehat{t}_{wy}^{(1)}, \ldots, \widehat{t}_{wy}^{(q)})$.

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. A. Plikusas, Calibrated estimators of the ratio. *Lithuanian Math. J.*, **41** (special issue), 457-462 (2001).

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

Thank you for attention!