Design based and model based calibration

Aleksandras Plikusas

Vilnius University

Valmiera, 24-29, August, 2012

Outline

- Introduction, notation, definitions
- Estimation of total
- Estimation of ratio
- Estimation of covariance ?

Introduction, notation

Finite population:

$$\mathcal{U} = \{u_1, u_2, \ldots, u_N\} = \{1, 2, \ldots .N\}.$$

Survey variables:

$y \quad : \{y_1, y_2, \ldots, y_N\}$
$z \quad : \{z_1, z_2, \ldots, z_N\}$

Parameters of interest:

$$t = \sum_{k=1}^{N} y_k, \quad \mu_y = \frac{1}{N} \sum_{k=1}^{N} y_k, \quad \mu_z = \frac{1}{N} \sum_{k=1}^{N} z_k,$$

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

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

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.

The problem

Known auxiliary variables:

$$a^{(1)}, \dots, a^{(J)}$$

$$u_k \rightarrow \quad \mathbf{a}_k = (a_k^{(1)}, \dots, a_k^{(J)})', \; k = 1, \dots, N$$

$$\mathbf{t_a} = \sum_{k=1}^{N} \mathbf{a}_k = \big(\sum_{k=1}^{N} a_k^{(1)}, \dots, \sum_{k=1}^{N} a_k^{(J)} \big)'$$

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

Definition 1. Estimator

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

is called calibrated if

a) it estimates the known total $\mathbf{t_a}$ without error:

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

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

$$L(w, d) = L(w_k, d_k, k \in s).$$

Model calibration approach

Model calibration for the estimation of totals is proposed by *Wu and Sitter* (2001).

Suppose that the relationship between y_i and known auxiliary a_i can be described by the linear regression model (or by some more general model)

$$y_i = \beta_0 + \beta_1 a_i + \varepsilon_{yi}$$

and

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 a_i$$

Model calibration approach

The model calibrated estimator of the total

$$\hat{t}_y^{(MC)} = \sum_{k \in s} w^{(MC)} y_k$$

is defined under the conditions

$$\sum_{k \in s} w_k^{(MC)} \hat{y}_k = \sum_{k=1}^{N} \hat{y}_k$$

$$L = \sum_{k \in s} \frac{(w_k^{(MC)} - d_k)^2}{d_k q_k} \quad \rightarrow \quad \min$$

Empirical comparison

Empirical coefficient of variation

Population No 1 (Wu & Sitter)

Estimator		a	b	a & b
HT	0.041554			
DC		0.038395	0.035283	0.033674
MC		0.038509	0.035805	0.033904

Simulation is made by A. Chaustov.

Empirical comparison

Empirical coefficient of variation

Population No 2 (Lithuanian Enterprises)

Estimator		a	b	a & b
HT	0.066378			
DC		0.044411	0.075138	0.043740
MC		0.048526	0.085014	0.049900

Empirical comparison

Empirical coefficient of variation

Population No 3 (Lithuanian Enterprises)

Estimator		a	b	a & b
HT	0.086103			
DC		0.053379	0.062127	0.048058
MC		0.056351	0.060125	0.046883

The problem

How to construct calibrated estimators when estimating some other finite population parameters?

For example:

ratio of two totals
finite population covariance
variance of the estimator of total (quadratic form)

Some possible solutions

1. In case a parameter is a function of the finite population totals, estimate totals using calibrated estimators and plug-in.

2. "Calibrate" functions of totals. Many possibilities.

Example. Calibrated estimators of the ratio with one weighting system, Plikusas (2001)

Known auxiliary variables:

$$\begin{aligned} &\text{for study variable } x: \quad a_1, a_2, \ldots, a_N \\ &\text{for study variable } y: \quad b_1, b_2, \ldots, b_N \end{aligned}$$

Totals $t_a = \sum_{k=1}^{N} a_k$ and $t_b = \sum_{k=1}^{N} b_k$ are known.
The values of study variables are known only for sampled population elements.

Example. Calibrated estimators of the ratio with one weighting system, Plikusas (2001), Krapavickatė and Plikusas (2005)

Consider calibrated estimators of the ratio of the following form

$$\widehat{R}_{w1}^{(cal)} = \frac{\sum_{k\in s} w_k y_k}{\sum_{k\in s} w_k z_k},$$

here the weights w_k
a) minimize the loss function

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

b) satisfy the calibration equation

$$\frac{\sum_{k\in s} w_k a_k}{\sum_{k\in s} w_k b_k} = \frac{\sum_{k=1}^{N} a_k}{\sum_{k=1}^{N} b_k}.$$

Calibrated estimators of the ratio with two weighting systems (Plikusas, 2003)

Consider estimators having the following shape:

$$\widehat{R}_{w2}^{(cal)} = \frac{\sum_{k\in s} w_k^{(1)} y_k}{\sum_{k\in s} w_k^{(2)} z_k},$$

Calibrated estimators of the ratio with two weighting systems (Plikusas, 2003)

Consider estimators having the following shape:

$$\widehat{R}_{w2}^{(cal)} = \frac{\sum_{k\in s} w_k^{(1)} y_k}{\sum_{k\in s} w_k^{(2)} z_k},$$

here the weights $w_k^{(1)}$ and $w_k^{(2)}$
a) minimize the loss function

$$L^* = \alpha \sum_{k\in s} \frac{(w_k^{(1)} - d_k)^2}{d_k q_k} + (1-\alpha) \sum_{k\in s} \frac{(w_k^{(2)} - d_k)^2}{d_k q_k}, \quad 0 < \alpha < 1;$$

Calibrated estimators of the ratio with two weighting systems (Plikusas, 2003)

Consider estimators having the following shape:

$$\widehat{R}_{w2}^{(cal)} = \frac{\sum_{k\in s} w_k^{(1)} y_k}{\sum_{k\in s} w_k^{(2)} z_k},$$

here the weights $w_k^{(1)}$ and $w_k^{(2)}$
a) minimize the loss function

$$L^* = \alpha \sum_{k\in s} \frac{(w_k^{(1)} - d_k)^2}{d_k q_k} + (1-\alpha) \sum_{k\in s} \frac{(w_k^{(2)} - d_k)^2}{d_k q_k}, \quad 0 < \alpha < 1;$$

b) satisfy the calibration equation

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

Model calibration approach for the ratio

We extend the method to the estimation of ratio.

Suppose that the relationship between y_i and a_i (z_i and b_i) can be described by the linear regression models

$$y_i = \beta_0 + \beta_1 a_i + \varepsilon_{yi}, \quad z_i = \gamma_0 + \gamma_1 b_i + \varepsilon_{zi},$$

Define

$$\widehat{R}_{MC}^{(cal)} = \frac{\sum_{k \in s} w_k^{(MC)} y_k}{\sum_{k \in s} w_k^{(MC)} z_k},$$

Define

$$\widehat{R}_{MC}^{(cal)} = \frac{\sum_{k\in s} w_k^{(MC)} y_k}{\sum_{k\in s} w_k^{(MC)} z_k},$$

here the weights $w_k^{(MC)}$
a) minimize the loss function

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

Define

$$\widehat{R}_{MC}^{(cal)} = \frac{\sum_{k \in s} w_k^{(MC)} y_k}{\sum_{k \in s} w_k^{(MC)} z_k},$$

here the weights $w_k^{(MC)}$
a) minimize the loss function

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

b) satisfy the calibration equation

$$\frac{\sum_{k \in s} w_k^{(MC)} \hat{y}_k}{\sum_{k \in s} w_k^{(MC)} \hat{z}_k} = \frac{\sum_{k=1}^{N} \hat{y}_k}{\sum_{k=1}^{N} \hat{z}_k},$$

here $\hat{y}_k$ and $\hat{z}_k$ are fitted values for y_k and z_k.

Comments

1. Simulation results show that the calibrated estimator with two weighting systems may be more efficient in most cases.

2. Model calibrated estimator is of the same efficiency if the relation of study variables and auxiliaries is strong.

3. If working model is not correct the design based calibration is more efficient. The same is true for the estimation of totals (or means).

4. Calibrated weights have explicit expressions

Estimation of the finite population covariance

Finite population covariance

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

Standard estimator

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

Calibrated estimators of the covariance

Auxiliary variables a and b with the population values
$a_1, a_2, \ldots, a_N$
$b_1, b_2, \ldots, b_N$

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

Calibration equations with one system of weights

I)

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

$$\widehat{\mu}_{aw} = \frac{1}{N}\sum_{k\in s} w_k a_k, \quad \widehat{\mu}_{bw} = \frac{1}{N}\sum_{k\in s} w_k b_k.$$

II)

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

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

III)

$$\sum_{k\in s} w_k a_k = \sum_{k=1}^{N} a_k, \quad \sum_{k\in s} w_k b_k = \sum_{k=1}^{N} b_k. \qquad (3)$$

Calibrated estimators with several weighting systems

Estimators of type

$$\widehat{Cov}_{mw}(y,z) = \frac{1}{N-1}\sum_{k\in s} w_k^{(1)}\Big(y_k - \frac{1}{N}\sum_{l\in s} w_l^{(2)} y_l\Big)\Big(z_k - \frac{1}{N}\sum_{l\in s} w_l^{(3)} z_l\Big).$$

Calibrated estimators with several weighting systems

Case **1.**

$$\widehat{Cov}_{mw}(a,b) = Cov(a,b). \tag{4}$$

Case **2.** The weights $w_k^{(1)}$, $w_k^{(2)}$, $w_k^{(3)}$ are defined from the equations:

$$\frac{1}{N-1}\sum_{k\in s} w_k^{(1)}(a_k-\mu_a)(b_k-\mu_b) = Cov(a,b), \tag{5}$$

$$\sum_{k\in s} w_k^{(2)} a_k = \sum_{k=1}^{N} a_k, \quad \sum_{k\in s} w_k^{(3)} b_k = \sum_{k=1}^{N} b_k. \tag{6}$$

Case **3.**
$w_k^{(1)}$ from (9).
$w_k^{(2)}$ and $w_k^{(3)}$ are derived from (6)

Calibrated estimators with several weighting systems

Case **4.** Estimator

$$\widehat{Cov}_{mw}(y,z) = \frac{1}{N-1}\sum_{k\in s} w_k^{(1)}\Big(y_k - \frac{1}{N}\sum_{l\in s} w_l^{(2)} y_l\Big)\Big(z_k - \frac{1}{N}\sum_{l\in s} w_l^{(2)} z_l\Big). \tag{7}$$

$w_k^{(1)}$ are defined from (5),
$w_k^{(2)}$ from

$$\sum_{k\in s} w_k^{(2)} a_k = \sum_{k=1}^{N} a_k, \qquad \sum_{k\in s} w_k^{(2)} b_k = \sum_{k=1}^{N} b_k. \tag{8}$$

Calibrated estimators with several weighting systems

Case **5.**
$w_k^{(1)}$ from (9)
$w_k^{(2)}$ from (10)

$$\frac{1}{N-1}\sum_{k\in s} w_k(a_k-\widehat{\mu}_{aw})(b_k-\widehat{\mu}_{bw}) = Cov(a,b), \qquad (9)$$

$$\sum_{k\in s} w_k^{(2)} a_k = \sum_{k=1}^{N} a_k, \qquad \sum_{k\in s} w_k^{(2)} b_k = \sum_{k=1}^{N} b_k. \qquad (10)$$

Case **6.**
$w_k^{(1)}$ from (5),
$w_k^{(2)}$ (9).

Estimator	RB	$Var \times 10^{-13}$	$RRMSE$	cv
$\rho(y,a) = 0.81$	$\rho(z,b) = 0.90$	$\rho(y,b) = 0.63$	$\rho(z,a) = 0.60$	
$\widehat{Cov}_{1w}^{(non)}(y,z)$	-0.049 5	2.749 3	0.093 5	0.083 5
$\widehat{Cov}_{1w}^{(tot)}(y,z)$	-0.079 6	5.313 3	0.136 0	0.119 8
$\widehat{Cov}_{1w}^{(lin)}(y,z)$	-0.006 5	2.212 9	0.071 5	0.071 6
$\widehat{Cov}_{mw}^{(1)}(y,z)$	-0.001 9	2.165 7	0.070 4	0.070 5
$\widehat{Cov}_{mw}^{(2)}(y,z)$	-0.004 9	2.119 4	0.069 8	0.070 0
$\widehat{Cov}_{mw}^{(3)}(y,z)$	-0.051 0	2.804 0	0.095 0	0.084 4
$\widehat{Cov}_{mw}^{(4)}(y,z)$	-0.004 6	2.121 1	0.069 8	0.070 0
$\widehat{Cov}_{mw}^{(5)}(y,z)$	-0.050 5	2.792 0	0.094 6	0.084 2
$\widehat{Cov}_{mw}^{(6)}(y,z)$	-0.005 0	2.107 8	0.069 6	0.069 8
$\widehat{Cov}(y,z)$	-0.073 5	10.386 1	0.170 8	0.166 5

Estimator	RB	$Var \times 10^{-13}$	$RRMSE$	cv
$\rho(y,a) = 0.21$	$\rho(z,b) = 0.90$	$\rho(y,b) = 0.63$	$\rho(z,a) = 0.15$	
$\widehat{Cov}_{1w}^{(non)}(y,z)$	-0.0635	6.7417	0.1395	0.1327
$\widehat{Cov}_{1w}^{(tot)}(y,z)$	-0.0743	5.2115	0.1321	0.1180
$\widehat{Cov}_{1w}^{(lin)}(y,z)$	-0.0858	9.4940	0.1706	0.1613
$\widehat{Cov}_{mw}^{(1)}(y,z)$	-0.0792	9.8254	0.1696	0.1629
$\widehat{Cov}_{mw}^{(2)}(y,z)$	-0.0814	9.3788	0.1676	0.1595
$\widehat{Cov}_{mw}^{(3)}(y,z)$	-0.0643	6.7424	0.1399	0.1328
$\widehat{Cov}_{mw}^{(4)}(y,z)$	-0.0784	9.2041	0.1650	0.1575
$\widehat{Cov}_{mw}^{(5)}(y,z)$	-0.0619	6.6470	0.1380	0.1315
$\widehat{Cov}_{mw}^{(6)}(y,z)$	-0.0805	9.4446	0.1677	0.1599
$\widehat{Cov}(y,z)$	-0.0738	9.7766	0.1668	0.1615

Estimator	RB	$Var \times 10^{-13}$	$RRMSE$	cv
$\rho(y,a) = 0.23$	$\rho(z,b) = 0.31$	$\rho(y,b) = 0.19$	$\rho(z,a) = 0.16$	
$\widehat{Cov}_{1w}^{(non)}(y,z)$	-0.0627	12.1333	0.1781	0.1778
$\widehat{Cov}_{1w}^{(tot)}(y,z)$	-0.0703	10.2911	0.1688	0.1651
$\widehat{Cov}_{1w}^{(lin)}(y,z)$	-0.0767	10.2916	0.1716	0.1663
$\widehat{Cov}_{mw}^{(1)}(y,z)$	-0.0764	10.2927	0.1715	0.1662
$\widehat{Cov}_{mw}^{(2)}(y,z)$	-0.0763	10.2829	0.1714	0.1661
$\widehat{Cov}_{mw}^{(3)}(y,z)$	-0.0666	11.4251	0.1749	0.1733
$\widehat{Cov}_{mw}^{(4)}(y,z)$	-0.0757	10.3007	0.1712	0.1662
$\widehat{Cov}_{mw}^{(5)}(y,z)$	-0.0660	11.4427	0.1748	0.1733
$\widehat{Cov}_{mw}^{(6)}(y,z)$	-0.0722	10.3695	0.1702	0.1661
$\widehat{Cov}(y,z)$	-0.0730	10.2602	0.1698	0.1654

Some comments

• Calibrated estimators of the covariance are more efficient provided at least one highly correlated auxiliary variable is available. Model calibrated estimators are efficient in case model is correct.

• All estimators are of the same quality in case of low correlated auxiliary variables.

• Linearized variance estimators are very approximate. Bootstrap variance estimator seems to be more precise.

• There are many different possibilities to construct the calibrated estimators of the covariance.

•

References

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

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

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

Plikusas, A. (2001) Calibrated estimators of the ratio. *Lithuanian Mathematical Journal*, (special iss.) **41**, 457-462.

A. Plikusas, D. Pumputis (2007) Calibrated estimators of the population covariance. *Acta Applicandae Mathematicae*, **97**(3) 177-187.

A. Plikusas, D. Pumputis, Estimation finite population covariance using calibration, *Nonlinear Analysis: Modelling and Control*, 15, (3), 325-340, 2010
C. Wu, R. Sitter, Model calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, **96**, 185-193.