HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

A two-level hybrid calibration technique for small area estimation

Risto Lehtonen (University of Helsinki)
Ari Veijanen (Statistics Finland)

SAE2017
Paris, 10-12 July 2017

Outline

- Some calibration methods for domain and small area estimation
- Monte Carlo experiment with real data
- Conclusions
- References

This is joint work with my colleague Dr Ari Veijanen of Statistics Finland

Some calibration methods for domain estimation

- Direct* design-based method
  1. Model-free calibration MFC

- Semi-direct* model-assisted methods
  2. Model-assisted calibration MC
  3. Single-level hybrid calibration HC1
  4. Two-level hybrid calibration HC2

(*) Direct estimator defined as in the US Federal Committee on Statistical Methodology (1993) paper and semi-direct as in Lehtonen & Veijanen (2012)

Notation

Domain totals

$$t_d = \sum_{k \in U_d} y_k, \quad U_d \subset U, \quad d = 1,...,D$$

Calibration estimators

$$\hat{t}_d = \sum_{k \in s_d} w_k y_k$$

$w_k = a_k F(\boldsymbol{\lambda}' \mathbf{z}_k)$ method-specific **calibration weight** for $k \in s_d$

$a_k = 1/\pi_k$ design weight, π_k is inclusion probability

Here $F(\boldsymbol{\lambda}' \mathbf{z}_k) = 1 + \boldsymbol{\lambda}' \mathbf{z}_k$

$\boldsymbol{\lambda}$ is a Lagrange coefficient vector

$\mathbf{z}_k$ is method-specific vector of **calibration variables**

$s_d = s \cap U_d$ domain sample (unplanned domains)

$s \subset U$ sample drawn from U with sampling design $p(s)$

Calibration weights in domain calibration

Using Lagrange multipliers we minimize:

$$\sum_{k \in s_d} \frac{\left(w_k - a_k\right)^2}{a_k} - \boldsymbol{\lambda}' \left(\sum_{k \in s_d} w_k \mathbf{z}_k - \sum_{k \in U_d} \mathbf{z}_k \right), \quad d = 1,\ldots,D$$

subject to **calibration (benchmarking) constraints**

$$\sum_{k \in s_d} w_k \mathbf{z}_k = \sum_{k \in U_d} \mathbf{z}_k,$$

where $\mathbf{z}_k$ is calibration vector

Calibrated weights are defined in:

$$w_k = a_k(1 + \boldsymbol{\lambda}' \mathbf{z}_k), \text{ where}$$

$$\boldsymbol{\lambda}' = \left(\sum_{i \in U_d} \mathbf{z}_i - \sum_{i \in s_d} a_i \mathbf{z}_i \right)' \left(\sum_{i \in s_d} a_i \mathbf{z}_i \mathbf{z}_i' \right)^{-1}, \quad d = 1,\ldots,D$$

1. Model-free calibration equations

Direct method

$$\sum_{k \in s_d} w_k \mathbf{z}_k = \sum_{k \in U_d} \mathbf{z}_k = \left(N_d, \sum_{k \in U_d} x_{1k}, ..., \sum_{k \in U_d} x_{Jk} \right)'$$

where

$\mathbf{z}_k = \mathbf{x}_k = (x_{0k}, x_{1k}, ..., x_{Jk})'$, $k \in U_d$, $d = 1, ..., D$

$\mathbf{x}_k$ is vector of auxiliary x-variable values for element $k \in U_d$

$x_{0k} = 1$ for all $k \in U_d$

NOTE:

Coherence (benchmarking) property $\sum_{k \in s_d} w_k \mathbf{x}_k = \sum_{k \in U_d} \mathbf{x}_k$,

for x-variable totals in domains U_d

2. Model-assisted calibration equation:
Semi-direct method

$$\sum_{k \in s_d} w_k \mathbf{z}_k = \sum_{k \in U_d} \mathbf{z}_k = \left(N_d, \sum_{k \in U_d} \hat{y}_k \right)'$$

where $\mathbf{z}_k = (1, \hat{y}_k)', \ k \in U_d, \ d = 1,...,D$

Fitted values $\hat{y}_k = f(\mathbf{x}_k'(\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}_d))$ with $\mathbf{x}_k = (x_{0k}, x_{1k}, ..., x_{Jk})', \ k \in U_d$

e.g. logistic mixed model for a binary y-variable

$$E_m(y_k | u_d) = \frac{\exp(\mathbf{x}_k'\boldsymbol{\beta} + u_d)}{1 + \exp(\mathbf{x}_k'\boldsymbol{\beta} + u_d)}, \ k \in U_d, \ d = 1,...,D$$

NOTE:

Coherence (benchmarking) property for x-variable totals in domains U_d is not met

3. Single-level hybrid calibration equation: Semi-direct method

$$\sum_{k \in s_d} w_k \mathbf{z}_k = \sum_{k \in U_d} \mathbf{z}_k = \left(N_d, \sum_{k \in U_d} \hat{y}_k, \sum_{k \in U_d} x_{1k}, \ldots, \sum_{k \in U_d} x_{jk} \right)'$$

where $\mathbf{z}_k = (1, \hat{y}_k, x_{1k}, \ldots, x_{jk})', \; k \in U_d, \; d = 1, \ldots, D$

$\hat{y}_k = f(\mathbf{x}_k'(\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}_d))$ with $\mathbf{x}_k = (x_{0k}, x_{j+1,k}, \ldots, x_{Jk})', \; k \in U_d$

NOTE:

Coherence (benchmarking) property for totals in domains U_d is met for x-variables selected in the MFC part

4. Two-level hybrid calibration equation

Semi-direct method

Basic idea: Two calibration equations

To be solved simultaneously for a single calibration weight variable

$$\sum_{k\in s_d} w_k \mathbf{z}_k^{(1)} = \sum_{k\in U_d} \mathbf{z}_k^{(1)} = \left(N_d, \sum_{k\in U_d} \hat{y}_k \right)' \quad \text{(MC part, lower level)}$$

$$\sum_{k\in r_d} w_k \mathbf{z}_k^{(2)} = \sum_{k\in R_d} \mathbf{z}_k^{(2)} = \left(\sum_{k\in R_d} x_{1k}, ..., \sum_{k\in R_d} x_{jk} \right)' \quad \text{(MFC part, higher level)}$$

where $\mathbf{z}_k^{(1)} = (1, \hat{y}_k)'$, $s_d = s \cap U_d$, $d = 1, ..., D$

$\mathbf{z}_k^{(2)} = (x_{1k}, ..., x_{jk})'$, $r_d = s \cap R_d$, $R_d \supset U_d$

$\hat{y}_k = f(\mathbf{x}_k'(\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}_d))$ with $\mathbf{x}_k = (x_{0k}, x_{j+1,k}, ..., x_{Jk})'$, $x_{0k} = 1$, $k \in U_d$

NOTE: MFC part and MC part are calibrated at **distinct levels**

Two-level hybrid calibration equation - 2

Redefine the domain estimator to include weights w_k

for all units in region r_d :

Extended y-variable

$y_k^* = y_k,\ k \in s_d,\ 0$ otherwise

The new domain estimator is $\hat{t}_d = \sum_{k \in r_d} w_k y_k^*$

Redefine the MC part: Extended predictions and x-variable

$\hat{y}_k^* = \hat{y}_k,\ k \in s_d,\ 0$ otherwise

$x_{0k}^* = 1,\ k \in s_d,\ 0$ otherwise

New z-vector is $\mathbf{z}_k^* = (x_{0k}^*, \hat{y}_k^*)'$

Two-level hybrid calibration equation - 3

We apply a new calibration equation for the MC part:

$$\sum_{k \in r_d} w_k \mathbf{z}_k^* = \sum_{k \in U_d} \mathbf{z}_k^{(1)} = \left(N_d, \sum_{k \in U_d} \hat{y}_k \right)'$$

Minimize function

$$\sum_{k \in r_d} \frac{(w_k - a_k)^2}{a_k} - \boldsymbol{\lambda}_1' \left(\sum_{k \in r_d} w_k \mathbf{z}_k^* - \sum_{k \in U_d} \mathbf{z}_k^{(1)} \right) - \boldsymbol{\lambda}_2' \left(\sum_{k \in r_d} w_k \mathbf{z}_k^{(2)} - \sum_{k \in R_d} \mathbf{z}_k^{(2)} \right)$$

Two-level HC estimator of domain total $t_d = \sum_{k \in U_d} y_k$ is given by

$$\hat{t}_d = \sum_{k \in r_d} w_k y_k^* = \sum_{k \in s_d} w_k y_k, \; d = 1,...,D$$

NOTE: Benchmarking property for x-variables included in $\mathbf{z}_k^{(2)}$

EXAMPLE: Poverty rate for regions

- **Design-based simulation experiment with real data**
- Research questions:
- How do MFC, MC and HC1 & HC2 compare in accuracy?
- Does the accuracy of MC change when introducing benchmarking constraints into MC by single-level HC1?
- To what extent two-level HC2 is capable to reduce the effect of instability possible still visible in the MFC part of single-level HC1, in the smallest domains?

Simulation design

Population *U* with one million elements (Statistics Finland)

Income data and indicators generated from variables describing labour force status (3 classes), age group (3 classes) and gender

Regional hierarchy (EC regional NUTS-breakdown):

NUTS4 sub-regions within NUTS3 regions

Domains of interest : 36 NUTS4 regions

Domains divided into 3 size classes by domain sample size

(Minor domains - Medium-sized domains - Major domains

Higher level regions : 7 NUTS3 regions

Simulation experiments

$K = 1000$ independent SRSWOR samples, $n = 2000$

Parameters and estimators

Target parameters : at - risk - of poverty rate in domains

$$r_d = t_d / N_d, \quad d = 1,...,36$$

where domain totals of poor people are given by

$$t_d = \sum_{k \in U_d} y_k$$

y_k is binary poverty indicator with values:

$y_k = 1$ In poverty, 0 otherwise

Calibration estimators :

$$\hat{t}_d = \sum_{k \in s_d} w_k y_k, \; d = 1,...,36$$

$$\hat{r}_d = \hat{t}_d / N_d, \quad d = 1,...,36$$

Special cases of z-vectors

Model-free calibration MFC

$$\mathbf{z}_k = \mathbf{x}_k = (x_{0k}, x_{1k}, x_{2k}, x_{3k}, x_{4k}, x_{5k})'$$

Model-assisted calibration MC

$$\mathbf{z}_k = (1, \hat{y}_k)', \quad \hat{y}_k = \exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}} + \hat{u}_d) / (1 + \exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}} + \hat{u}_d))$$

$$\mathbf{x}_k = (x_{0k}, x_{1k}, x_{2k}, x_{3k}, x_{4k}, x_{5k})'$$

Hybrid calibration HC1

$\hat{y}_k$ from logistic mixed model with $\mathbf{x}_k = (x_{0k}, x_{3k}, x_{4k}, x_{5k})'$

$$\mathbf{z}_k = (1, \hat{y}_k, x_{1k}, x_{2k})'$$

Hybrid calibration HC2

$$\mathbf{z}_k^{(1)} = (1, \hat{y}_k)'$$

$\hat{y}_k$ from logistic mixed model with $\mathbf{x}_k = (x_{0k}, x_{3k}, x_{4k}, x_{5k})'$

$$\mathbf{z}_k^{(2)} = (x_{1k}, x_{2k})'$$

Assisting models

Logistic mixed models for binary response variable y

$$E_m(y_k \mid u_d) = \frac{\exp(\mathbf{x}_k'\boldsymbol{\beta} + u_d)}{1 + \exp(\mathbf{x}_k'\boldsymbol{\beta} + u_d)}, \quad k \in U_d,\ d = 1,...,36$$

where $\mathbf{x}_k = (1, x_{1k}, ..., x_{Ak})'$ for $k \in U_d$ (auxiliary data)

u_d are domain-level random intercepts, $u_d \sim N(0, \sigma_u^2)$

Estimate $\boldsymbol{\beta}$ from sample data set (`lme4`, GLIMMIX)

Calculate estimates $\hat{u}_d$, $d = 1,...,D$ and predicted values:

$$\hat{y}_k = \frac{\exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}} + \hat{u}_d)}{1 + \exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}} + \hat{u}_d)}$$

calculated for all $k \in U_d$, $d = 1,...,36$

Quality measure of estimators

Accuracy

Relative root mean squared error RRMSE (%)

$$RRMSE(\hat{\theta}_d) = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(\hat{\theta}_d(s_i) - \theta_d)^2}\ /\ \theta_d\ ,\ d = 1,...,36$$

where

$\hat{\theta}_d(s_i)$ is estimate from sample s_i for domain d

θ_d is known parameter value in domain d

NOTE: All methods are nearly design unbiased

Table 1 Average relative root mean squared error (RRMSE) (%) of calibration estimators of poverty rate in domains over domain sample size classes

Estimator	Assisting model & calibration scheme		Expected domain sample size			All
			Minor <25	Medium 25-50	Major >50	
Direct estimators						
Model-free calibration	NUTS4	$\mathbf{z}_k = (1, x_{1k}, \ldots, x_{5k})'$	61.1	40.4	20.1	47.3
Semi-direct estimators						
Model: $E_m(y_k \mid u_d) = \exp(\mathbf{x}_k'\boldsymbol{\beta} + u_d) / (1 + \exp(\mathbf{x}_k'\boldsymbol{\beta} + u_d))$, $\mathbf{x}_k = (1, x_{1k}, x_{2k}, x_{3k}, x_{4k}, x_{5k})'$						
Model MC calibration	NUTS4	$\mathbf{z}_k = (1, \hat{y}_k)'$	54.1	37.6	19.8	43.0
Model: $E_m(y_k \mid u_d) = \exp(\mathbf{x}_k'\boldsymbol{\beta} + u_d) / (1 + \exp(\mathbf{x}_k'\boldsymbol{\beta} + u_d))$, $\mathbf{x}_k = (1, x_{3k}, x_{4k}, x_{5k})'$						
Hybrid HC1 calibration	NUTS4	$\mathbf{z}_k = (1, \hat{y}_k, x_{1k}, x_{2k})'$	58.0	39.1	20.1	45.4
Hybrid HC2 calibration	NUTS4	$\mathbf{z}_k^{(1)} = (1, \hat{y}_k)'$	54.2	38.1	20.2	43.3
	NUTS3	$\mathbf{z}_k^{(2)} = (x_{1k}, x_{2k})'$				

Conclusions

- Two-level hybrid calibration tends to outperform single-level HC in accuracy for small domains
- This may happen if the estimation of model-free part at the domain level appears unstable in the single-level HC because of small domain sample size
- Calibration to higher regional level in the MFC part of HC2 can involve better stability because of larger regional sample sizes
- In this case, the new two-level hybrid calibration method may offer a safe choice

  ---
- Hybrid calibration may also provide potentials for adjustment for unit nonresponse (further research)

References

Deville J.-C. and Särndal C.-E. (1992) Calibration estimators in survey sampling. *JASA* 87, 376-382.

Estevao V.M. and Särndal C.-E. (2004) Borrowing strength is not the best technique within a wide class of design-consistent domain estimators. *JOS* 20, 645–669.

Hidiroglou M.A. and Estevao V.M. (2016) A comparison of small area and calibration estimators via simulation. *Joint Issue of Statistics in Transition and Survey Methodology*,17, 133–154.

Lehtonen R. and Veijanen A. (2012) Small area poverty estimation by model calibration. *Journal of the Indian Society of Agricultural Statistics* 66, 125-133.

Lehtonen & Veijanen (2015) Small area estimation by calibration methods. WSC 2015 of the ISI, Rio de Janeiro.

Lehtonen R. and Veijanen A. (2016) Design-based methods to small area estimation and calibration approach. In: Pratesi M. (Ed.) *Analysis of Poverty Data by Small Area Estimation*. Chichester: Wiley.

Montanari G.E. and Ranalli M.G. (2009) Multiple and ridge model calibration. Proceedings of Workshop on Calibration and Estimation in Surveys 2009. Statistics Canada.

Särndal C.-E. (2007) The calibration approach in survey theory and practice. *SMJ* 33, 99–119.

Wu C. and Sitter R.R. (2001) A model-calibration approach to using complete auxiliary information from survey data. *JASA* 96, 185–193. (with corrigenda)

Thank you for your attention