Conditional restriction estimator for domains

Natalja Lepik, Kaja Sõstra and Imbi Traat

University of Tartu, Estonia
Statistics of Estonia

Survey sampling, sampling theory, finite population

Often a functional relationship exists between population parameters

- quarterly totals have to sum up to the yearly total
- domain totals have to sum up to the population total
- totals of smaller domains have to sum up to the totals of larger domains

Survey estimates usually do not satisfy the relationships known for the parameters:

- estimators are random
- estimators from different surveys
- estimators from the same survey, but not additive (domains)
- different estimation methods for the parameters

However, users usually want that survey estimates are consistent between themselves – satisfy the same restrictions known for parameters.

In this talk

- General Restriction estimator (Knottnerus, 2003)
- Conditional Restriction estimator

  applied for domains
- some properties
- simulation study

General Restriction (GR) estimator

- is a new estimator constructed on the bases of initial estimators
- is a vector
- restrictions are satisfied
- covariance matrix is explicitly given
- known long ago
- discovered for survey sampling purposes

Let the parameter vector

$$\theta = (\theta_1, \ldots, \theta_k)'$$

be initially unbiasedly estimated by

$$\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k)'$$

Let

$$\mathbf{V} = E(\widehat{\theta} - \theta)(\widehat{\theta} - \theta), \text{ nonsingular.}$$

Suppose, parameters have to satisfy restrictions

$$\mathbf{R}\theta = \mathbf{c},$$

where $\mathbf{R} : r \times k$ matrix of rank r and $\mathbf{c} : r \times 1$ vector of constants.

For example

$$\mathbf{R} = (1, 1, \ldots, 1), \text{ or } \mathbf{R} = (-1, 1, \ldots, 1) \text{ and } \mathbf{c} = 0$$

The GR-estimator

$$
\begin{aligned}
\widehat{\theta}_{gr} &= \widehat{\theta} + \mathbf{K}\left(\mathbf{c} - \mathbf{R}\widehat{\theta}\right) \\
\mathbf{K} &= \mathbf{V}\mathbf{R}'\left(\mathbf{R}\mathbf{V}\mathbf{R}'\right)^{-1} \\
\mathbf{V}_{gr} &\equiv \text{Cov}\left(\widehat{\theta}_{gr}\right) = \left(\mathbf{I} - \mathbf{K}\mathbf{R}\right)\mathbf{V}.
\end{aligned}
$$

Some simple properties,

$$\mathbf{R}\widehat{\theta}_{gr} = \mathbf{c},$$

$$\widehat{\theta}_{gr} = \widehat{\theta}, \text{ if } \mathbf{R}\widehat{\theta} = \mathbf{c},$$

$$E\widehat{\theta}_{gr} = \theta$$

Special features of

$$\begin{aligned}\widehat{\theta}_{gr} &= \widehat{\theta} + \mathbf{K}\left(\mathbf{c} - \mathbf{R}\widehat{\theta}\right) \\ \mathbf{K} &= \mathbf{V}\mathbf{R}'\left(\mathbf{R}\mathbf{V}\mathbf{R}'\right)^{-1}\end{aligned}$$

due to the finite population sampling

If $\mathbf{V}$ is not known, the $\widehat{\theta}_{gr}$ is not estimator.

Replacing $\mathbf{V}$ by $\hat{\mathbf{V}}$ gives

$$\widehat{\widehat{\theta}}_{gr} = \widehat{\theta} + \hat{\mathbf{V}}\mathbf{R}'\left(\mathbf{R}\hat{\mathbf{V}}\mathbf{R}'\right)^{-1}\left(c - \mathbf{R}\widehat{\theta}\right)$$

Expanding $\widehat{\widehat{\theta}}_{gr}$ into Taylor series at the point $(\theta, \mathbf{V})$ gives the following linear term:

$$
\begin{aligned}
\widehat{\widehat{\theta}}_{gr} &\approx \widehat{\widehat{\theta}}_{gr}\Big|_{(\theta,\mathbf{V})} + \left[\frac{d\widehat{\widehat{\theta}}_{gr}}{d\widehat{\mathbf{V}}}\right]'_{(\theta,\mathbf{V})} \text{vec}\left(\widehat{\mathbf{V}} - \mathbf{V}\right) + \left[\frac{d\widehat{\widehat{\theta}}_{gr}}{d\widehat{\mathbf{V}}}\right]'_{(\theta,\mathbf{V})} \text{vec}\left(\widehat{\theta} - \theta\right) \\
&= \widehat{\theta} + \mathbf{K}\left(\mathbf{c} - \mathbf{R}\widehat{\theta}\right) = \widehat{\theta}_{gr}
\end{aligned}
$$

Further properties

The $\widehat{\theta}_{gr}$ is a linear minimum variance estimator of θ, given $\widehat{\theta}$ and given the information that $\mathbf{c} - \mathbf{R}\theta = 0$.

Can be shown by using projection theory.

The $\widehat{\theta}_{gr}$ is more effective than $\widehat{\theta}$ can be simply seen from

$$\mathbf{V} - \mathbf{V}_{gr} = \mathbf{KRV} = \mathbf{VR}'(\mathbf{RVR}')^{-1}\mathbf{RV} > 0,$$

because matrix of type

$$\mathbf{AA}'$$

is positive definite if $\mathbf{A}$ is of full rank (true by assumptions here).

Further properties

If the initial $\widehat{\theta}$ is biased, $E\left(\widehat{\theta}\right) = \theta + \mathbf{B}$, then the restriction estimator $\widehat{\theta}_{gr}$ still exists (satisfies restrictions), but biased:

$$E\left(\widehat{\theta}_{gr}\right) = \theta + (\mathbf{I} - \mathbf{KR})\mathbf{B}.$$

It can be shown that

$$MSE\left(\widehat{\theta}_{gr}\right) = (\mathbf{I} - \mathbf{KR}) \cdot MSE(\widehat{\theta})$$

attains its minimum for

$$\mathbf{K} = MSE(\widehat{\theta}) \cdot \mathbf{R}' \left(\mathbf{R} \cdot MSE(\widehat{\theta}) \cdot \mathbf{R}'\right)^{-1}.$$

The bias of $\widehat{\theta}_{gr}$ satisfies the following restrictions

$$\mathbf{R}(\mathbf{I} - \mathbf{KR})\mathbf{B} = \mathbf{0}$$

Further properties

$$\begin{aligned} \widehat{\theta}_{gr} &= \widehat{\theta} + \mathbf{K}\left(\mathbf{c} - \mathbf{R}\widehat{\theta}\right) \\ \mathbf{K} &= \mathbf{V}\mathbf{R}'\left(\mathbf{R}\mathbf{V}\mathbf{R}'\right)^{-1} \\ \mathrm{Cov}\left(\widehat{\theta}_{gr}\right) &= \left(\mathbf{I} - \mathbf{K}\mathbf{R}\right)\mathbf{V}. \end{aligned}$$

Another matrix representation through covariance matrices

$$\mathbf{K} = \mathrm{Cov}(\widehat{\theta}, \mathbf{R}\widehat{\theta})\ \mathrm{Cov}^{-1}(\mathbf{R}\widehat{\theta})$$

$$\mathrm{Cov}(\widehat{\theta}_{gr}) = \mathrm{Cov}(\widehat{\theta}, \mathbf{R}\widehat{\theta})\ \ \mathrm{Cov}^{-1}(\mathbf{R}\widehat{\theta})\ \mathrm{Cov}(\widehat{\theta}, \mathbf{R}\widehat{\theta})$$

Conditional restriction estimator

In Statistical Agencies, after publishing main estimates

- a need occurs for additional estimates
- they should be consistent with the published ones
- the published ones can not be changed

Conditional restriction estimator

- find the restriction estimator so that the published numbers appear in restrictions as fixed constants

- the variance formula gives now the conditional variance (underestimates)

- find the unconditional variance!

Conditional restriction estimator for domain estimation

We have initial estimators $\widehat{\Theta}_1$ and $\widehat{\Theta}_2$ for domain parameters

We want to put restrictions without changing $\widehat{\Theta}_1$

Find $\widehat{\Theta}_2^{gr}$ so that

$$\mathbf{R}\widehat{\Theta}_2^{gr} = \widehat{\Theta}_1$$

Corresponding GR-estimator is

$$\widehat{\Theta}_2^{gr} = \widehat{\Theta}_2 + \mathbf{K}(\widehat{\Theta}_1 - \mathbf{R}\widehat{\Theta}_2),$$

where $\mathbf{K} = \mathbf{V_2R'(RV_2R')^{-1}}$.

What about Cov$(\widehat{\theta}_{gr})$ now?

Let $\mathbf{V_1} = \text{Cov}(\widehat{\Theta}_1)$, $\mathbf{V_2} = \text{Cov}(\widehat{\Theta}_2)$, $\mathbf{V_{21}} = \text{Cov}(\widehat{\Theta}_2, \widehat{\Theta}_1)$

The unconditional covariance matrix is

$$\begin{aligned} \text{Cov}(\widehat{\Theta}_2^{gr}) &= \text{Cov}(\mathbf{K}\widehat{\Theta}_1 + \mathbf{P}\widehat{\Theta}_2) = \\ &= \mathbf{PV_2} + \mathbf{KV_1K'} + \mathbf{PV_{21}K'} + (\mathbf{PV_{21}K'})', \end{aligned}$$

where $\mathbf{P} = \mathbf{I} - \mathbf{KR}$

Simulation study

Population is based on Estonian LFS survey data.

Population size: 2000 persons (1192 households)

Number of domains $D = 3$

Target variables:

- monthly salary (thousand kroons)
- higher education (binary variable)

Table 1. Population characteristics

Domain	# of persons	Total salary	Total education
1	1 019	4 999	129
2	733	4 614	209
3	248	1 396	36
Total	2 000	11 010	374

Simulations

10,000 independent SI-samples were drawn from the population.

GR-estimates of domains $\widehat{\theta}_1^{gr}$, $\widehat{\theta}_2^{gr}$, $\widehat{\theta}_3^{gr}$ were calculated

- based on ratio estimators for domains $\widehat{\theta}_1$, $\widehat{\theta}_2$, $\widehat{\theta}_3$
- using estimated population total for restriction $\widehat{\theta}_U$
- using known covariance matrix

Consistency problem

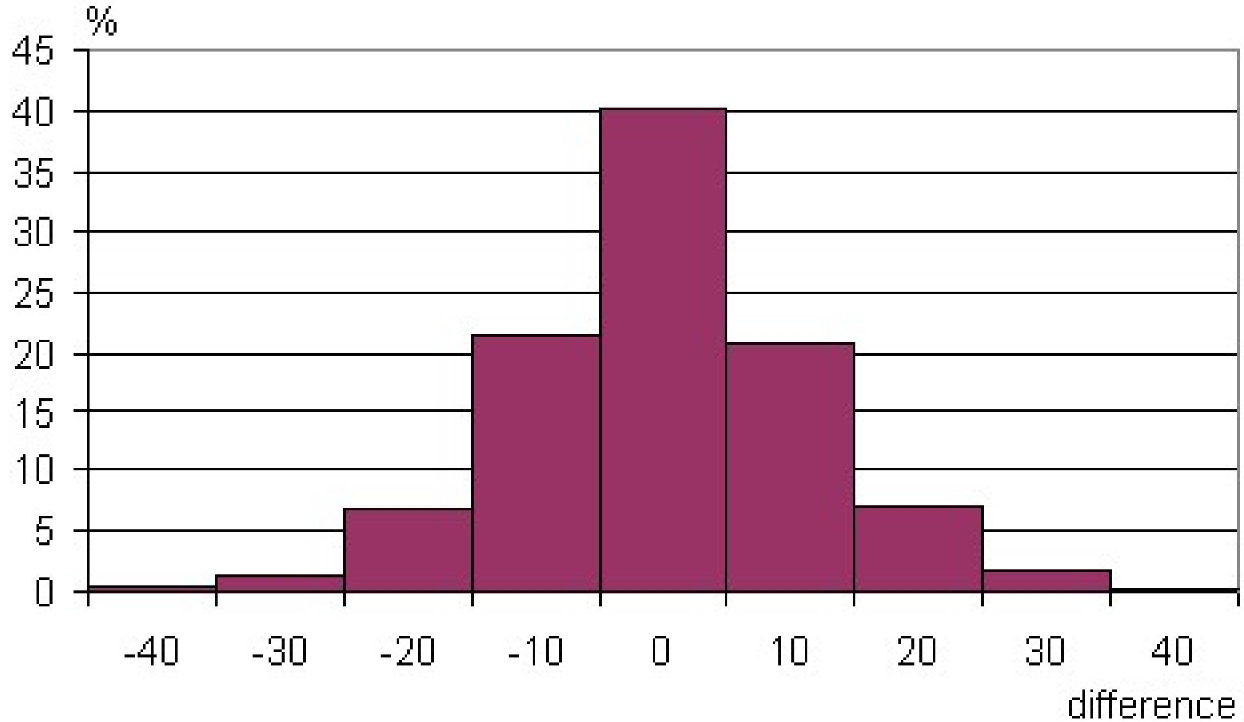

Figure 1. Difference between the sum of domain total estimates and estimated population total:
$\widehat{\theta}_1 + \widehat{\theta}_2 + \widehat{\theta}_3 - \widehat{\theta}_U$

SI-design, binary variable

Sample	Parameter	D1	D2	D3	Sum	U
1	$\widehat{\boldsymbol{\theta}}$	144	277	29	450	470
	$\widehat{\boldsymbol{\theta}}^{GR}$	152	287	31	470	470
2	$\widehat{\boldsymbol{\theta}}$	138	153	40	331	330
	$\widehat{\boldsymbol{\theta}}^{GR}$	138	152	40	330	330
3	$\widehat{\boldsymbol{\theta}}$	170	199	22	391	370
	$\widehat{\boldsymbol{\theta}}^{GR}$	162	188	20	370	370
		...	...	...	...	...
Mean	$\widehat{\boldsymbol{\theta}}$	130	209	36	375	374
	$\widehat{\boldsymbol{\theta}}^{GR}$	130	209	36	375	374
True		129	209	36	374	374

Simulation results (1)

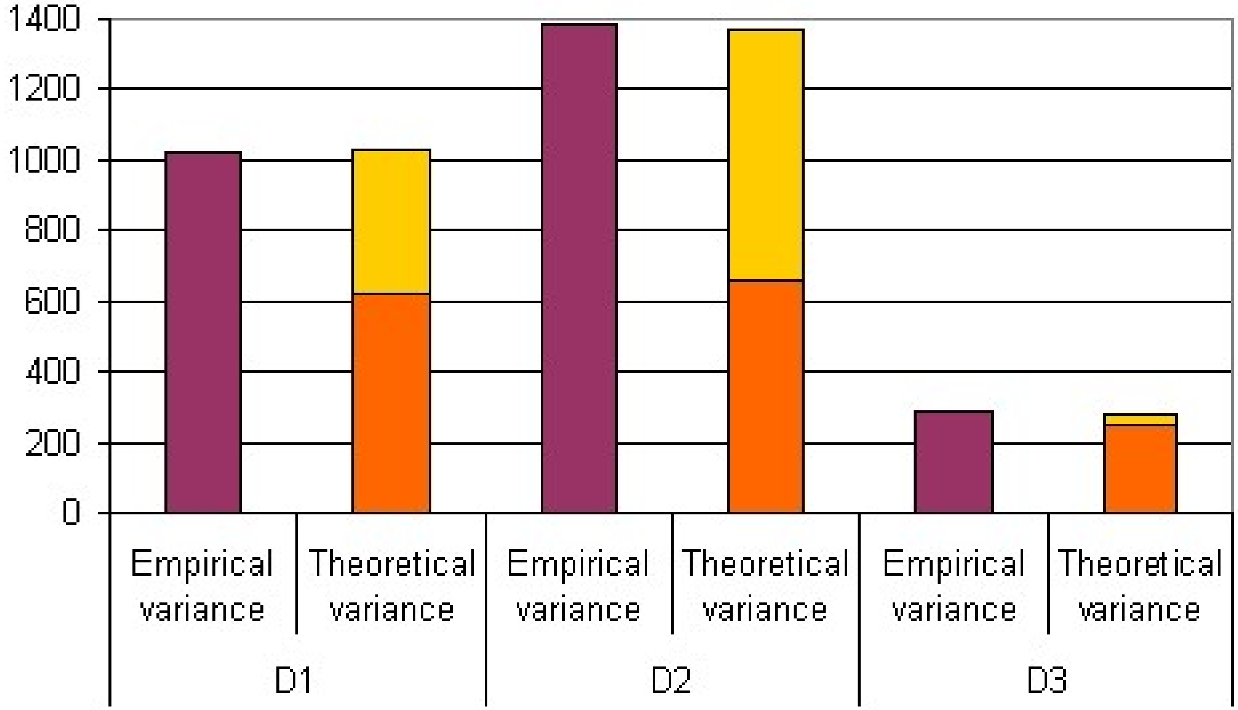

Figure 2. Empirical and theoretical variance of estimated domain total for binary variable:

$$\mathrm{Cov}(\hat{\Theta}_2^{gr}) \;= \mathbf{PV_2} + \mathbf{KV_1K'} + \mathbf{PV_{21}K'} + (\mathbf{PV_{21}K'})'$$

Simulation results (2)

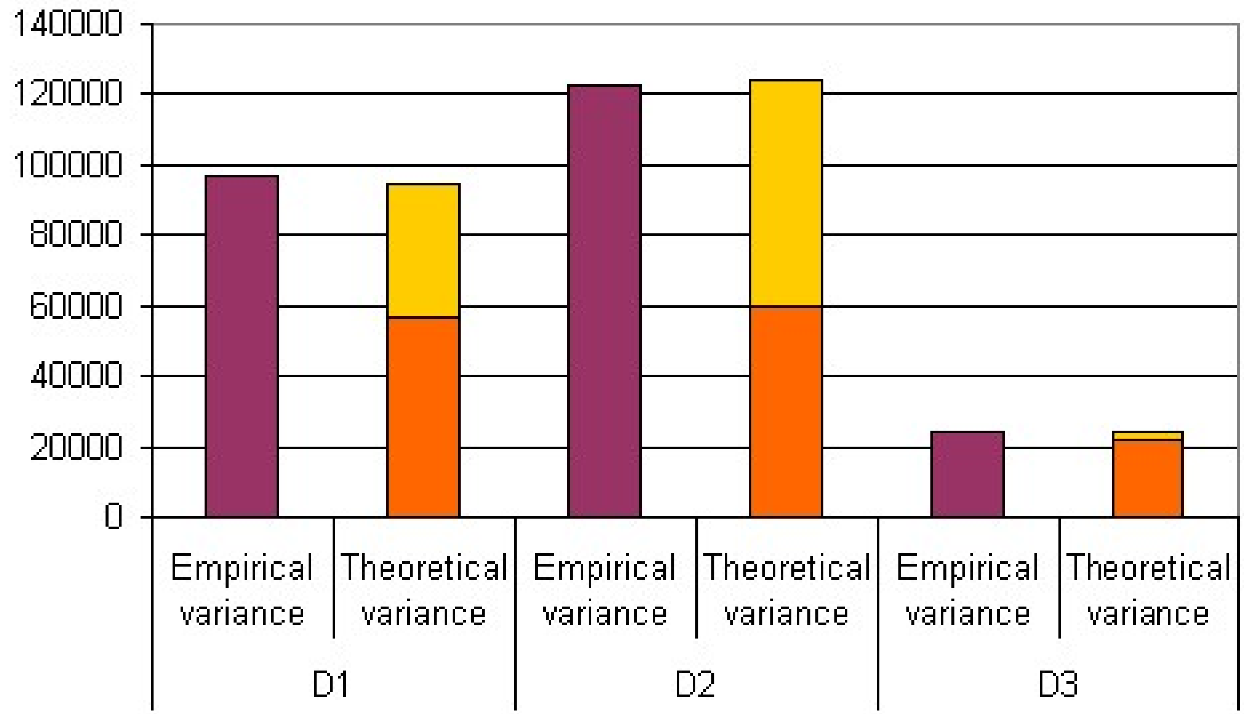

Figure 3. Empirical and theoretical variance of estimated domain total for continuous variable:

$$\mathrm{Cov}(\hat{\Theta}_2^{gr}) = \mathbf{PV_2} + \mathbf{KV_1K'} + \mathbf{PV_{21}K'} + (\mathbf{PV_{21}K'})'$$

References

Knottnerus, P. (2003). Sample Survey Theory. Some Pythagorean Perspectives. New York: Springer

Rao, C.R. (1965). Linear Statistical Inference and Its Applications. New York: Wiley

Thank You!