Estimation Under Restrictions Built Upon Biased Initial Estimators

Natalja Lepik[1]

[1]University of Tartu, e-mail: natalja.lepik@ut.ee

Abstract

The users of official statistics often require that sample-based estimates satisfy certain restrictions. In the domain's case it is required that the estimates of domain totals sum up to the population total or to its estimate. The general restriction estimator (GR) proposed by Knottnerus (2003) is described in this paper, which uses an unbiased initial estimators for its construction. Also three new estimators that satisfies the linear restriction are proposed and compared. We allow the initial estimators for them to be biased.
Keywords: Survey sampling, restriction estimator

1 Introduction

Nowadays, demand on accurate statistics of population sub-groups or domains increases. This statistics can be obtained from surveys, or, sometimes, aggregated from registers. It may happen that even if the register contains variables under interest, it does not contain identifies of the domains under our particular interest. As follows, these domain totals can not be produced from that register, they need to be estimated from a survey. The survey has to collect information on the same study variable but together with domain identifiers. As a result, the consistency problem occurs, the domain estimates from the survey do not sum up to the totals available from the registers. Analogical problem arises in the multi-survey situation, where some study variables are common in two or more surveys. Domain estimates from one survey do not sum up to the estimates of larger domains (or population totals) from another survey. Yet, there is one more situation where the consistency problem occurs. Domains themselves and the population total may be estimated by conceptually different estimators in the same survey. As a result, the domain totals do not sum up to the population total, or to the relevant larger domains.

The described inconsistency is annoying from the statistics users viewpoint. Statisticians know that the relationships between population parameters do not necessarily hold for the estimates in a sample. They also know that any auxiliary information incorporated into estimators may increase precision of these estimators. In our situation known relationships between population parameters is a kind of the auxiliary information. Involving this information into estimation process presumably improves estimates. Our goal is to define consistent domain estimators that are more accurate than the initial inconsistent domain estimators.

The problem is not new, consistency of estimators has been considered for some time. For example, if consistency is required between two surveys or between a survey and a register, some authors (Zieschang 1990, Renssen and Nieuwenbroek 1997, Traat and Särndal 2009, Dever and Valliant 2010) have proposed classical calibration approach as a solution. In this approach, the common variables are considered as additional auxiliary variables, and consistency requirement is presented in terms of calibration constraints. Other authors (Kroese and Renssen 1999, Knottnerus and Van Duin 2006) use different calibration approach for this situation, called repeated weighting. They re-calibrate the initially calibrated estimators to satisfy the consistency constraints with outside information.

Yet another approach is proposed by Knottnerus (2003). His estimator is based on the unbiased initial estimators and is unbiased itself. The advantage of the GR estimator is the variance minimizing property

in a class of linear estimators. Sõstra (2007) has developed the GR estimator for estimating domain totals under summation restriction. Optimality property of the domain GR estimator is studied in Sõstra and Traat (2009). In all these works, the unbiased or asymptotically unbiased initial estimators are assumed.

It is well known that there are many useful estimators that are biased. For example, the model-based small area estimators are design-biased. The synthetic estimator can be biased on the domain level. Even the widely used GREG estimator is only asymptotically unbiased. In this paper we will allow the vector of initial estimators $\hat{\boldsymbol{\theta}}$ to be biased, and will construct three new restriction estimators, based on the biased initial estimators.

2 Estimation under restriction

Let finite population U be divided into D non-overlapping domains U_d, $d \in \mathcal{D} = \{1, 2, ..., D\}$. We are interested in some domain parameters, for example domain totals, $t^d = \sum_{i \in U_d} y_i$ with y_i being the value of study variable for object i. It is natural that domain totals sum up to the population total, $\sum_{d=1}^{D} t^d = t = \sum_{i \in U} y_i$.

2.1 Knottnerus' approach

In general case, we denote the parameter vector under study by $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)'$, it satisfies linear restrictions:

$$\mathbf{R}\boldsymbol{\theta} = \boldsymbol{c}, \tag{1}$$

where $\mathbf{R}$ is an $r \times k$ matrix of rank r and $\boldsymbol{c}$ is the r-dimensional vector of known constants.

In a case of domain totals, where $\sum_{d=1}^{D} t^d = t$,

$$\mathbf{R} = (1, 1, ..., 1, -1)_{1 \times (D+1)},\ \boldsymbol{\theta} = (t_y^1, t_y^2, ..., t_y^D, t_y)' \text{ and } \boldsymbol{c} = 0,$$

or alternatively,

$$\mathbf{R} = (1, 1, ..., 1)_{1 \times D},\ \boldsymbol{\theta} = (t_y^1, t_y^2, ..., t_y^D)' \text{ and } \boldsymbol{c} = t_y.$$

Let $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, ..., \hat{\theta}_k)$ be the vector of estimators of $\boldsymbol{\theta}$ that do not necessarily satisfy the linear restriction (1), i.e. $\mathbf{R}\hat{\boldsymbol{\theta}} \neq \boldsymbol{c}$, in general. Knottnerus (2003, p. 328-329) proposes the following restriction estimator to solve this problem.

Assume that $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, ..., \hat{\theta}_k)'$ is unbiased for the parameter vector $\boldsymbol{\theta}$ with the variance $\mathbf{V}$, such that $\mathbf{RVR}'$ can be inverted. Then the general restriction estimator $\hat{\boldsymbol{\theta}}_{GR}$ that satisfies restrictions (1) for $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{GR}$, and the variance $\mathbf{V}_{GR}$ of this estimator are:

$$\hat{\boldsymbol{\theta}}_{GR} = \hat{\boldsymbol{\theta}} + \mathbf{K}(\boldsymbol{c} - \mathbf{R}\hat{\boldsymbol{\theta}}), \tag{2}$$

$$\mathbf{V}_{GR} = \mathbb{C}\mathrm{ov}(\hat{\boldsymbol{\theta}}_{GR}) = (\mathbb{I} - \mathbf{KR})\mathbf{V}, \tag{3}$$

where $\mathbb{I}$ is the $k \times k$ identity matrix and

$$\mathbf{K} = \mathbf{VR}'(\mathbf{RVR}')^{-1}. \tag{4}$$

Since $\mathbf{RK}$ is the identity matrix, it is easy to check that $\hat{\boldsymbol{\theta}}_{GR}$ satisfies restrictions (1):

$$\mathbf{R}\hat{\boldsymbol{\theta}}_{GR} = \mathbf{R}\hat{\boldsymbol{\theta}} + \mathbf{RK}(\boldsymbol{c} - \mathbf{R}\hat{\boldsymbol{\theta}}) = \boldsymbol{c}.$$

Knottnerus (2003, p. 332) shows that $\hat{\boldsymbol{\theta}}_{GR}$ is optimal in a class of estimators that are linear in $\hat{\boldsymbol{\theta}}$ and satisfy restrictions (1). In this class, $\hat{\boldsymbol{\theta}}_{GR}$ has minimum variance (in Löwner ordering). For example, other estimators in this class can be received by replacing $\mathbf{V}$ in the expression of $\mathbf{K}$ by any arbitrary $k \times k$ matrix $\mathbf{V}^*$, such that $\mathbf{RV}^*\mathbf{R}$ can be inverted. But the resulting estimators have bigger variance than

$\hat{\boldsymbol{\theta}}_{GR}$. In Sõstra (2007, p. 45) it is also shown that $\hat{\boldsymbol{\theta}}_{GR}$ is never less efficient than the initial estimator $\hat{\boldsymbol{\theta}}$, $\mathbf{V}_{GR} \leq \mathbf{V}$ in the sense of Löwner ordering.

Without loss of generality, we further consider linear restrictions in the form

$$\mathbf{R}\boldsymbol{\theta} = \mathbf{0}. \tag{5}$$

With $\boldsymbol{c} = \mathbf{0}$, the Knottnerus' GR estimator simplifies to the form

$$\hat{\boldsymbol{\theta}}_{GR} = (\mathbb{I} - \mathbf{KR})\hat{\boldsymbol{\theta}}. \tag{6}$$

For biased estimators the accuracy of the estimator is ordinarily measured by its mean square error. The GR-estimator (6) with biased initial estimator $\hat{\boldsymbol{\theta}}$ is not optimal any more for $\boldsymbol{\theta}$ in the sense of MSE. Although it still satisfies restrictions (5), it may have bigger mean square error than that of the initial estimator. For further details see Lepik (2011, p. 36).

In the following section we allow initial estimator to be biased, and we define three different restriction estimators for this case.

2.2 Restriction estimators handling bias

Assume that estimator $\hat{\boldsymbol{\theta}}$ is biased for $\boldsymbol{\theta}$,

$$\mathbb{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta} + \boldsymbol{b}, \tag{7}$$

where $\boldsymbol{b}$ is a vector of biases.

The first restriction estimator with biased initial estimators is defined in the following proposition.

Proposition 1. *The estimator*

$$\hat{\boldsymbol{\theta}}_{GR1} = (\mathbb{I} - \boldsymbol{KR})(\hat{\boldsymbol{\theta}} - \boldsymbol{b}), \tag{8}$$

with $\boldsymbol{K} = \boldsymbol{VR}'(\boldsymbol{RVR}')^{-1}$ *is unbiased for* $\boldsymbol{\theta}$*. Its variance is*

$$\mathbb{C}ov(\hat{\boldsymbol{\theta}}_{GR1}) = (\mathbb{I} - \boldsymbol{KR})\,\boldsymbol{V}, \tag{9}$$

and it is the optimal estimator among all linear estimators in $(\hat{\boldsymbol{\theta}} - \boldsymbol{b})$ *that satisfy restriction (5).*

For the proofs of this result and the following propositions see Lepik (2011, pp. 38-42).

Similarly to Knottnerus GR estimator our $\hat{\boldsymbol{\theta}}_{GR1}$ requires quantities that are usually unknown in practise, here the bias $\boldsymbol{b}$ and the variance $\mathbf{V}$. If $\mathbf{V}$ and $\boldsymbol{b}$ are replaced with consistent estimators, $\hat{\boldsymbol{\theta}}_{GR1}$ is consistent itself.

Below we define an estimator that is free of the knowledge of $\boldsymbol{b}$, satisfies restrictions and is more accurate than the initial estimator $\hat{\boldsymbol{\theta}}$, in $\mathbb{M}$SE terms.

Proposition 2. *The estimator, satisfying restrictions (5), but based on the mean square error* $\boldsymbol{M}$ *of the initial estimator* $\hat{\boldsymbol{\theta}}$*, is*

$$\hat{\boldsymbol{\theta}}_{GR2} = (\mathbb{I} - \boldsymbol{K}^*\boldsymbol{R})\hat{\boldsymbol{\theta}}, \tag{10}$$

where $\boldsymbol{K}^* = \boldsymbol{MR}'(\boldsymbol{RMR}')^{-1}$*. The bias of the* $\hat{\boldsymbol{\theta}}_{GR2}$ *is*

$$\boldsymbol{b}(\hat{\boldsymbol{\theta}}_{GR2}) = (\mathbb{I} - \boldsymbol{K}^*\boldsymbol{R})\boldsymbol{b}, \tag{11}$$

and the mean square error matrix is

$$\mathbb{M}SE(\hat{\boldsymbol{\theta}}_{GR2}) = (\mathbb{I} - \boldsymbol{K}^*\boldsymbol{R})\boldsymbol{M}. \tag{12}$$

Furthermore,

$$\mathbb{M}SE(\hat{\boldsymbol{\theta}}_{GR2}) \leq \boldsymbol{M} \tag{13}$$

in the sense of Löwner ordering.

The third estimator with its properties is proposed in the following proposition.

Proposition 3. *The restriction estimator*

$$\hat{\boldsymbol{\theta}}_{GR3} = (\mathbb{I} - \boldsymbol{K}^* \boldsymbol{R})(\hat{\boldsymbol{\theta}} - \boldsymbol{b}) \tag{14}$$

with $\boldsymbol{K}^* = \boldsymbol{M}\boldsymbol{R}'(\boldsymbol{R}\boldsymbol{M}\boldsymbol{R}')^{-1}$ *satisfies restrictions (5) and is unbiased for* $\hat{\boldsymbol{\theta}}$*. It's MSE is the covariance of the estimator and is equal to*

$$\mathbb{M}SE(\hat{\boldsymbol{\theta}}_{GR3}) = (\mathbb{I} - \boldsymbol{K}^* \boldsymbol{R})\, \boldsymbol{V} (\mathbb{I} - \boldsymbol{K}^* \boldsymbol{R})'. \tag{15}$$

Furthermore,

$$\mathbb{M}SE(\hat{\boldsymbol{\theta}}_{GR3}) \leq \boldsymbol{M}. \tag{16}$$

It easy to ensure that GR estimator (6) is the particular case of the estimators $\hat{\boldsymbol{\theta}}_{GR1}$, $\hat{\boldsymbol{\theta}}_{GR2}$ and $\hat{\boldsymbol{\theta}}_{GR3}$, if the vector of initial estimators has the zero bias, $\boldsymbol{b} = \mathbf{0}$. These estimators have higher accuracy than the initial estimator $\hat{\boldsymbol{\theta}}$ in a term of MSE. The next result compares the accuracy of all four estimators.

Proposition 4. *The mean square error matrices of the restriction estimators* $\hat{\boldsymbol{\theta}}_{GR1}$*,* $\hat{\boldsymbol{\theta}}_{GR2}$*,* $\hat{\boldsymbol{\theta}}_{GR3}$ *and the initial estimator* $\hat{\boldsymbol{\theta}}$ *can be ordered (in the sense of Löwner ordering) as following:*

$$\mathbb{M}SE(\hat{\boldsymbol{\theta}}_{GR1}) \leq \mathbb{M}SE(\hat{\boldsymbol{\theta}}_{GR3}) \leq \mathbb{M}SE(\hat{\boldsymbol{\theta}}_{GR2}) \leq \mathbb{M}SE(\hat{\boldsymbol{\theta}}). \tag{17}$$

2.3 Some thoughts for the future research

Estimators GR1 and GR3 requires the knowledge of the bias $\boldsymbol{b}$. In practise it is usually not known, sometimes can be estimated. The behavior of the estimators $\hat{\hat{\boldsymbol{\theta}}}_{GR1} = (\mathbb{I} - \mathbf{K}\mathbf{R})(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{b}})$ and $\hat{\hat{\boldsymbol{\theta}}}_{GR3} = (\mathbb{I} - \mathbf{K}^*\mathbf{R})(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{b}})$ is not studied yet.

Analogical situation is with the quantities $\mathbf{V}$ and $\mathbf{M}$. If to replace these quantities with their unbiased estimates, then the ordering of the MSEs of GR1, GR2 and GR3 is not necessarily hold.

References

Dever, J.A., Valliant, R.L. (2010) A comparison of variance estimators for poststratification to estimated control totals. *Survey Methodology*, 36(1), pp. 45-56.

Knottnerus, P. (2003) *Sample Survey Theory. Some Pythagorean Perspectives.* Wiley, New York

Knottnerus, P., van Duin, C. (2006). Variances in Repeated Weighting With an Application to the Dutch Labour Force Survey. *Journal of Official Statistics*, 22, pp. 565-584.

Kroese, A.H., Renssen, R.H. (1999). Weighting and Imputation at Statistics Netherland. *Proceedings of the IASS Conference on Small Area Estimation*, Riga, 109-120.

Lepik, N. (2011) *Estimation of domains under restrictions built upon generalized regression and synthetic estimators. Doctoral Dissertation.* Tartu

Renssen, R.H., Nieuwenbroek, N.J. (1997), Aligning Estimates for Common Variables in two or More Sample Surveys, *Journal of the American Statistical Association*, 92, 368-374.

Sõstra, K. (2007) *Restriction estimation for domains. Doctoral Dissertation.* Tartu

Sõstra, K., Traat, I. (2009) Optimal domain estimation under summation restriction. *Journal of Statistical Planning and Inference* vol. 139, pp. 3928-3941

Traat, I., Särndal, C.E. (2009). Domain Estimators Calibrated on Information from Other Surveys. *Research Report* No. 2009-1, Vol. 15, Department of Mathematics and Mathematical Statistics, Umea University, Sweden.

Zieschang, K.D. (1990), Sample Weighting Methods and Estimation of Totals in the Consumer Expenditure Survey. *Journal of the American Statistical Association*, 85, 986-1001.