ON DESIGN MEAN SQUARE ERROR ESTIMATION FOR MODEL-BASED SMALL AREA ESTIMATORS

A. Čiginas[1,2]

[1] Vilnius University, Lithuania
e-mail: andrius.ciginas@mif.vu.lt

[2] Statistics Lithuania, Lithuania
e-mail: andrius.ciginas@stat.gov.lt

Abstract

Estimating the means or totals in domains of a survey population with small sample sizes, indirect model-based estimators are often more efficient than direct ones. In practice, it is important to have mean square error (MSE) estimators for the former estimation derived under a design-based approach, which is typical for direct estimation applied to domains with large samples. We consider the design MSE estimation for empirical best linear unbiased predictors based on the Fay–Herriot model. In this case, unbiased MSE estimators are known as unstable in the literature. We combine them with some biased but less variable estimators of the design MSEs and show the gain in the simulation study.

Keywords: conditional mean square error, area-level model, empirical best linear unbiased predictor, composite estimation.

1 Introduction and some results

We estimate the means of a survey variable in M sampled domains (areas) of a finite population. Let $\hat{\theta}_i^{\mathrm{d}}$ be a design-unbiased estimator of the mean θ_i in the ith area with $\mathrm{E}(\hat{\theta}_i^{\mathrm{d}} \,|\, \theta_i) = \theta_i$ and the sampling variance $\mathrm{var}(\hat{\theta}_i^{\mathrm{d}} \,|\, \theta_i) = \psi_i$ is assumed to be known. This variance can be large if the domain sample size is small.

Suppose that, for each area, the auxiliary information is available as a vector $\mathbf{z}_i$ of known characteristics, which are linearly associated with unknown parameter θ_i. Then, to improve the direct estimation, famous area-level Fay–Herriot model can be used to build the best linear unbiased predictors (Rao and Molina, 2015, Section 6.1.1)

$$\tilde{\theta}_i^{\mathrm{H}} = \gamma_i \hat{\theta}_i^{\mathrm{d}} + (1-\gamma_i)\mathbf{z}_i'\tilde{\boldsymbol{\beta}} \quad \text{with} \quad \gamma_i = \sigma_v^2/(\psi_i + \sigma_v^2), \qquad i = 1, \ldots, M, \tag{1}$$

and

$$\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\psi_i, \sigma_v^2) = \left[\sum_{i=1}^{M} \mathbf{z}_i\mathbf{z}_i'/(\psi_i + \sigma_v^2)\right]^{-1} \left[\sum_{i=1}^{M} \mathbf{z}_i\hat{\theta}_i^{\mathrm{d}}/(\psi_i + \sigma_v^2)\right],$$

where σ_v^2 is the variance of random area effects, which is asumed to be known. Predictors (1) are the linear combinations of the direct estimators $\hat{\theta}_i^{\mathrm{d}}$ and the regression-synthetic estimators $\tilde{\theta}_i^{\mathrm{S}} := \mathbf{z}_i'\tilde{\boldsymbol{\beta}}$ with the weights γ_i.

Replacing σ_v^2 by an estimator $\hat{\sigma}_v^2$ in (1), we obtain empirical best linear unbiased predictors (EBLUPs) $\hat{\theta}_i^{\mathrm{H}}$ of the means θ_i, $i = 1, \ldots, M$. In practice, the design variances ψ_i are also unknown and therefore they are evaluated from external sources or by smoothing their direct estimates. Let us denote the evaluated variances by $\hat{\psi}_i^{\mathrm{s}}$.

Model MSE of EBLUPs $\hat{\theta}_i^{\mathrm{H}}$ is often used to measure the variability of the predictors. On the other hand, if the accuracy of the direct estimators is evaluated using the design MSE in domains with sufficiently large sample sizes, then it makes sense to use the same measure also for EBLUPs applied in the survey (Rao et al., 2018). However, estimation of the design (conditional) MSE

$$\mathrm{MSE}(\hat{\theta}_i^{\mathrm{H}}) = \mathrm{E}[(\hat{\theta}_i^{\mathrm{H}} - \theta_i)^2 \,|\, \theta_i] \tag{2}$$

is also a small area estimation problem because (approximately) design-unbiased estimators of (2) can be very unstable and take negative values for small sample sizes. It happens for the estimators of (2) proposed in Rivest and Belmonte (2000), Datta et al. (2011), and for elementary estimators considered in Pfeffermann and Ben-Hur (2019).

As an alternative to the unbiased estimators, one can use, according to Pfeffermann and Ben-Hur (2019), the naïve estimators

$$\mathrm{mse}_{\mathrm{n}}(\hat{\theta}_i^{\mathrm{H}}) = \hat{\gamma}_i^2 \hat{\psi}_i^{\mathrm{s}} + (1-\hat{\gamma}_i)^2 (\hat{\theta}_i^{\mathrm{H}} - \mathbf{z}_i' \tilde{\boldsymbol{\beta}}(\hat{\psi}_i^{\mathrm{s}}, \hat{\sigma}_v^2))^2, \qquad i = 1, \ldots, M, \tag{3}$$

of (2), where $\hat{\gamma}_i = \hat{\sigma}_v^2/(\hat{\psi}_i^{\mathrm{s}} + \hat{\sigma}_v^2)$. These estimators are biased but more stable and positive.

We propose another estimation of (2). We apply the results of Čiginas (2021) to design-based compositions (1) and then replace the unknown parameters by their empirical versions. First, we derive the estimator $\hat{\gamma}_i(1-\hat{\gamma}_i)\hat{\psi}_i^{\mathrm{s}}$ of the part of the squared bias of (2). Second, in line with the assumptions used in Čiginas (2021), we approximate $\mathrm{var}(\tilde{\theta}_i^{\mathrm{H}} \,|\, \theta_i) \approx \gamma_i^2 \psi_i + (1-\gamma_i)^2 \,\mathrm{var}(\tilde{\theta}_i^{\mathrm{S}} \,|\, \theta_i)$. Estimating this approximation and adding it to the estimated bias part, we arrive to

$$\mathrm{mse}_{\mathrm{b}}(\hat{\theta}_i^{\mathrm{H}}) = \hat{\gamma}_i \hat{\psi}_i^{\mathrm{s}} + (1-\hat{\gamma}_i)^2 \hat{\sigma}^2(\hat{\theta}_i^{\mathrm{S}}), \qquad i = 1, \ldots, M, \tag{4}$$

where $\hat{\sigma}^2(\hat{\theta}_i^{\mathrm{S}})$ denotes an estimator of the design variance $\mathrm{var}(\tilde{\theta}_i^{\mathrm{S}} \,|\, \theta_i)$.

The reference Rao et al. (2018) suggests linearly combine unbiased MSE estimators with biased ones like (3) using the estimated $\hat{\gamma}_i$ in the weighting. We compare some of that combinations numerically and present these results at the conference.

References

Čiginas, A. (2021) Design-based composite estimation rediscovered. `arXiv:2108.05052 [stat.ME]`.

Datta, G. S., Kubokawa, T., Molina, I., Rao, J. N. K. (2011) Estimation of mean squared error of model-based small area estimators. *Test*, **20**, 367–388.

Pfeffermann, D., Ben-Hur, D. (2019) Estimation of randomisation mean square error in small area estimation. *International Statistical Review*, **87**, 31–49.

Rao, J. N. K., Molina, I. (2015) *Small Area Estimation.* John Wiley, New Jersey.

Rao, J. N. K., Rubin-Bleuer, S., Estevao, V. M. (2018) Measuring uncertainty associated with model-based small area estimators. *Survey Methodology*, **44**, 151–166.

Rivest, L.- P., Belmonte, E. (2000) A conditional mean squared error of small area estimators. *Survey Methodology*, **26**, 67–78.