# Small area estimation in practice

Vilma Nekrašaitė – Liegė[1]

[1] Vilnius Gediminas Technical University, Statistics Lithuania, Lithuania
e-mail: nekrasaite.vilma@gmail.com

## Abstract

The study of GREG and Synthetic estimators are made in this paper. Results of simulation, in which Lithuanian private accommodation statistics database were used, are showed.

## 1 Introduction

Nowadays the statistics become very important not only for public, but and for private sector. The data are needed not only for population, but and for subgroups of the population. Such areas for which a "separate" estimate of the total (mean, etc.) of study variable $y$ is required is called a domain. If there is small sample size in domain, we called it small area (Rao, J. N. K. (2003)). Because of the smallness of sample sizes in the areas, direct survey estimates for small areas has unacceptably large standard errors. That's why alternative methods for small area estimation were developed.

Two types of estimators for domains: Generalized regression (GREG) estimators and Synthetic (SYN) estimators, will be discussed in this paper.

## 2 Notation

Let us denote by: $U = 1, 2, ..., N$ – the finite population, consisting of $N$ units; $U_d$, $d = 1, ..., D$ – domain population, consisting of $N_d$ units ($\cup_{d=1}^{D} U_d = U$, $\sum_{d=1}^{D} N_d = N$); $\mathbf{s} = 1, 2, ..., n \subset U$ – the sample consisting of $n$ units out of the $N$ units of $U$; $\mathbf{s}_d$, $d = 1, ..., D$ – domain sample, consisting of $n_d$ units ($\cup_{d=1}^{D} \mathbf{s}_d = \mathbf{s}$, $\sum_{d=1}^{D} n_d = n$); $y$ – the study variable, the values $y_k$ are known just for the elements of a sample $\mathbf{s}$; $\mathbf{x}$ – the vector of auxiliary variables, the values $\mathbf{x}_k$ are known for all units; $\pi_k$ – inclusion probability for unit $k$; $w_k = \pi_k^{-1}$ – sampling weight for unit $k$.

For the research such assumptions were made: $y_1, ..., y_N$ are realizations of independent random variables $Y_1, ..., Y_N$; $\mathbf{E}(Y_k) = \mathbf{B}^T \mathbf{x}_k$ and $\mathbf{V}(Y_k) = \sigma_k$.

## 3 Generalized regression estimator

Generalized regression (GREG) estimator, which use model as assisting tool, is a standard fixed effects regression model fitted at the population level, thus

$$\hat{t}_y = \hat{t}_{y_{HT}} + (t_{\mathbf{x}} - \hat{t}_{\mathbf{x}_{HT}})' \hat{\mathbf{B}}. \qquad (1)$$

Here

$$\hat{t}_{y_{HT}} = \sum_{k \in \mathbf{s}} w_k y_k. \tag{2}$$

$$\hat{t}_{\mathbf{x}_{HT}} = \{\sum_{k \in \mathbf{s}} w_k x_{1k}, \sum_{k \in \mathbf{s}} w_k x_{2k}, ..., \sum_{k \in \mathbf{s}} w_k x_{Jk}\}'. \tag{3}$$

$$\hat{\mathbf{B}} = \Big(\sum_{k \in \mathbf{s}} \frac{\mathbf{x}_k \mathbf{x}_k'}{\sigma_k^2 \pi_k}\Big)^{-1} \sum_{k \in \mathbf{s}} \frac{\mathbf{x}_k \mathbf{y}_k}{\sigma_k^2 \pi_k}. \tag{4}$$

To obtain point estimates for domains we use

$$\hat{t}_{d_y} = \hat{t}_{d_{y_{HT}}} + (t_{d_\mathbf{x}} - \hat{t}_{d_{\mathbf{x}_{HT}}})'\hat{\mathbf{B}}. \tag{5}$$

where $\hat{\mathbf{B}}$ satisfies (4) equation. Such form GREG estimator is called indirect GREG estimator. From the theory (Särndal, C.-E., Swensson, B., Wretman, J. (1992) and Lehtonen, R., Pahkinen, E. (2004)) we know, that GREG estimators are design unbiased, but with the large variance for small domain.

# 4 Synthetic estimator

The other type of estimators for domains is Synthetic estimators, which is design biased but has small variance for small domain (Rao, J. N. K. (2003) and Lehtonen, R., Pahkinen, E. (2004)). The idea of synthetic estimator is that small areas are similar, so we could "borrow strength" from all small areas to in order to construct an estimate for any single small area:

$$\hat{t}_{d_y} = \frac{N_d}{N} \hat{t}_{y_{HT}}. \tag{6}$$

Using such synthetic estimator we assume, that the mean of small area is the same as the mean of population.
If we have auxiliary information, then synthetic estimator can be written in such form:

$$\hat{t}_{d_y} = \sum_{U_d} \hat{\mathbf{B}} \mathbf{x}_k. \tag{7}$$

Here $\hat{\mathbf{B}}$ satisfies (4) equation.
Synthetic estimator is obviously a good choice for many situations since: with synthetic estimator we will always predict values for study variables, even for empty domains.

# 5 Simulation study

Lithuanian private accommodation statistics database was used for simulation study. Database includes 133 records for 2006 year. Every record consists of such variables: county of residence, income for 2006 year, number of guests, number of nights and etc.
1000 samples were drawn from database by simple random sampling of 80 elements. Total number of guests were estimated in every county. These estimators were used for estimation:

- Horvitz - Thompson (H-T) estimator ((2) equation);
- Synthetic (SYN1) estimator ((6) equation);

- Synthetic (SYN2) estimator ((7) equation);
- Generalized regression (GREG) estimator ((5) equation).

The last two estimators used income for 2006 year as auxiliary information, because this information we are getting from administrative sources every year for each unit.

Two measures were applied to compare the performance of the different estimators for $M = 1000$ simulation, the means absolute relative bias

$$ARB = \left| \frac{1}{M} \sum_{m=1}^{M} \frac{\hat{t}_d^{(m)} - t_d}{t_d} \right| \tag{8}$$

and the relative root means square error

$$RMSE = \frac{\sqrt{\frac{1}{M} \sum_{m=1}^{M} (\hat{t}_d^{(m)} - t_d)^2}}{t_d} \tag{9}$$

here $\hat{t}_d^{(m)}$ is the predicted value of the total number of guests from $m$-th simulation in county $d$ and the $t_d$ refers to the true population total number of guests in the same county.

In the table below the results of three counties are showed. Each of these counties belong to different domain size class.

Table 1: Simulation results

| Estimator | Domain sample size class | Domain total in population | Estimate of domain total | Absolute relative bias $ABR\%$ | Relative root MSE $RMSE\%$ |
|---|---|---|---|---|---|
| H-T | $0 - 10$ | $1,132$ | $1,152$ | 1.5 | 25.0 |
| | $11 - 20$ | $8,341$ | $8,213$ | 1.2 | 18.3 |
| | $> 20$ | $6,101$ | $6,176$ | 1.7 | 41.7 |
| SYN1 | $0 - 10$ | $1,132$ | $1,389$ | 52.8 | 53.0 |
| | $11 - 20$ | $8,341$ | $3,933$ | 32.7 | 35.2 |
| | $> 20$ | $6,101$ | $8,098$ | 22.7 | 25.8 |
| SYN2 | $0 - 10$ | $1,132$ | 754 | 26.2 | 26.9 |
| | $11 - 20$ | $8,341$ | $6,154$ | 28.6 | 30.3 |
| | $> 20$ | $6,101$ | $7,843$ | 33.3 | 34.0 |
| GREG | $0 - 10$ | $1,132$ | $1,133$ | 0.3 | 16.4 |
| | $11 - 20$ | $8,341$ | $8,320$ | 0.1 | 12.8 |
| | $> 20$ | $6,101$ | $6,094$ | 0.2 | 16.9 |

From the table we can see, that for synthetic estimators $ARB$ is large, but variance is small (relative root means square error is almost the same as absolute relative bias). For Horvitz - Thompson and GREG estimator the $ARB$ is small, but $RMSE$ much bigger then $ARB$.

Such simulation showed, that GREG estimator is most efficient estimator for Lithuanian private accommodation statistics even if sample size in domain is very small.

## References

Lehtonen, R., Pahkinen, E. (2004) *Practical Methods for Design and Analysis of Complex Surveys.* Wiley, Chichester.

Rao, J. N. K. (2003) *Small Area Estimation.* Wiley, New York.

Särndal, C.-E., Swensson, B., Wretman, J. (1992) *Model Assisted Survey Sampling.* Springer - Verlag, New York.