Using auxiliary information for domain estimation

Karolin Toompere[1]

[1] University of Tartu, Estonia
e-mail: karolin.toompere@ut.ee

Abstract

In this paper two estimators for subgroups of population or domains are studied: the GREG estimator and the synthetic estimator. While the GREG estimator produces nearly unbiased estimates for domains, it has a larger variability than the synthetic estimator. The synthetic estimator is more stable but it gives biased estimates. A simulation study is used to illustrate this problem.

1 Introduction

Usually in the surveys the estimates are found not only for the population but also for the subgroups or domains. Often there is enough data to find the population level estimates but to find good estimates in the domains we would need more observations. Therefore it is important to use the best methods available to get good estimates in spite of the small number of observations. One method that uses auxiliary information and gives nearly unbiased estimates at the domain level is the generalized regression (GREG) estimator. GREG estimator is favoured in practice because of its good properties. Since it has quite large sampling variability and to ensure stable estimates a large sample is needed, sometimes synthetic estimator is used instead. Synthetic estimator is the sum of predicted values in the domain and it uses values of study variable outside the domain. Synthetic estimator has a small variance but it can be biased.

2 Notation

Let the population $U = 1, \dots, i, \dots, N$ be divided into D domains, $U_1, \dots, U_d, \dots, U_D$ and let N_d be the size of domain U_d.

$$U = \cup_{d=1}^{D} U_d$$

and

$$N = \sum_{d=1}^{D} N_d$$

We want to estimate the domain total

$$t_d = \sum_{U_d} y_i \qquad d = 1, \ldots, D,$$

y_i denotes the study variable for observation i.

A sample s with sample size n is drawn from population U. The part of the sample that falls into U_d is denoted by s_d and its size by n_d.

We also have auxiliary information available. Let x_i be the auxiliary vector for observation i.

3 Regression estimator

If we have auxiliary information then we can use it to get in domains estimates with better quality. One possibility is to use the GREG estimator. The regression model is fitted describing the relationship between the study variable and the auxiliary variables. The fitted values are used to find the estimates.

GREG estimator for domains is given by the formula:

$$\hat{t}_{dr} = \sum_{U_d} \hat{y}_i + \sum_{s_d} w_i e_i, \tag{1}$$

where

$$\hat{y}_i = x_i^T \hat{B}$$

are predicted values,

$$e_i = y_i - \hat{y}_i$$

are residuals,

$$\hat{B} = \sum_s \frac{w_i x_i x_i^T}{\sigma_i^2} \sum_s \frac{w_i x_i y_i}{\sigma_i^2}$$

is regression coefficient vector and w_i are sampling weights

The formula (1) consists of sample fitted y-values and the adjusting term containing residuals. To calculate the predicion term

$$\sum_{U_d} \hat{y}_i = (\sum_{U_d} x_i)' \hat{B}$$

we need to know the sums of the auxiliary variables in the domains $\sum_{U_d} \mathbf{x}_i$.

4 Synthetic estimator

The first term of the regression estimator:

$$\hat{t}_{dsy} = \sum_{s_d} \hat{y_k} = (\sum_{U_d} \mathbf{x_k})'\hat{\mathbf{B}}$$

can be also considered as an estimator for the domain totalt_d. It is the sum of predictions in the domain and it is called the synthetic estimator. If the adjusting term of regression estimator (1) is nonzero, the synthetic estimator is biased but its variance is small and it is sometimes used to estimate in very small domains. The predicted values of the objects are calculated using the sample outside the s_d.

5 Simulation study

In the simulation study the StatVillage (Schwartz,1997) data was used. It is based on a real data of the census of Canada. In the StatVillage there are 1024 households that are divided into 128 blocks, in every block there are 8 houses. 36 different variables are measured for each household.

I looked three different domains with different sizes. In first domain there are northern households, in the second the middle ones and in the third domain the southern households. The sizes of the households are given in the Table 1.

A simple random sample with sample size 100 was drawn R=1000 times. The aim was to estimate the total income in the domains and the total number of rooms in the domains. The auxiliary information included in addition to the sizes of the domains the number of people in the households. For each domain the synthetic estimates of the domain totals were found. Also two GREG estimates were found - first one, where

Table 1: Domains used in the simulation study

domain	**blocks**	**nr of blocks**	**nr of households**
domain1	blocks 1-13	13	104
domain2	blocks 14-51	38	304
domain3	blocks 52-128	77	616

the auxiliary information was at the population level (GREG1) and the second one, where we also knew the auxiliary information at the domain level (GREG2).

The absolute relative biases (2) and relative root mean square errors (3) were calculated, also means (4), standard deviations (5) and coefficients of variation (6).

$$ARB = |\frac{1}{R}\sum_{r=1}^{R}\frac{\hat{t_d} - t_d}{t_d}| \tag{2}$$

$$RRMSE = \frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{t}_d - t_d)^2}}{t_d} \tag{3}$$

$$MEAN = \frac{1}{R}\sum_{r=1}^{R}\hat{t}_d \tag{4}$$

$$STD = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(\hat{t}_d - \frac{1}{R}\sum_{r=1}^{R}\hat{t}_d)^2} \tag{5}$$

$$CV = \frac{STD}{MEAN} \tag{6}$$

Table 2: The results of the simulation. Estimating the total income in the domains

	MEAN	STD	CV	ARB	RRMSE
Synthetic					
domain1	6 830 327	487 892	0.071	0.509	0.511
domain2	19 017 508	1 247 292	0.066	0.177	0.185
domain3	32 421 448	1 902 749	0.059	0.528	0.535
GREG1					
domain1	13 894 245	4 051 286	0.292	0.004	0.292
domain2	23 220 025	3 370 185	0.145	0.000005	0.146
domain3	21 155 013	2 113 703	0.099	0.0007	0.099
GREG2					
domain1	13 995 006	1 825 704	0.130	0.011	0.132
domain2	23 065 072	744 768	0.032	0.001	0.032
domain3	21 251 922	1 329 510	0.063	0.0006	0.063

The results are given in Tables 2 and 3. It is seen that like expected the bias of synthetic estimates are larger than the bias of the GREG estimates. The GREG using auxiliary information at the population level and at the domain level had both small biases. For the smallest domain the synthetic estimator gave estimates with considerably smaller variance than others. When estimating the total number of rooms, synthetic estimator gave less variable results than the GREG2. When estimating the total income the GREG at domain level gave less variable results at larger domains than the synthetic estimator. The relative root mean square errors also show that synthetic estimator is in general less variable.

Table 3: The results of the simulation. Estimating the total number of rooms in the domains

	MEAN	STD	CV	ARB	RRMSE
Synthetic					
domain1	769	23	0.029	0.152	0.154
domain2	2183	62	0.028	0.051	0.057
domain3	4013	123	0.031	0.067	0.074
GREG1					
domain1	900	255	0.283	0.006	0.281
domain2	2301	338	0.147	0.002	0.147
domain3	3763	330	0.087	0.0006	0.088
GREG2					
domain1	917	105	0.115	0.012	0.117
domain2	2300	98	0.043	0.001	0.088
domain3	3762	153	0.041	0.0002	0.041

References

Schwartz,C.-J. (1997) StatVillage: An On-Line, WWW-Accessible, Hypothetical City Based on Real Data for Use in an Introductory Class in Survey Sampling. *Journal of Statistics Education* **v.5, n.2.**

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