Liet. mat. rink. LMD darbai, **50**, 2009, 304–309

Small area estimation in the case of nonresponse

Vilma NEKRAŠAITĖ-LIEGĖ
Vilnius Gedimino Technical University
Saulėtekio 11, LT-10223 Vilnius, Lithuania
e-mail: nekrasaite.vilma@gmail.com

Abstract. In this paper the effect of model and nonresponse adjustment on different types of estimators for the totals of small area domains is examined. The empirical results are based on Monte Carlo simulations with repeated samples drawn from a finite population constructed from the real data from the Lithuanian Business Survey.

Keywords: population, sample, small area, nonresponse, response probability, estimator, model.

Introduction

Small area estimation becomes more and more important in recent years. The main reason of this, is the demand of reliable small area statistics. To choose the right model is very important, when small area estimators are used [3,4]. Traditional sampling theory assumes complete measurement for all sampling units [7], but in true surveys we are faced with nonresponse. Some methods have been developed for correcting the consequences of unit nonresponse for the total of population [1,7], but they were never used in small area estimation. In this paper both the choice of model and nonresponse adjustment for small area estimation is examined.

1. Notations

Let us denote a finite population $U = 1, 2, \ldots, N$, consisting of N units. This population can be divided into D nonoverlaping domains U_d, $d = 1, \ldots, D$, consisting of N_d units. A sample $\mathbf{s}$ consisting of n units is selected from the population U, $\mathbf{s} = 1, 2, \ldots, n \subset U$. This sample also can be divided into D nonoverlaping sample domains $\mathbf{s}_d$, $d = 1, \ldots, D$, consisting of n_d units. Let I_d be an indicator for the domain membership such that $I_d = 1$ if unit k from the population U is in d domain, $k \in U_d$ and $I_d = 0$ in other cases. Each unit k has an inclusion probability $\pi_k = \mathbf{P}(k \in \mathbf{s})$ or sampling weight $w_k = \pi_k^{-1}$. For different reasons there are missing units in the sample s. Let us denote the responding set by r. Instead of original sample size n we receive complete responses for $n^{(r)}$. Let a response probability for each unit be $\kappa_k = \mathbf{P}(k \in \mathbf{s}^{(r)},\ k \in \mathbf{s})$, where $\mathbf{s}^{(r)}$ is a responded sample.

Let y be a study variable, which values y_k are known just for the elements of a sample $\mathbf{s}$ and $\mathbf{x}$ be a vector of auxiliary variables, which values $\mathbf{x}_k$ are known for all units in U. Let $t_d = \sum_{k \in U_d} y_k$ be a domain total – parameter of interest.

2. Models and estimators

2.1. Models

Let say that $y_1, \dots, y_N$ are realizations of independent random variables $Y_1, \dots, Y_N$, $Y_k = \mathbf{x}'_k\beta + \varepsilon_k$, $k = 1, \dots, N$. Here $\mathbf{x}_k$ is the vector of auxiliary variables, β is the vector of regression coefficients and ε_k are residuals with variances $\mathbf{V}(\varepsilon_k) = \sigma_k^2$. Three special cases of this model is used:

1) *common model.* Here $\mathbf{x}_k = (1, x_{1k}, x_{2k}, \dots, x_{Jk})'$ and $\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_J)$. The model has common intercepts and common slopes for all domains.
2) *model with domain-intercepts.* Here $\mathbf{x}_k = (I_{k1}, \dots, I_{kd}, \dots, I_{kD}, x_{1k}, x_{2k}, \dots, x_{Jk})'$ and $\beta = (\beta_{01}, \dots, \beta_{0d}, \dots, \beta_{0D}, \beta_1, \beta_2, \dots, \beta_J)$. The model has separate intercepts and common slopes for all domains.
3) *mixed model with random intercepts.* Here $\mathbf{x}_k = (1, x_{1k}, x_{2k}, \dots, x_{Jk})'$ and $\beta = (\beta_0 + \mathbf{u}_d, \beta_1, \beta_2, \dots, \beta_J)$. Here $\mathbf{u}_d$ is a vector of random effects defined at the domain level. So the model has random effects for intercepts and common slopes for all domains.

The estimate of β for common model and model with domain-intercepts is calculated using generalized weighted least squares (GWLS) method and for the mixed model with random intercepts – restricted maximum likelihood (REML) method where weights are incorporated.

2.2. Estimators

The parameter of interest is domain total. There is made an assumption, that number of elements in each domain (N_d) is known. There are two types of domains – planned and unplanned. For planned domains [8] the sample size n_d in domain sample $s_d \in U_d$ is fixed in advance. In this research domains are unplanned, that means the sample size n_d in domain sample $s_d \in U_d$ is random.

It is possible to group all estimators in to three groups:

1) *design-based estimators.* Design-based estimators use information about the sampling design by means of sampling weights. A statistical model here is used as an assisting tool to incorporate auxiliary information into the estimation procedure. Such estimators are Horvitz–Thompson (HT) estimator $\hat{t}_d = \sum_{k \in \mathbf{s_d}} w_k y_k$ [2,5] and generalized regression (GREG) estimator $\hat{t}_d = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in \mathbf{s}_d} w_k (y_k - \hat{y}_k)$ [7]. Here $\hat{y}_k$ is fitted value calculated by a given model ($\hat{y}_k = \mathbf{x}'_k\beta$).
2) *model-dependent estimators.* Model-dependent estimators use fitted values $\hat{y}_k$, where $\hat{\beta}$ is calculated using unweighed methods, for example REML. Such estimators are synthetic (SYN) estimator $\hat{t}_d = \sum_{k \in U_d} \hat{y}_k$ [6] and empirical best linear unbiased prediction (EBLUP) estimator $\hat{t}_d = \sum_{k \in U_d \setminus \mathbf{s}_d} \hat{y}_k + \sum_{k \in \mathbf{s}_d} y_k$ [6].
3) *model-based estimators.* Model-based estimators use fitted values $\hat{y}_k$, where $\hat{\beta}$ is calculated using weighed methods, for example GWLS. Such estimators are pseudo-synthetic (SYN-P) estimator $\hat{t}_d = \sum_{k \in U_d} \hat{y}_k$ [6] and pseudo-EBLUP (EBLUP-P) estimator $\hat{t}_d = \sum_{k \in U_d \setminus \mathbf{s}_d} \hat{y}_k + \sum_{k \in \mathbf{s}_d} y_k$ [6].

3. Adjustment based on nonresponse

The most common method of handling nonresponse is reweighting, where original inclusion probabilities π_k are deflated by the response probabilities κ_k and new sampling weight $w_k = (\pi_k \kappa_k)^{-1}$ is obtained.

The original response probability is never known in practice, so there are several methods to estimate it. One of them is called weighting-class, where the sample is divided into homogenous and mutually exclusive groups g, $g = 1, \dots, G$ with the same response probability for the unit in the same group:

$$\hat{\kappa}_k = \hat{\kappa}_g = \frac{\sum_{j \in \mathbf{s}_g^{(r)}} w_j}{\sum_{j \in \mathbf{s}_g} w_j} = \frac{\hat{N}_g^{(r)}}{\hat{N}_g}. \tag{1}$$

The other method of estimating response probability is to use logistic regression model for explaining the response mechanism [1]:

$$\hat{\kappa}_k = \frac{\exp\{\hat{\theta}_k\}}{1 + \exp\{\hat{\theta}_k\}} = \frac{\exp\{\hat{B}\mathbf{x}_k\}}{1 + \exp\{\hat{B}\mathbf{x}_k\}}. \tag{2}$$

4. Simulation study

4.1. Population

Lithuanian business statistics data is used for simulation study. The population includes 6110 enterprisers, which filled questionnaire in 2008. Every record consists of such variables: region of residence, income for the first quarter of 2008 year, number of employees in the same quarter, classification of enterpriser (NACE code).

For the study variable income is chosen. Let us notate this variable as y_k for the kth enterpriser, $k = 1, \dots, N = 6110$. The parameter of interest is total income in each region (domain total – t_d). There are 17 regions of interest. To improve quality of estimations two auxiliary variables were used: number of employees (x_1) and indicator of the NACE code (x_{2j}, $j = 1, \dots, 6$).

1000 independent samples of 250 elements are drawn from the population by simple random sampling without replacement (SRS). Several type of estimators are calculated using different type of models (see Table 1).

Each estimator is calculated three times using different response probability estimators. In the first case it is saying that there is no nonresponse ($\kappa_k = \kappa = 1$). In other cases 85% response rate is generated using survey's properties (response rate depends

Table 1. Models and estimators

Model	Estimators	Notations
Common model	GREG, SYN-P	C-GREG, C-SYN
Model with domain-intercepts	GREG, SYN-P	D-GREG, D-SYN
Mixed model with random intercepts	GREG, SYN-P, EBLUB-P	M-GREG, M-SYN, M-EBLUP

Table 2. Simulation results, when there is no nonresponse

Estimator	Domain sample size class			
	Minor $0-7$		Medium $8-15$	
	$MABR, \%$	$MRRMSE, \%$	$MABR, \%$	$MRRMSE, \%$
H-T	0.9	77.8	0.3	50.6
C-GREG	0.9	27.5	0.8	16.4
C-SYN	37.4	37.8	25.4	25.8
D-GREG	0.8	21.4	0.2	13.9
D-SYN	6.0	9.7	2.1	6.9
M-GREG	1.7	21.7	0.8	13.7
M-SYN	6.1	10.6	1.9	7.5
M-EBLUP	5.8	10.4	1.8	7.5

Table 3. Results, when weighted-class method is used for κ_k estimation

Estimator	Domain sample size class			
	Minor $0-7$		Medium $8-15$	
	$MABR, \%$	$MRRMSE, \%$	$MABR, \%$	$MRRMSE, \%$
H-T	5.4	78.9	1.3	52.1
C-GREG	1.7	29.1	1.7	17.2
C-SYN	37.6	38.0	25.7	26.2
D-GREG	2.1	22.6	0.9	14.7
D-SYN	6.0	10.1	2.1	7.2
M-GREG	2.0	23.0	1.3	14.8
M-SYN	6.3	10.7	2.0	7.5
M-EBLUP	6.1	10.5	2.0	7.3

Table 4. Results, when logistic regression model is used for κ_k estimation

Estimator	Domain sample size class			
	Minor $0-7$		Medium $8-15$	
	$MABR, \%$	$MRRMSE, \%$	$MABR, \%$	$MRRMSE, \%$
H-T	5.7	77.8	0.9	51.4
C-GREG	1.9	28.9	1.7	19.4
C-SYN	37.7	38.1	25.7	26.2
D-GREG	2.0	22.5	1.2	15.9
D-SYN	6.0	10.1	2.0	7.1
M-GREG	2.1	22.9	1.6	14.7
M-SYN	6.4	10.8	2.1	7.5
M-EBLUP	6.2	10.6	3.5	7.9

on region, number of employees and classification of enterpriser). Both weighting-class method (see Eq. (1)) and logistic regression model Eq. (2)) are used to estimated response probability. Using weighting-class method units are grouped by region, and for the logistic regression model auxiliary vector $\mathbf{x}_k = (1, x_1, x_{21}, x_{22}, x_{25})'$ is used.

4.2. Accuracy measures

Two accuracy measures are applied to compare the performance of the different estimators for $M = 1000$ simulation: the absolute relative bias

$$ARB(\hat{t}_d) = \frac{\left|\frac{1}{M}\sum_{m=1}^{M}\hat{t}_d^{(m)} - t_d\right|}{t_d} \tag{3}$$

and the relative root means square error

$$RRMSE(\hat{t}_d) = \frac{\sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{t}_d^{(m)} - t_d)^2}}{t_d}. \tag{4}$$

Here $\hat{t}_d^{(m)}$ is the predicted value of the total from mth simulation in region d and the t_d refers to the true population in the same region.

There are 17 regions of interest, so for the better comparison these regions are grouped into three domain sample size class by the average number of elements in sample domain (minor $0 - 7$, medium $8 - 15$, major > 15). A mean of absolute relative bias (MARB) and a mean of relative root means square error (MRRMSE) in each class are calculated.

4.3. Simulation results

The results are presented just for two domain sample size classes, because the results in small domain are more interesting then in large one (see Tables 2–4).

5. Conclusions

As you can see from the tables both methods of nonresponse adjustment can be use. Model with domain-intercepts and mixed model with random intercepts are better choise, then common model. Model-based estimators are biased but their MRRMSE is smaller than that of design-based estimators, so what estimator to choose depends on what result you want – unbiased (GREG) or with smaller mean square error (SYN-P or EBLUP-P).

References

1. A. Ekholm, S. Laaksonen. Weighting via response modeling in the Finnish household budget survey. *Journal of Official Statistics*, 7(3):325–337, 1991.
2. D.G. Horvitz, D.J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685, 1952.
3. R. Lehtonen, C.-E. Särndal, A. Veijanen. The effect of model choice in estimation for domains, including small domains. *Survey Methodology*, 29:33–44, 2003.

4. R. Lehtonen, C.-E. Särndal, A. Veijanen. Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition*, 7:649–673, 2005.
5. R.D. Narain. On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3:169–174, 1951.
6. J.N.K. Rao. *Small Area Estimation*. Wiley, New York, 2003.
7. C.-E. Särndal, B. Swensson, J. Wretman. *Model Assisted Survey Sampling*. Springer-Verlag, New York, 1992.
8. M.P. Singh, J. Gambino, H.J. Mantel. Issues and strategies for small area data. *Survey Methodology*, 20:3–14, 1994.

REZIUMĖ

V. Nekrašaitė-Liegė. Mažų sričių vertinimas neatsakymų atveju

Šiame straipsnyje nagrinėjamas modelio ir neatsakymų vertinimo būdo parinkimo efektyvumas skirtingų tipų sumos įverčiams mažose srityse. Empiriniai rezultatai paremti Monte Karlo simulaicijomis su pasikartojančiomis imtimis, kurios buvo renkamos iš populiacijos, sukonstruotos remiantis tikrais duomenimis, gautais iš Lietuvos įmonių tyrimo.

Raktiniai žodžiai: populiacija, imtis, maža sritis, neatsakymai, atsakymo tikimybė, įvertis, modelis.