Lietuvos matematikos rinkinys. LMD darbai
51 tomas, 2010, 414–419

ISSN 0132-2818
www.mii.lt/LMR/

Weighting and imputation comparison in small area estimation

Vilma Nekrašaitė-Liegė

Vilniaus Gedimino Technical University
Saulėtekio 11, LT-10223 Vilnius
E-mail: nekrasaite.vilma@gmail.com

Abstract. In this paper, different methods of nonresponse adjustment for the totals of small area domains are examined. To improve quality of estimations linear model with random parameters at domain level is used. The empirical results are based on Monte Carlo simulations with repeated samples drawn from a finite population constructed from the Lithuanian survey on short-term statistics on service.

Keywords: small area, response probability, imputation, model.

Introduction

In true surveys, we are faced with nonresponse. Nonresponse not only mean less efficient estimates because of reduced sample size, but also standard-complete methods cannot be immediately used to analyze data. There are several methods to correct the consequences of unit nonresponse [1, 6] and they are examined in this paper for the small areas. The focus on small area is made because there is relatively little of an impartial comparison between nonresponse treatment [4]. Also for the small area estimation an important aspect is to choose the right estimator and model [2, 3]. During the previous research [4] it was showed, that the linear model with random parameters at domain level is a good choice, but which estimator to use is still an open question. That is why, two different estimators (generalized regression estimator [8] and empirical best linear unbias predictor [5]) are investigated in this paper.

1 Population

Let $U = \{1, 2, \ldots, k, \ldots, N\}$ denote a finite population with N units. This population is divided into D nonoverlaping domains U_d, $d = 1, \ldots, D$, consisting of N_d units. A sample $\mathbf{s}$ of the size n units is selected from the population U, $\mathbf{s} = \{s_1, s_2, \ldots, s_n\} \subset U$. Each unit k has an inclusion probability $\pi_k = \mathbf{P}(k \in \mathbf{s})$ or a sampling weight $w_k = \pi_k^{-1}$. For different reasons there are missing units in the sample $\mathbf{s}$. Let a response probability for each unit be $\varkappa_k = \mathbf{P}(k \in \mathbf{s}^{(r)}, k \in \mathbf{s})$, where $\mathbf{s}^{(r)} \subset \mathbf{s}$ is a responded sample.

Let us denote y as a study variable, which values y_k are known just for the elements of a response sample $\mathbf{s}^{(r)}$ and $\mathbf{x} = (x_1, x_2, \ldots, x_J)'$ as 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. It is assumed that the number of the elements in each

domain U_d, $d = 1, \ldots, D$, is known, but the domains are not used in the sample design. This means that the sample part in each domain, $\mathbf{s} \cap U_d$, has a random size.

2 Model and estimators

Let values $y_1, \ldots, y_N$ of a study variable y be realizations of independent random variables $Y_1, \ldots, Y_N$, which satisfy the following general linear model [7]:

$$Y_k = \beta_0 + \sum_{d=1}^{D} I_{dk} u_d + \sum_{j=1}^{J} \beta_j x_{jk} + \varepsilon_k, \quad k \in U. \tag{1}$$

Here β_0 and β_j, $j = 1, \ldots, J$, are regression coefficients, u_d, $d = 1, \ldots, D$, are random parameters that are related to the corresponding domain and I_{dk}, $d = 1, \ldots, D$, $k = 1, \ldots, N$, are domain indicators ($I_{dk} = 1$, if $k \in U_d$ and $I_{dk} = 0$ otherwise). The errors ε_k and random parameters u_d are assumed to be independent and identically distributed Gaussian random variables with mean 0 and variances σ^2 and σ_1^2 respectively. Using restricted maximum likelihood (REML) method with incorporated weights [7] model's parameters are estimated and predicted values $\hat{y}_k = \hat{\beta}_0 + \sum_{d=1}^{D} I_d \hat{u}_d + \sum_{j=1}^{J} \hat{\beta}_j x_{jk}$ are computed for all $k \in U$.

These predicted values are used to estimate two different domain total estimators: generalized regression (GREG) estimator [8]

$$\hat{t}_{d_G} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in \mathbf{s} \cap U_d} w_k (y_k - \hat{y}_k), \quad d = 1, \ldots, D, \tag{2}$$

and empirical best linear unbiased predictor (EBLUP estimator) [5]

$$\hat{t}_{d_E} = \sum_{k \in U_d \setminus \mathbf{s}} \hat{y}_k + \sum_{k \in \mathbf{s} \cap U_d} y_k, \quad d = 1, \ldots, D. \tag{3}$$

These estimators in the case of significant nonresponse rate should be corrected.

3 Methods for nonresponse adjustment

Weighting and imputation are the two main methods used to correct for bias due to nonresponse and to make efficient use of data.

3.1 Weighting

Using weighting method the original inclusion probabilities π_k are deflated by the response probabilities $\varkappa_k$ and new sampling weights $w_k = (\pi_k \varkappa_k)^{-1}$, $k \in U$, are obtained. The original response probability is never known in practice, so there are several methods to estimate it. One of them is called a weighting-class, where the response sample $\mathbf{s}^{(r)}$ and sample $\mathbf{s}$ are divided into G mutually exclusive and homogenous (with respect of the response rate) groups $\mathbf{s}_g^{(r)}$ and $\mathbf{s}_g$, $g = 1, \ldots, G$, with the same response probability $\varkappa_k$ for the unit in the same group:

$$\hat{\varkappa}_k = \frac{\sum_{j \in \mathbf{s}_g^{(r)}} w_j}{\sum_{j \in \mathbf{s}_g} w_j} = \frac{\hat{N}_g^{(r)}}{\hat{N}_g}, \quad k \in \mathbf{s}. \tag{4}$$

Another method for estimating the response probability is to apply a logistic regression model [1]:

$$\hat{\varkappa}_k = \frac{\exp\{\hat{B}\mathbf{x}_k\}}{1+\exp\{\hat{B}\mathbf{x}_k\}}, \quad k \in \mathbf{s}. \tag{5}$$

Here $\hat{B}$ is the maximum likelihood estimator of the parameters of the logistic regression model based on the data $\{(z_k, x_k), k \in \mathbf{s}\}$ where $z_k = 1$, if $k \in \mathbf{s}^{(r)}$, and $z_k = 0$ otherwise.

When weighting methods for nonresponse adjustment are applied in the estimation of the domain total, the correction of estimators (2) and (3) should be made by replacing sampling weights w_k with $\hat{w}_k = (\pi_k \hat{\varkappa}_k)^{-1}$, $k \in \mathbf{s}$, not only in equations (2) and (3), but also in calculation of the parameters of the model (1).

3.2 Imputation

Another method to adjust nonresponse is to impute a value for the missing unit. There are many types of imputation methods, which can be divided into three main groups:

1) Logical imputation (deductive). It is a part of the editing process and is used when reliable, explicit solution exists given appropriate assumptions.

2) Real donor imputation. Here the imputed observation value is borrowed from another respondent. The most common real donor imputations are the nearest neighbors and the random donor imputation. For the nearest neighbor imputation, a missing value y_k is imputed by choosing that value y_l which corresponds to the value $\mathbf{x}_l$ closest to $\mathbf{x}_k$. The closest value is determined by the distance between any two response values ($d_{kl} = \sqrt{\sum_{j=1}^{J}(x_{kj} - x_{lj})^2}$, $k \in \mathbf{s}\backslash\mathbf{s}^{(r)}$, $l \in \mathbf{s}^{(r)}$). For the random donor imputation the data are divided into homogenous groups by a suitable method and the donors are chosen randomly within these groups.

3) Model based imputation. Here the imputed observation value is calculated using the model with the coefficients estimated from the response sample $\mathbf{s}^{(r)}$. The most common method is a regression imputation.

Imputation methods can also be classified as a single imputation (when one value is imputed instead of missing one) or multiple imputation. Multiple imputation produces several imputed datasets and instead of the missing value a mean of imputed datasets is used.

Let us denote a new variable y^* which values y_k^* are equal to y_k, if $k \in \mathbf{s}^{(r)}$, or y_k^{imp}, if $k \in U\backslash\mathbf{s}^{(r)}$. Here y_k^{imp} can be a single value, if single imputation is used, or the mean of imputed datasets, if multiple imputation is used. Then estimators (2) and (3) can be written as follows:

$$\hat{t}_{d_G} = \sum_{k\in U_d} \widehat{y^*}_k + \sum_{k\in\mathbf{s}\cap U_d} w_k(y_k^* - \widehat{y^*}_k), \quad d = 1, \ldots, D, \tag{6}$$

and

$$\hat{t}_{d_E} = \sum_{k\in U_d\backslash\mathbf{s}} \widehat{y^*}_k + \sum_{k\in\mathbf{s}\cap U_d} y_k^*, \quad d = 1, \ldots, D. \tag{7}$$

4 Simulation study

For the simulation experiment, a real population from Statistics Lithuania is used. The quarterly survey on short-term statistics on service has been taken. The population includes $N = 1660$ enterprisers, which filled questionnaire in the first quarter of 2008 year. Every record consists of such variables: region of residence, income for the first quarter of 2008 year, number of employees in the same quarter, value-added tax (VAT), classification of economic activities in the European community (NACE).

The income is chosen as the study variable y. Let y_k denote the value of y for the kth enterpriser, $k = 1, \ldots, N$. The parameter of interest is total income in each region (domain total – t_d). There are $D = 14$ regions of interest. To improve the quality of the estimators 7 auxiliary variables were used: number of employees (x_1), VAT (x_2) and indicator of the NACE ($x_3 - x_7$). 1000 independent samples of 80 elements are drawn from the population by simple random sampling without replacement (SRS).

The GREG and the EBLUP estimators are used to estimate the domain total. Each estimate is calculated several times using different methods of nonresponse adjustment. These differences are denoted by adding two letters, $LL \in \{WC, LR, RD, NN, CR, DR\}$, and number, $R \in \{0, 1, 2\}$, at the end of the estimate's name (GREG-LLR or EBLUP-LLR). The meaning of these abbreviations is described below. The weighting-class method (WC) and the logistic regression model (LR) are applied to estimate response probability. Also, the performance of different imputation methods (random donor (RD), nearest neighbors (NN), regression imputation using the common model (CR) [2] and regression imputation using the model with domain-intercepts (DR) [2]) is investigated. For weighting-class, random donor and nearest neighbors methods units are grouped by the number of employees and indicator of the NACE. For the logistic regression model auxiliary vector $\mathbf{x}$ with values $\mathbf{x}_k = (1, x_{1k}, x_{2k}, x_{3k}, x_{5k})'$ is used. For the regression imputation the mean of imputed datasets with 5 values is applied. For the common model auxiliary vector $\mathbf{x}$ with values $\mathbf{x}_k = (1, x_{1k}, x_{2k}, x_{3k}, x_{4k}, x_{5k}, x_{6k})'$ is used. For the model with domain-intercepts auxiliary vector $\mathbf{x}$ with values $\mathbf{x}_k = (I_{1k}, \ldots, I_{Dk}, x_{1k}, x_{2k}, x_{3k}, x_{4k}, x_{5k}, x_{6k})'$ is applied. Here $I_{dk} = 1$ if unit k belongs to d domain and $I_{dk} = 0$ otherwise, $d = 1, \ldots, D$.

Each of these methods of nonresponse adjustment is applied for two populations constructed from the real population (indicated by the number $R = 0$) by generating different response rate. The response rates of 89% and 79% are generated for the first ($R = 1$) and the second ($R = 2$) populations, respectively. These rates represents the response rate in the survey (actually the response rate depends on region, number of employees and NACE).

To compare the results two accuracy measures are applied for $M = 1000$ simulations: the absolute relative bias $ARB(\hat{t}_d) = |\frac{1}{M}\sum_{m=1}^{M} \hat{t}_d^{(m)} - t_d|/t_d$ and the relative root means square error $RRMSE(\hat{t}_d) = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{t}_d^{(m)} - t_d)^2}/t_d$, $d = 1, \ldots, D$. 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 14 regions of interest, so for the better comparison these regions are grouped into two domain sample size class by the average number of elements in domain sample (minor 0–5 and medium 6–10). A mean of absolute relative bias (MARB) and a mean of relative root means square error (MRRMSE) in each class are calculated (see Tables 1–2).

Table 1. Results for different methods of nonresponse adjustment using GREG estimator.

Estimator	Domain sample size class			
	Minor 0–5		Medium 6–10	
	MABR, %	*MRRMSE*, %	*MABR*, %	*MRRMSE*, %
GREG-0	1.2	34.9	1.2	28.3
GREG-WC1	1.1	39.0	1.6	32.0
GREG-LR1	1.3	37.6	1.4	30.3
GREG-RD1	6.4	64.1	2.5	48.9
GREG-NN1	4.7	55.7	3.3	42.7
GREG-CR1	1.1	36.5	1.7	30.9
GREG-DR1	0.9	36.6	1.6	30.8
GREG-WC2	3.2	36.9	2.3	31.9
GREG-LR2	3.0	34.8	2.1	29.5
GREG-RD2	8.9	70.5	7.3	58.7
GREG-NN2	5.5	58.3	3.3	43.2
GREG-CR2	3.2	40.8	3.5	35.5
GREG-DR2	3.5	43.0	3.6	34.8

Table 2. Results for different methods of nonresponse adjustment using EBLUP estimator.

Estimator	Domain sample size class			
	Minor 0–5		Medium 6–10	
	MABR, %	*MRRMSE*, %	*MABR*, %	*MRRMSE*, %
EBLUP-0	6.0	17.6	3.1	17.2
EBLUP-WC1	6.0	18.1	3.3	19.4
EBLUP-LR1	5.8	16.9	2.9	17.5
EBLUP-RD1	8.4	43.0	5.9	39.8
EBLUP-NN1	9.6	36.4	7.8	32.0
EBLUP-CR1	6.0	20.2	2.8	18.0
EBLUP-DR1	5.8	20.8	3.0	18.4
EBLUP-WC2	5.6	22.4	4.2	21.6
EBLUP-LR2	5.5	20.0	3.9	19.2
EBLUP-RD2	11.1	50.7	6.9	47.2
EBLUP-NN2	9.2	41.2	6.2	35.9
EBLUP-CR2	6.3	22.2	3.2	21.0
EBLUP-DR2	6.2	21.6	3.0	20.6

5 Conclusions

The results in the tables show that the real donor imputation (nearest neighbor and random donor) are the worst methods, since they increase bias and MRRMSE more than the others methods. The weighting methods (weighting-class and logistics regression) yield the best results in small area estimation with nonresponse. Alike the other nonresponse adjustment methods, they do not so much depend on the nonresponse rate.

References

[1] A. Ekholm and S. Laaksonen. Weighting via response modeling in the finnish household budget survey. *Journal of Official Statistics*, **7**(3):325–337, 1991.

[2] R. Lehtonen, C.-E. Sarndal and A. Veijanen. The effect of model choice in estimation for domains, including small domains. *Survey Methodology*, **29**:33–44, 2003.

[3] R. Lehtonen, C.-E. Sarndal and 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.

[4] V. Nekrašaitė-Liegė. Small area estimation in the case of nonresponse. *Lietuvos matematikos rinkinys. LMD darbai*, **50**:304–309, 2009.

[5] J.N.K. Rao. *Small Area Estimation*. Wiley, New York, 2003.

[6] DB. Rubin. *Multiple imputation for nonresponse in surveys*. Wiley, New York, 1987.

[7] A. Saei and R. Chambers. Small area estimation under linear and generalized linear mixed models with time and area effects. *EURAREA Consortium 2004, Project Reference Volume*, 2004. Available from Internet: http://eprints.soton.ac.uk/8165/. [Cited 25 May 2010]. (Article with URL and lastchecked).

[8] C.-E. Sarndal, B. Swensson and J. Wretman. *Model Assisted Survey Sampling*. Springer-Verlag, New York, 1992.

REZIUMĖ

Persvėrimo ir įrašymo metodų palyginimas mažose srityse

V. Nekrašaitė-Liegė

Šiame straipsnyje nagrinėjami sumos įvertiniai mažose srityse, kai neatsakymai vertinami skirtingais metodais. Įvertinių kokybei pagerinti naudojamas tiesinis modelis su atsitiktiniais parametrais. Empiriniai rezultatai paremti Monte Karlo simuliacijomis su pasikartojančiomis imtimis, kurios buvo renkamos iš populiacijos sukonstruotos remiantis Lietuvos paslaugų įmonių statistiniu tyrimu.

Raktiniai žodžiai: maža sritis, atsakymo tikimybė, įrašymas, modelis.