Liet. matem. rink., T. 43, spec. nr., 480–485
© 2003 Matematikos ir informatikos institutas

Dealing with nonresponse in business surveys

Audrius INDRIULIONIS (Statistics Lithuania)
e-mail: audriusi@mail.std.lt

1. Introduction

Nonresponse is present in all surveys. The literature provides various statistical methods for treating nonresponse: imputation, reweighting, subsampling of nonrespondents and other ones. The paper considers the case study of three methods of estimation of response probability: simple reweighting, exponential and logistic regression methods. Real data of the Lithuanian business register are used for simulation.

2. Notation

Let us consider a finite population of N elements $U = \{1, 2, \dots, N\}$ and let y be a variable of our interest taking values $y_1, y_2, \dots, y_N$, where y_k is the value of y for the k-th population element. We are estimating the population total

$$t_y = \sum_{k=1}^{N} y_i.$$

A stratified simple random sample is considered. The population U is divided into H strata

$$U = U_1 \cup U_2 \cup \ldots \cup U_H,\ U_k \cap U_l = \emptyset, \quad k \neq l.$$

Denote by y_{hk} values of the variable y in the stratum U_h and write

$$t_y = \sum_{h=1}^{H} \sum_{k \in U_h} y_{hk}.$$

Let $s_h \subset U_h$ be a sample in the stratum U_h and consider the Horvitz–Thompson estimator of t_y in the stratified sample

$$\hat{t}_y = \sum_{h=1}^{H} \sum_{k \in s_h} \frac{y_{hk}}{\pi_{hk}}.$$

Here $\pi_{hk} = n_h/N_h$ is the inclusion probability of the k-th element from the stratum h. Let $r_h \subset s_h$ be a set of responded elements in the stratum h. Denote

$$\pi^*_{hk} = P\{k \in s_h \& k \in r_h\} = P\{k \in s_h\}P\{k \in r_h \mid k \in s_h\} = \pi_{hk}\theta_{hk}.$$

Let us consider an unbiased estimator of t_y

$$\hat{t}^*_y = \sum_{h=1}^{H} \sum_{k \in r_h} \frac{y_{hk}}{\pi_{hk}\theta_{hk}} \tag{2.1}$$

that can be used in the presence of nonresponse. In practice π_{hk} are known and conditional response probabilities θ_{hk} are not known. The estimates $\hat{\theta}_{hk}$ are used in (2.1) instead of θ_{hk}. We consider three methods of estimation of θ_{hk}.

3. Methods for estimating the response probability θ_{hk}

3.1. *Reweighting*

In the case of a stratified simple random sample θ_{hk} can be estimated by

$$\hat{\theta}_{hk} = \hat{\theta}_h = \frac{n_h^{(r)}}{n_h}, \quad k \in U_h,$$

where n_h is the number of sampled elements in the stratum h, $n_h^{(r)}$ is the number of responded elements in the stratum h. The reweighting estimator of total is

$$\hat{t}^{REW}_y = \sum_{h=1}^{H} \frac{N_h}{n_h^{(r)}} \sum_{k \in r_h} y_{hk}.$$

3.2. *Exponential function*

Let x_{hk} and z_{hk} denote respectively the number of employees and the salary fund of the k-th enterprise in the stratum h. The values of x_{hk} and z_{hk} are known for all population elements. Suppose the conditional response probability θ_{hk} has the shape

$$\theta_{hk} = 1 - \mathrm{e}^{-(\alpha+\beta x_{hk}+\gamma z_{hk})},$$

where α, β, γ are unknown parameters. It is estimated as follows

$$\hat{\theta}^*_{hk} = 1 - \mathrm{e}^{-(\hat{\alpha}+\hat{\beta} x_{hk}+\hat{\gamma} z_{hk})},$$

where $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}$ are the maximum likelihood estimators of α, β, γ. In this case, the estimator of the total t_y is

$$\hat{t}_y^{EXP} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k\in r_h} \frac{y_{hk}}{\hat{\theta}_{hk}^{*}}.$$

3.3. *Logistic regression function*

The third case considered is the logistic regression model for θ_{hk} with x_{hk} and z_{hk} as auxiliary variables:

$$\theta_{hk} = \frac{1}{1+\mathrm{e}^{-(\alpha+\beta x_{hk}+\gamma z_{hk})}}.$$

In this case θ_{hk} is estimated by

$$\hat{\theta}_{hk}^{**} = \frac{1}{1+\mathrm{e}^{-(\hat{\alpha}+\hat{\beta} x_{hk}+\hat{\gamma} z_{hk})}},$$

where $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}$ are maximum likelihood estimators of α, β, γ. The estimators are calculated using SAS procedure. The estimator of total is

$$\hat{t}_y^{RLOG} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k\in r_h} \frac{y_{hk}}{\hat{\theta}_{hk}^{**}}.$$

3.4. *Calibrated estimators*

The calibration of weights is a statistical method for improving estimators using auxiliary information. The basic idea of calibration of weights of the estimators of totals is proposed in [1]. Calibrated estimators may be used in the presence of nonresponse which are introduced in [2].

3.5. *Estimation of variance*

It is supposed that the elements respond independently. The variance estimator of the estimator of total in two phase sample design, given in [3], is used. In the case of a stratified simple random sample, this estimator can be written like this:

$$\widehat{D}(\hat{t}_y) = \sum_{h=1}^{H} \frac{N_h^2}{n_h}\Big(1-\frac{n_h}{N_h}\Big)\frac{1}{n_h-1}\left[\sum_{k\in r_h}\frac{y_{hk}^2}{\theta_{hk}^2} - \frac{1}{n_h}\Big(\sum_{k\in r_h}\frac{y_{hk}}{\theta_{hk}}\Big)^2\right] + \sum_{h=1}^{H}\frac{N_h^2}{n_h^2}\sum_{k\in r_h}\Big(\frac{1}{\theta_{hk}^2}-\frac{1}{\theta_{hk}}\Big)y_{hk}^2.$$

Taking $\hat{\theta}_{hk}$, $\hat{\theta}^{*}_{hk}$, $\hat{\theta}^{**}_{hk}$ instead of θ_{hk}, we get the variance estimators of $\hat{t}_y^{REW}$, $\hat{t}_y^{EXP}$ and $\hat{t}_y^{RLOG}$, respectively. The randomness of $\hat{\theta}_{hk}$, $\hat{\theta}^{*}_{hk}$, $\hat{\theta}^{**}_{hk}$ is not taken into account.

4. Simulation results

4.1. *Sampling probabilities and response rates*

The simulated response rates are quite similar to those that appear in the real business surveys. Different sample sizes are $n_1 = 189$, $n_2 = 78$, $n_3 = 50$. The population size is 361. We examined two types of auxiliary information in our simulation: strongly correlated with a study variable (the coefficient of correlation greater than 0.8) and with the coefficient of correlation less than 0.5. In Figs. 1, 2 and 3, 4, the average standard error of estimates and root mean square error of estimates with the strong and weak coefficients of correlation are displayed, respectively.

4.2. *The measures of accuracy*

The simulated estimator of root mean square error $\widehat{RMSE}_{SIM}(\hat{t}_y)$ and the simulated estimator of the coefficient of variation $\widehat{CV}_{SIM}(\hat{t}_y)$ are taken as the measures of accuracy:

$$\widehat{RMSE}_{SIM}(\hat{t}_y) = \left[\widehat{D}_{SIM}(\hat{t}_y) + \widehat{BIAS}^2(\hat{t}_y)\right]^{\frac{1}{2}}.$$

Here

$$\widehat{D}_{SIM}(\hat{t}_y) = \frac{1}{M}\sum_{j=1}^{M}\widehat{D}(\hat{t}_y^{(j)}),$$

and

$$\widehat{BIAS}_{SIM}(\hat{t}_y) = \frac{1}{M}\sum_{j=1}^{M}(t_y - \hat{t}_y^{(j)}).$$

The simulated estimator of coefficient of variation is:

$$\widehat{CV}_{SIM}(\hat{t}_y) = \frac{1}{M}\sum_{j=1}^{M}\frac{\sqrt{\widehat{D}(\hat{t}_y^{(j)})}}{\hat{t}_y^{(j)}}.$$

Here M denotes the number of repeated samples. In our case, $M = 1000$, $\hat{t}_y^{(j)}$ and $\widehat{D}(\hat{t}_y^{(j)})$, $j = 1, \ldots, M$, denote the estimate of t_y and that of the variance of $\hat{t}_y$ for the j-th sample, respectively.

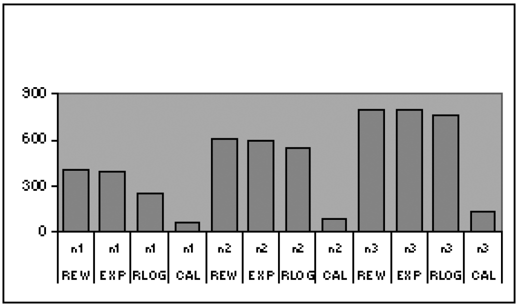

Fig. 1. Average standard error of estimates (Strong correlation).

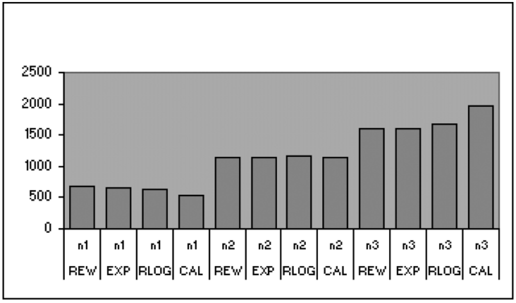

Fig. 2. Average standard error of estimates (Weak correlation).

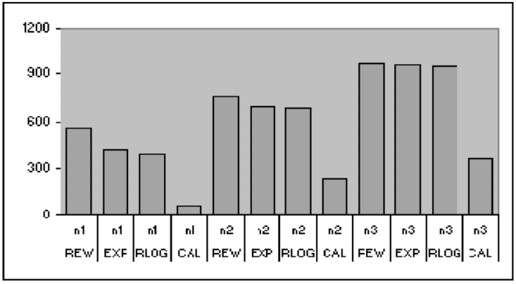

Fig. 3. Root mean square error of estimates (Strong correlation).

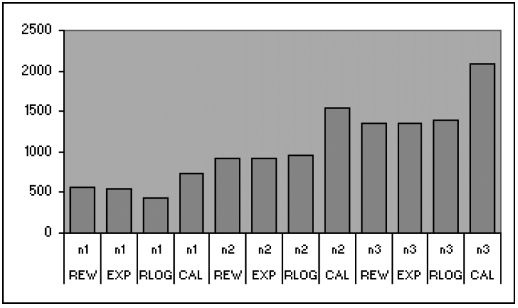

Fig. 4. Root mean square error of estimates (Weak correlation).

5. Some conclusions

- Exponential and logistic regression methods give the better results than simple reweighting method.
- Some research is needed to decide which of those methods – exponential or logistic regression – is better.
- The effect of calibrated estimators are shown in Figs. 1, 3 with strong coefficient of correlation.

References

[1] J.-C. Deville, C.-E. Särndal, Calibration estimators in survey sampling, *Journal of the American Statistical Association*, **50**, 376–382 (1992).

[2] S. Lundström, *Calibration as Standart Method for Treatment of Nonresponse*, Doctoral dissertation, Stockholm University (1997).

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

Neatsakymų vertinimas įmonių tyrimuose

A. Indriulionis

Šiame darbe buvo palyginti skirtingi atsakymo į apklausą tikimybių vertinimo būdai. Paprastasis persvėrimo, eksponentinis ir logistinės regresijos metodai buvo naudojami norint gauti sumų įverčius, kai į apklausą atsako ne visi į imtį išrinkti elementai. Atsakymų į apklausą lygiai panašūs į stebimus realiuose įmonių tyrimuose.