ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 216, 2008

Wojciech Gamrot[*]

ON SOME COMPOSITE ESTIMATOR OF THE POPULATION MEAN

ABSTRACT. In this paper an estimator of the finite population mean in the unit nonresponse situation is proposed. It is constructed as a combination of the well-known regression estimator derived from the linear model and a reweighting-type estimator based on a logistic regression model. Combination weights depend on goodness of fit of respective models. Hence, the estimator for which the corresponding model better describes observed sample data dominates in the combination. Some Monte Carlo simulation results revealing its properties are presented.

Key words: nonresponse, regression, weighting adjustment.

I. INTRODUCTION

Consider a finite and fixed population U of size N. A mean value $\overline{Y} = N^{-1}\sum_{i\in U} y_i$ of some characteristic Y taking values $y_1,...,y_N$, is to be estimated. A sample s of size n is drawn from U according to the sampling design p(s) determining the inclusion probabilities of the first order denoted by π_i for $i, j \in U$. Assume stochastic nonresponse that does not depend on the sample. Hence, an individual response probability ρ_i may be associated with each unit and the sample s is randomly divided into subsets: s_1 and s_2, containing responding and non-responding units respectively. Under nonresponse, the well-known Horvitz-Thompson estimator of the population mean is biased, when computed solely on the basis of responding units. Bethlehem (1988) considers the following modification of this estimator:

$$\overline{y}_{MHT} = \sum_{i\in s_1} \frac{y_i}{\pi_i} \Big/ \sum_{i\in s_1} \frac{1}{\pi_i} \tag{1}$$

[*] Ph.D. Department of Statistics, University of Economics, Katowice.

and shows that its bias is aproximately equal to: $B(\overline{y}_{MHT}) = C_U(y_i,\rho_i)/\overline{\rho}$, where $\overline{\rho} = N^{-1}\sum_{i\in U}\rho_i$ and $C_U(y_i,\rho_i) = N^{-1}\sum_{i\in U}(y_i - \overline{Y})(\rho_i - \overline{\rho})$. Hence, the lower the covariance between y_i and ρ_i, the lower the bias. This estimator will be denoted by the symbol MHT.

Consider the superpopulation model ξ, stating that values $y_1,...,y_N$ are realizations of independent random variables $Y_1,...,Y_N$, satisfying:

$$\begin{cases} E_\xi(Y_i) = \boldsymbol{\beta}\mathbf{x}_i \\ V_\xi(Y_i) = \sigma^2 \end{cases} \tag{2}$$

for i=1,..,N. The vector $\boldsymbol{\beta} = [\beta_1,...,\beta_k]$ and scalar σ^2 are model parameters, while $\mathbf{x}_i = [x_1,...,x_H]'$ denotes vector of auxiliary characteristics $X_1,...,X_H$ associated with i-th unit. Denoting $\mathbf{X} = [\mathbf{x}_1,...,\mathbf{x}_N]$ and $\mathbf{y} = [y_1,...,y_N]$ and applying ordinary least squares we obtain the best linear unbiased (with respect to ξ) estimator of $\boldsymbol{\beta}$:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{3}$$

The quantity **b** may be estimated from the sample by the statistic:

$$\hat{\mathbf{b}} = \left(\sum_{i\in s_1}\frac{\mathbf{x}_i\mathbf{x}_i'}{\pi_i}\right)^{-1}\left(\sum_{i\in s_1}\frac{\mathbf{x}_i y_i}{\pi_i}\right). \tag{4}$$

Consider the regression estimator of the population mean

$$\overline{y}_{REG} = \overline{y}_{MHT} + \hat{\mathbf{b}}(\overline{\mathbf{x}}_{HT} - \overline{\mathbf{x}}_{MHT}) \tag{5}$$

where $\overline{\mathbf{x}}_{HT} = N^{-1}\sum_{i\in s}(\mathbf{x}_i/\pi_i)$ and $\overline{\mathbf{x}}_{MHT} = \sum_{i\in s_1}(\mathbf{x}_i/\pi_i)\Big/\sum_{i\in s_1}(1/\pi_i)$. It is more accurate than $\overline{y}_{MHT}$ when (2) accurately reflects reality. It will be denoted further by the symbol REG.

Another approach to construct nonresponse-corrected population mean estimator relies on assumed dependencies between auxiliary variables and response probabilities. These relations are represented by parametric models,

such as logistic model (see Rizzo et al. 1996, Ekholm and Laaksonen 1991), stating that units respond independently with probabilities:

$$\rho_i = \frac{1}{1+e^{\lambda x_i'}} \tag{6}$$

for $i \in U$, where $\boldsymbol{\lambda} = [\lambda_1, ..., \lambda_H]$ is a parameter vector. Its maximum likelihood estimate $\hat{\boldsymbol{\lambda}} = [\hat{\lambda}_1, ..., \hat{\lambda}_H]$ may be obtained using iterative methods considered e.g. by Minka (2001). Consequently, by replacing unknown ρ_i's with estimates $\hat{\rho}_i = (1+e^{\hat{\lambda} x_i'})^{-1}$ we obtain the following weighting-adjustment mean value estimator:

$$\bar{y}_{RHO} = \frac{1}{N} \sum_{i \in s_1} \frac{y_i}{\pi_i \hat{\rho}_i}. \tag{7}$$

This estimator should be more acurate than $\bar{y}_{MHT}$ when the model (6) accurately describes the behavior of ρ_i's. In the following study it will be denoted by the symbol RHO.

II. COMPOSITE ESTIMATOR

The attractiveness of both regression and reweighting-type estimator depends on the ability of underlying models to describe the behavior of y_i or ρ_i. One may attempt to measure this ability. The goodness of fit of the regression model (2) may be measured by means of the respondent subset determination coefficient given by formula:

$$R = \left(\sum_{i \in s_1} (\hat{\mathbf{b}}^* \mathbf{x}_i - \frac{1}{n_1} \sum_{j \in s_1} \hat{\mathbf{b}}^* \mathbf{x}_j)^2 \right) \Big/ \left(\sum_{i \in s_1} (y_i - \frac{1}{n_1} \sum_{j \in s_1} y_j)^2 \right). \tag{8}$$

Moreover, the goodness of fit of the logistic model may be measured by the log-likelihood function $\ln L = \sum_{i \in s_1} \ln \hat{\rho}_i + \sum_{i \in s_2} \ln (1 - \hat{\rho}_i)$ or by the standarized quantity:

$$R_L = 2(\sqrt[n]{\ln L} - 0.5) \tag{9}$$

that shall take values from the <0,1> interval. Let us now consider a composite estimator:

$$\bar{y}_{COM} = \alpha_{REG}\bar{y}_{REG} + \alpha_{RHO}\bar{y}_{RHO} \tag{10}$$

where $\alpha_{REG} = R/(R+R_L)$ and $\alpha_{RHO} = R_L/(R+R_L)$ are weights proportional to the goodness of fit of respective model. The composite estimator should behave like regression estimator when auxiliary information is more suitable for linear model, and behave like weighting adjustment estimator when available auxiliary information is more appropriate for logistic model. Hence the composite estimator should inherit the virtues of both. In the following paragraphs it will be denoted by the symbol COM.

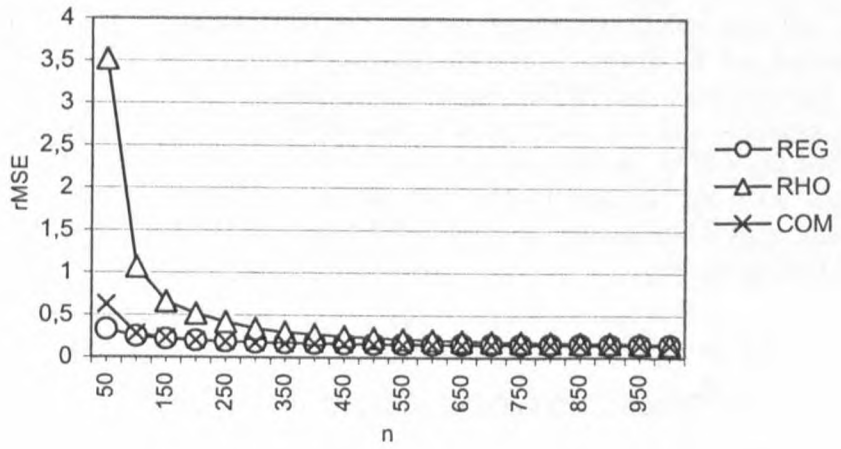
III. SIMULATION RESULTS

A simulation study was carried out to examine the properties of four estimators: MHT, REG, RHO and COM. Experiments were executed using pseudo-random number generator of multivariate Gaussian distribution. Four variables: Y, X_1, X_2, X_3 were generated for 10000 population units with Y being variable under study, X_1 being auxiliary variable for regression estimator, X_2 being auxiliary variable for logistic model and X_3 being unknown to sampler, determining individual response probabilities according to univariate logistic model: $\rho_i = (1+e^{X_3})^{-1}$. Simple samples were repeatedly drawn without replacement from the population. Survey behavior of each unit was independently simulated assuming the response probability equal to ρ_i. All estimators were computed using resulting incomplete data and their empirical distributions were examined. Three simulation experiments were carried out for correlation matrices between variables respectively equal to:

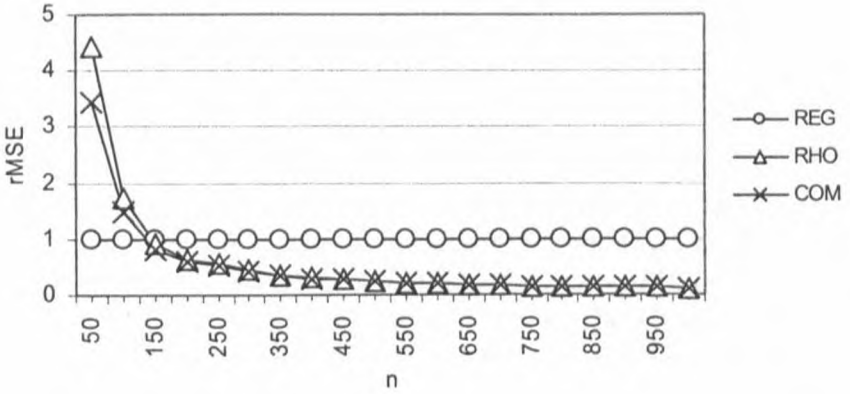
$$R_1 = \begin{bmatrix} 1 & 0.7 & 0.75 & 0.75 \\ 0.7 & 1 & 0.7 & 0.7 \\ 0.75 & 0.7 & 1 & 0.75 \\ 0.75 & 0.7 & 0.75 & 1 \end{bmatrix}, \quad R_2 = \begin{bmatrix} 1 & 0 & 0.9 & 0.9 \\ 0 & 1 & 0 & 0 \\ 0.9 & 0 & 1 & 0.9 \\ 0.9 & 0 & 0.9 & 1 \end{bmatrix},$$

$$R_3 = \begin{bmatrix} 1 & 0.9 & 0 & 0.9 \\ 0.9 & 1 & 0 & 0.9 \\ 0 & 0 & 1 & 0 \\ 0.9 & 0.9 & 0 & 1 \end{bmatrix}. \tag{11}$$

All standard deviations were set to one. Mean value vector was always equal to $\boldsymbol{\mu} = [10,10,10,0]$. Matrix R_1 represents the situation when auxiliary data is suitable for both REG and RHO estimators. With R_2 it is suitable only for RHO. With R_3 it is suitable only for REG. All simulations were carried out for n = 50, 100, ..., 1000. Efficiency of estimators relative to the MHT estimator is shown on graphs 1-3. In all three experiments it is computed for any estimator T as $\mathrm{rMSE(T)} = \mathrm{MSE(T)} / \mathrm{MSE}(\bar{y}_{\mathrm{MHT}})$.

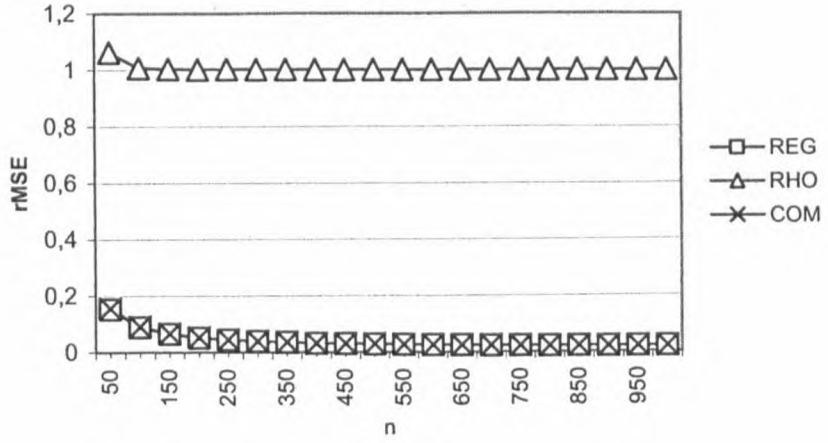


Pic. 1. The relative efficiency as a function of sample size n for correlation matrix R_1

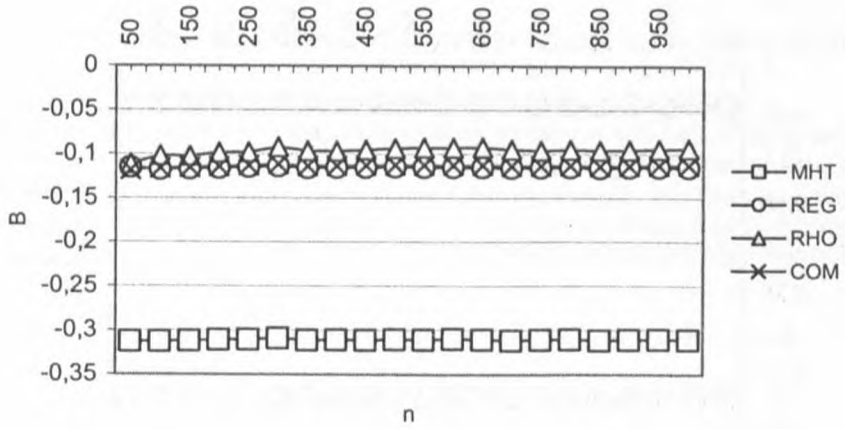


Pic. 2. The relative efficiency as a function of sample size n for correlation matrix R_2

For all estimators the relative efficiency diminishes with growing sample size and then stabilizes for large values of n, with notable exception of REG estimator and R_2 where it is approximately constant. Estimators REG and RHO are more accurate than MHT when their respective models fit well to the data. The estimator COM is more accurate than MHT when at least one of these models fits well to the data. One may say that this estimator is more robust with respect to model misspecification than REG and RHO. Moreover for large sample sizes it has usually the lowest MSE, although the advantage over REG and RHO is modest.

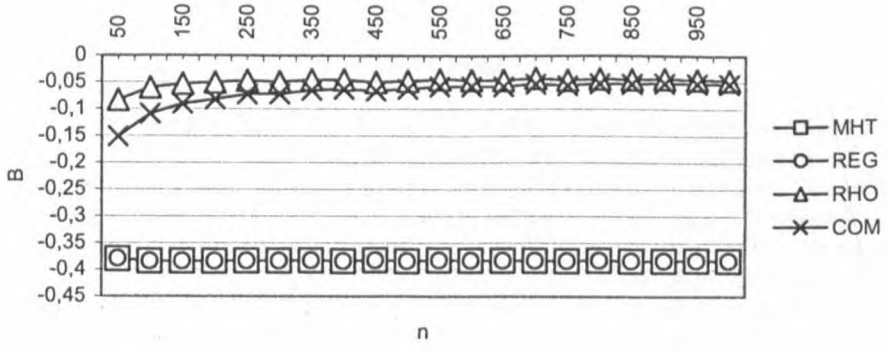


Pic. 3. The relative efficiency as a function of sample size n for correlation matrix R_3

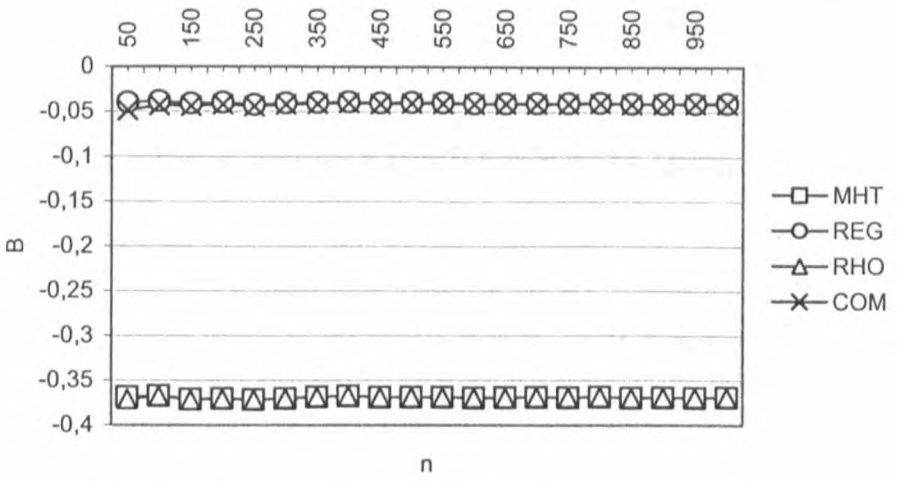


Pic. 4. The bias as a function of sample size n for correlation matrix R_1

The bias of all estimators is shown on graphs 4-6. All of them are biased negatively. The bias is constant or slowly diminishes with growing sample size n to stabilize when n is large enough. In absolute terms, the bias of MHT estimator was the highest in most cases. The estimators REG and RHO provide substantial bias reduction when auxiliary information is suitable and respective models fit well to the data. Otherwise, their bias is very close to the one of MHT estimator. The bias of the COM estimator is contained between the biases of REG, RATIO and MHT. Usually it does not differ much from the lowest observed bias. In all experiments it is significantly lower than the bias of MHT so the composite estimator provides substantial bias reduction when at least one model fits the data well.



Pic. 5. The bias as a function of sample size n for correlation matrix R_2



Pic. 6. The bias as a function of sample size n for correlation matrix R_3

CONCLUSIONS

All simulations were carried out assuming strong dependency between the variable under study and response probabilities, which is highly unwelcome from the estimation viewpoint. Both regression and weighting-adjustment estimators allow to reduce bias and improve accuracy provided that respective model fits the data well. The proposed composite estimator reduces the bias and improves accuracy when any of these two models fits well.

REFERENCES

Bethlehem J.G. (1988) Reduction of Nonresponse Bias Through Regression Estimation *Journal of Official Statistics* Vol 4. No. 3. 1988, 251-160.

Ekholm A. Laaksonen S. (1991) Weighting via Response Modelling in the Finnish Household Budget Survey, *Journal of Official Statistics* Vol 7. No 3. 325-338.

Minka T.P.(2001) *Algorithms for Maximum Likelihood Logistic Regression* Technical Report URL: http://www.stat.cmu.edu/tr/tr758/tr758.pdf.

Rizzo L. Kalton G. Brick J.M. (1996) A Comparison of Some Weighting Adjustment Methods for Panel Nonresponse, *Survey Methodology*. Vol 22. No 1. 43-53.

Wojciech Gamrot

O PEWNYM ESTYMATORZE ZŁOŻONYM ŚREDNIEJ W POPULACJI

W artykule zaproponowno estymator złożony średniej w populacji skończonej przy brakach odpowiedzi. Jest on kombinacją estymatora regresyjnego opartego na modelu liniowym i estymatora wykorzystującego ważenie danych opartego na modelu logistycznym. Wagi kombinacji uzależniono od miar dobroci dopasowania tych modeli do danych. Przedstawiono wyniki symulacji wykonanych dla zbadania jego własności.