ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 194, 2005

*Wojciech Gamrot**

ON APPLICATION OF LOGISTIC REGRESSION TO MEAN VALUE ESTIMATION IN TWO-PHASE SAMPLING FOR NONRESPONSE

Abstract

The phenomenon of nonresponse in a sample survey usually leads to bias in estimates of population parameters. One of the techniques applied as a countermeasure for nonresponse is based on two-phase (or double) sampling. Usually a linear combination of mean value estimates obtained in both phases of the survey is used as an estimate of population mean value of the characteristic under study. In this paper alternative estimators for two-phase sampling scheme using estimates of response probabilities obtained on the basis of logistic regression model are considered. The results of Monte Carlo simulation study comparing the properties of these estimators are presented. In the simulations, the data from the Polish 1996 Agricultural Census were used.

Key words: stochastic nonresponse, response probabilities, logistic regression, mean value estimation.

I. INTRODUCTION – TWO-PHASE SAMPLING

Let us assume that the mean value $\overline{Y}$ of some characteristic Y in the population U of the size N is to be estimated and to accomplish this a simple random sample s of the size n is drawn without replacement from U, according to the sampling design $p(\cdot)$, given by formula:

$$\mathrm{P}(s) = \binom{N}{n}^{-1}; \tag{1}$$

We admit, as in the paper of Cassel et al. (1983), that nonresponse mechanism is a stochastic one. This means that each i-th population unit has some

* Ph.D., Department of Statistics, University of Economics in Katowice.

unknown probability ρ_i of responding if it is included in the sample. Hence, the phenomenon of nonresponse may be treated as an additional phase of sample selection, governed by some unknown probability distribution $q(s_2|s)$. Särndal et al. (1992) call the probability distribution $q(s_2|s)$ the *response distribution.*

Depending on response probabilities, during an attempt to collect data some units respond and some do not. Hence, the sample s can be divided into two sets s_1 and s_2, of the sizes $0 \leqslant n_1 \leqslant n$ and $0 \leqslant n_2 \leqslant n$ containing responding and non-responding units respectively. Consequently, we have $s_1 \cup s_2 = s$, $s_1 \cap s_2 = \varnothing$ and $n_1 + n_2 = n$. Sizes n_1 and n_2 are random variables, whose distributions depend on unknown response probabilities. As it has been shown by Lessler and Kalsbeek (1992), estimates of population mean based only on observations from the set s_1 will be biased. To reduce bias, a subsampling scheme proposed by Hansen and Hurwitz (1946) may be applied. According to this scheme, a second phase of the survey is performed. In this second phase, a simple subsample u of the size $n_u = cn_2$ (where $0 < c \leqslant 1$ is a constant fixed in advance) is selected without replacement from among n_2 units of the set s_2. The probability of selection for a certain subsample may be expressed as:

$$P_2(u|n_2) = \binom{n_2}{n_u}^{-1}; \tag{2}$$

All units included in the subsample are then re-contacted, and we assume that appropriate effort is made to obtain data for all subsampled units. Hence, under this procedure the probabilities ρ_i apply only to the first phase of the survey.

II. CLASSICAL MEAN VALUE ESTIMATOR IN TWO-PHASE SAMPLING

As it has been shown by Särndal et al., (1992), for the sampling scheme described above and stochastic nonresponse, the following statistic is an unbiased estimator of population mean:

$$\bar{y}_s = w_1 \bar{y}_1 + w_2 \bar{y}_u \tag{3}$$

where $\bar{y}_1 = \frac{1}{n_1} \sum_{i \in s_1} y_i$, $\bar{y}_2 = \frac{1}{n_2} \sum_{i \in s_2} y_i$ and $\bar{y}_u = \frac{1}{n_u} \sum_{i \in u} y_i$ are mean values of the characteristic under study in the sets s_1, s_2 and u respectively, whereas

$w_1 = \frac{n_1}{n}$ and $w_2 = \frac{n_2}{n}$ denote fractions of sets s_1 and s_2 in the initial sample s. Let us also define: $S^2 = \frac{1}{N-1}\sum_{i \in U}(y_i - \overline{Y})^2$ and $S_{s_2}^2 = \frac{1}{n_2 - 1}\sum_{i \in s_2}(y_i - \overline{y}_2)^2$. The variance of $\overline{y}_s$ depends on unknown response probabilities, and can be expressed as:

$$V(\overline{y}_s) = \frac{N-n}{Nn}S^2 + E_p E_q\left(w_2 \frac{1/c - 1}{n} S_{s_2}^2 | s\right); \tag{4}$$

where E_p denotes expectation with respect to the first stage sampling design[1], and E_q denotes expectation with respect to the response distribution. In particular, when deterministic nonresponse appears, the population U may be divided into two strata U_1 and U_2 such that U_1 contains respondents and U_2 contains nonrespondents. Consequently, the variance may be expressed as:

$$V(\overline{y}_s) = \frac{N-n}{Nn}S^2 + W_2 \frac{1/c - 1}{n} S_{s_2}^2; \tag{5}$$

where W_2 is the population nonrespondent fraction and S_2^2 is the variance of the characteristic under study in nonrespondent stratum U_2. In further study, the estimator $\overline{y}_s$ will be denoted by the symbol S.

III. ALTERNATIVE MEAN VALUE ESTIMATOR USING AUXILIARY INFORMATION

The estimator (3) is constructed as a linear combination of mean value estimates obtained in both phases of survey. In the case of deterministic nonresponse, the weights of this combination may be treated as the estimates of respondent and nonrespondent stratum fractions W_1 and W_2. However, there are other possible ways to construct these weights on the basis of available auxiliary information, which lead to the estimator:

$$\overline{y}_L = \alpha \overline{y}_1 + (1 - \alpha)\overline{y}_u \tag{6}$$

In particular, Wywiał (2001a) suggests to assess the parameter α as:

$$\hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N} \hat{\rho}_i \tag{7}$$

[1] Here: simple random sampling without replacement.

where $\hat{\rho}_i$ is the estimate of individual response probability ρ_i for the i-th unit. The estimates $\hat{\rho}_i$ are obtained by assuming that for any population unit the probability of response is given by the following function of auxiliary variables:

$$\rho_i = \frac{1}{1 + e^{\beta x_i}}; \quad i = 1...N; \tag{8}$$

where $\boldsymbol{\beta} = [\beta_0...\beta_k]$ denotes the vector of unknown parameters and $\mathbf{x}_i = [x_{i0}...x_{ik}]^T$ denotes the vector of auxiliary variables corresponding to the i-th population unit. We will assume that $x_{i0} = 1$ for $i = 1...N$, which means that β_0 may be treated as intercept. Assume that $J_i = 1$ if i-th unit responds and $J_i = 0$ if it does not. The parameters β can be estimated on the basis of the response behaviour observed in the initial sample s by minimizing the likelihood function (see Chow, 1995):

$$L = \prod_{i \in s} \left[\left(\frac{1}{1 + e^{\beta x_i}} \right)^{J_i} \left(\frac{e^{\beta x_i}}{1 + e^{\beta x_i}} \right)^{1 - J_i} \right]; \tag{9}$$

This is equivalent to the maximization of log-likelihood:

$$\log(L) = \sum_{i \in s_1} \ln \left(\frac{1}{1 + e^{\beta x_i}} \right) + \sum_{i \in s_2} \ln \left(\frac{e^{\beta x_i}}{1 + e^{\beta x_i}} \right) \tag{10}$$

Assuming that partial derivatives of this expression with respect to parameters β_i equal to zero we obtain the system of nonlinear equations, whose solution $\hat{\beta}$ is treated as an estimate of the parameter vector β (see Theil, 1979). The solution $\hat{\beta}$ may be found by using iterative methods, discussed e.g. by Minka (2001). In particular, a gradient projection method proposed by J.S. Rosen (1969) may be used. Finally, substituting $\hat{\beta}$ instead of β in the formula (8) it is possible to compute the estimates $\hat{\rho}_i$ of individual response probabilities, the parameter $\hat{\alpha}$, and the mean value estimate $\bar{y}_L$. In general, the estimator $\bar{y}_L$ is biased, but we may expect its variance to be significantly lower than the variance of $\bar{y}_S$. In further study the estimator $\bar{y}_L$ will be denoted by the symbol L.

In this paper a modified version of this estimator based on rounded (discretized) response probabilities $\hat{\rho}'_i$ is also proposed. The discretization is achieved by transforming the estimates of response probabilities according to the formula:

$$\hat{\rho}_i' = \begin{cases} 1 & \text{for} \quad \hat{\rho}_i \geqslant 0.5 \\ 0 & \text{for} \quad \hat{\rho}_i < 0.5 \end{cases} \tag{11}$$

and then applying transformed values instead of $\hat{\rho}_i$ in expressions (7) and (6). Such approach is in some sense justified, when it is expected that nonresponse mechanism is deterministic, or approximately deterministic. In such case, the proposed procedure resembles the application of discrimination method, to divide population into clusters of respondents and nonrespondents, and assess their population proportions, as proposed by Wywiał (2001b). In further study the modified version of the estimator $\overline{y}_L$ will be denoted by the symbol R.

It is worth noting that assumptions similar to the one given by (8) making use of multivariate logistic curve are often considered in the context of nonresponse modelling (see Cassel et al., 1983; Ekholm and Laaksonen, 1991; or Gao et al., 2000), but they typically constitute the basis for construction of weighting adjustments in single-phase sampling.

IV. ANOTHER ESTIMATOR INVOLVING ESTIMATES OF RESPONSE PROBABILITIES – AN EXTENSION OF THE SINGLEPHASE WEIGHTING ADJUSTMENT

The typical way the estimates of response probabilities are used in single-phase sampling is to construct individual weights, for each observation. As indicated by Särndal et al. (1992) and Bethlehem (1988), for an arbitrary sampling design the weight for i-th unit is usually set to $1/\pi_i\hat{\rho}_i$ where π_i is the inclusion probability associated with this unit. In the case of simple random sampling without replacement, the inclusion probability of the first order is equal to n/N so the mean value estimator takes the form:

$$\overline{y}_{1f} = \frac{1}{N} \sum_{i \in s_1} \frac{y_i}{n\hat{\rho}_i} \tag{12}$$

The use of estimates $\hat{\rho}_i$ instead of exact response probabilities introduces some bias, but we may hope that this bias is modest if response probabilities are estimated with sufficient accuracy. However, this estimator cannot be used if any of the estimates $\hat{\rho}_i$ is equal to zero. We propose the way to overcome this obstacle using the two-phase sampling procedure described above. Let us note that the (conditional) response probability for any i-th unit included in the first-phase sample s may be expressed as:

$$\rho_i' = \rho_i + c(1 - \rho_i); \tag{13}$$

where ρ_i is the probability of this unit responding at the first phase and $c(1-\rho_i)$ represents the probability of this unit not responding at the first phase, but being included in the subsample and consequently responding at the second phase. This allows to rewrite the estimator (12) in the form:

$$\overline{y}_W = \sum_{i \in s_1 \cup u} \frac{y_i}{n(\hat{\rho}_i + c(1-\hat{\rho}_i))}; \tag{14}$$

It is easy to notice that the expression in denominator is always positive, provided that $c > 0$. Let us also stress that sampling units from both set s1 and subsample u are used in computation of mean value estimates. For this estimator the logistic regression model will be used again as a means of estimating response probabilities. In further study the estimator $\overline{y}_w$ will be denoted by the symbol *W*.

V. COMPARISON OF ESTIMATORS BY MEANS OF MONTE CARLO SIMULATION

A simulation study was performed to compare the accuracy of the four estimators presented above. The data obtained from Polish Agricultural Census in 1996 for certain municipalities of the Dąbrowa Tarnowska district represented the population under study during simulations. The total of 2422 units were used in simulation. The variable under study, denoted by Y was total sales of the farm in the year 1995. The auxiliary variables were the farm area (in acres) – X_1, the number of pigs in the farm – X_2, and the number of cattle stock in the farm – X_3. The Pearson linear correlation coefficients between these variables are shown in the following table:

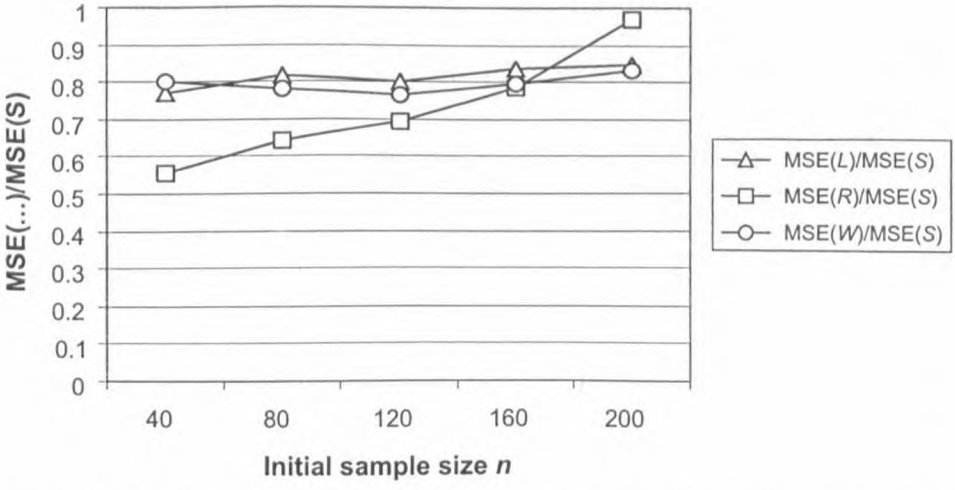
Table 1. Correlation coefficients between variable under study and auxiliary variables

	Y	X_1	X_2	X_3
Y	1	0.63	0.52	0.50
X_1	0.63	1	0.58	0.67
X_2	0.52	0.58	1	0.62
X_3	0.50	0.67	0.62	1

For every population unit the response probabilities were generated according to the model given by expression (8), and predefined parameter vector β. The experiments were carried out by repeatedly drawing without

replacement simple random samples from the population. To represent the stochastic nonresponse mechanism, for each unit included in any sample an independent random trial was executed with the probability of success equal to this unit's response probability. A unit was assumed to respond if the outcome of the trial was a success and treated as nonrespondent otherwise. For the resulting set of nonrespondents a simple subsample of the size equal to the 30% of the first-phase nonrespondent number was drawn without replacement, and all the four estimators denoted by letters S, L, R, W were computed. On the basis of computed estimates, the mean square error of each estimator was evaluated.

Experiment 1. In the first experiment only one auxiliary variable X_1 was used. Response probabilities were generated for an arbitrarily chosen parameter vector $\beta = [-4, 0.003]$. Consequently, the average response probability in the population was 0.89. Simulations were executed for the sample size $n = 40, 80, ..., 200$. For every value of n a total of 10 000 samples were drawn from the population. The relative accuracy (the proportion of MSE of any estimator to the MSE of the standard estimator S) is shown on the Graph 1. Each point on the graph therefore corresponds to 10 000 computed estimates.



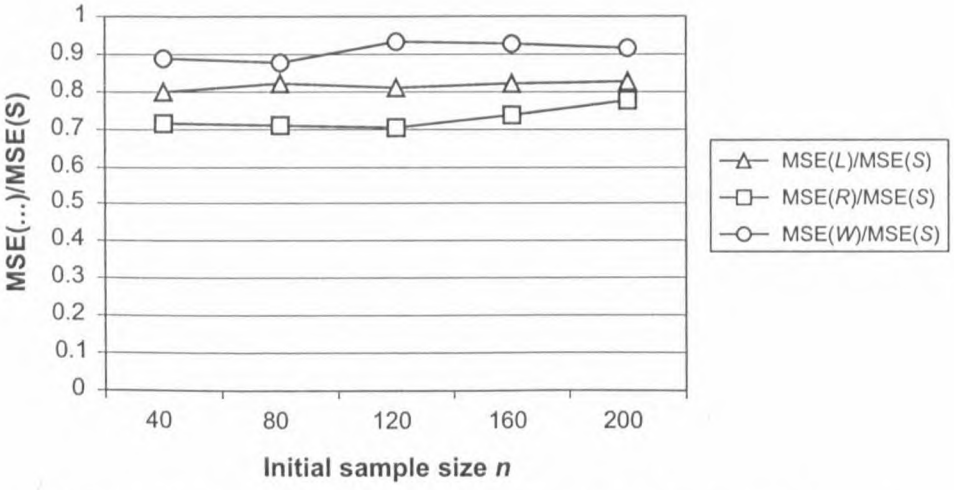
Graph 1. The dependence between the initial sample size n and relative accuracy of the estimators for single auxiliary variable

As it can be seen on the graph, for any value of n for which the simulations were executed, each of the estimators: L, R and W had lower MSE than the standard two-phase estimator S. For small initial sample sizes the estimator R was the best (in terms of MSE). However, the MSE

of this estimator grew rapidly, and for larger samples the estimators L and W were more accurate than R. The results suggest that for the sample size large enough, the MSE of estimator R may exceed the MSE of standard estimator S. The relative accuracy of the strategies L and W was rather stable and it was approximately equal to 80%. In most cases, the strategy W had lower MSE than the strategy L. It is worth noting that for small sample sizes the estimator R behaves well, despite non-deterministic character of response mechanism.

Experiment 2. In the second experiment three auxiliary variables X_1, X_2 and X_3 were used. Response probabilities were generated for an arbitrarily chosen parameter vector $\beta = [-6, 0.002, 0.146, 0.348]$, with values $\beta_1 ... \beta_3$ inversely proportional to the mean values of corresponding auxiliary variables. Consequently, the average response probability in the population was 0.85. Simulations were executed for the sample size $n = 40, 80, ..., 200$. For every value of n a total of 10 000 samples were drawn from the population. The relative accuracy (the proportion of MSE of any estimator to the MSE of the standard estimator S) is shown on the Graph 2. Each point on the graph therefore corresponds to 10 000 computed estimates.

In this experiment, again each of the estimators: L, R and W had lower MSE than the standard two-phase estimator S, for any value of n for which the simulations were executed. The lowest MSE was observed for the estimator R, and the highest for the estimator W.



Graph 2. The dependence between the initial sample size n and relative accuracy of the estimators for three auxiliary variables

It's worth noting that the addition of two auxiliary variables did not influence significantly the relative accuracy of the estimator L, but the behavior of two other estimators changed. The MSE of the estimator R grows slower with the increase of initial sample size, whereas the MSE of the estimator W is stable, but greater than it was for only one auxiliary variable, probably because of weaker correlation between variable X_1 (X_2) and the variable under study.

VI. SUMMARY

In general, the estimators considered in this paper are biased. For the price of bias, lower values of MSE may be achieved. It should be stressed however that, in order to improve the MSE suitable auxiliary information is needed. In the case of estimators L and R auxiliary characteristics have to be observed for all population units to compute the estimates. For the estimator W auxiliary characteristics should only be observed for all the units included in the initial sample, so this estimator may be applied in situations where the estimators R and L are not applicable due to lack of data on auxiliary characteristics in the whole population. Moreover, the simulation results presented here are based on assumption, that functional form of response mechanism is known. In practice, such knowledge usually comes from sources external to the survey, and it may be inaccurate or simply false. In such case a model misspecification error occurs, and as a result the MSE of the estimators may be greater. When the functional form of response mechanism is unknown, some nonparametric methods may also be used to estimate the response probabilities.

REFERENCES

Bethlehem J.G. (1988), Reduction of non-response bias through regression estimation, *Journal of Official Statistics*, **4**, 251–260.

Cassel C.M., Särndal C.E., Wretman J.H. (1983), Some uses of statistical models in connection with the nonresponse problem, [in:] *Incomplete Data in Sample Surveys*, eds. W.G. Madow, I. Olkin, Academic Press, New York.

Chow G.C. (1995), *Ekonometria*, Wyd. Nauk. PWN, Warszawa.

Ekholm A., Laaksonen S. (1991), Weighting via response modelling in the finnish household budget survey, *Journal of Official Statistics*, 7, 3, 325–338.

Gao S., Hui S.L., Hall K.S., Hendrie H.C. (2000), Estimating disease prevalence from two-phase surveys with non-response at the second phase, *Statistics in Medicine*, No 19, 2101–2114.

Hansen M.H., Hurwitz W.N. (1946), The problem of nonresponse in sample surveys, *Journal of the American Statistical Society*, 517–529.

Lessler J.T., Kalsbeek W.D. (1992), *Nonsampling Error in Surveys*, John Wiley & Sons, New York.
Minka T.P. (2001), Algorithms for maximum likelihood logistic regression, technical report, Carnegie Mellon University http://www.stat.cmu.edu/tr/tr758/tr758.pdf
Rosen J.B. (1969), The gradient projection method for nonlinear programming, *Journal of Society for Industrial and Applied Mathematics*, **8**, 1, 181–217.
Särndal C.E., Swensson B., Wretman J.H. (1992), *Model Assisted Survey Sampling*, Springer Verlag, New York.
Theil H. (1979), *Zasady ekonometrii*, PWN, Warszawa.
Wywiał J. (2001a), Estimation of population mean on the basis of non-simple sample when non-response error is present, *Statistics in Transition*, **5**, 3, 443–450.
Wywiał J. (2001b), On estimation of population mean in the case when nonrespondents are present, *Taksonomia – Klasyfikacja danych, teoria i zastosowania*, No 8, Prace Naukowe AE Wrocław.

Wojciech Gamrot

O ZASTOSOWANIU REGRESJI LOGISTYCZNEJ DO OCENY WARTOŚCI PRZECIĘTNEJ Z WYKORZYSTANIEM SCHEMATU LOSOWANIA DWUFAZOWEGO W PRZYPADKU BRAKÓW ODPOWIEDZI

Streszczenie

Wystąpienie niekompletności obserwacji badanej cechy w badaniu statystycznym zazwyczaj prowadzi do obciążenia uzyskanej oceny badanego parametru populacji. Jedna z technik stosowanych dla przeciwdziałania temu zjawisku opiera się na wykorzystaniu schematu losowania dwufazowego. Jako estymator wartości przeciętnej w populacji wykorzystuje się zazwyczaj kombinację liniową ocen wartości przeciętnej uzyskanych w pierwszym i drugim etapie (fazie) badania. W niniejszym referacie podjęto próbę zbadania własności alternatywnych strategii estymacji wykorzystujących schemat losowania dwufazowego, uwzględniających w konstrukcji estymatora oceny prawdopodobieństw uzyskania odpowiedzi od poszczególnych jednostek populacji uzyskane przy wykorzystaniu modelu regresji logistycznej. Porównanie własności wymienionych strategii przeprowadzono drogą symulacji komputerowej, przy wykorzystaniu danych uzyskanych w wybranych gminach powiatu Dąbrowa Tarnowska podczas spisu rolnego w roku 1996.