Mathematical Modelling and Analysis
Volume 12 Number 3, 2007, pages 369–377
© 2007 Technika ISSN 1392-6292 print, ISSN 1648-3510 online

MODEL–BASED ESTIMATOR OF TOTAL EXPENDITURE ON ENVIRONMENTAL PROTECTION

D. KRAPAVICKAITĖ

Institute of Mathematics and Informatics,
Vilnius Gediminas Technical University
Akademijos 4, LT-08663 Vilnius, Lithuania,
Saulėtekio av. 11, LT-10223 Vilnius, Lithuania
E-mail: krapav@ktl.mii.lt

Received October 04, 2005; revised March 31, 2007; published online September 15, 2007

Abstract. Expenditure of enterprises on environmental protection has highly positive or zero values. Thus, distribution of such a variable is skewed, and its population variance is high. As a consequence the design-based estimator of a total expenditure has a high variance. The censored regression model (*tobit* model) for transformed study variable is used in this paper. The values of the study variable of non-sampled elements are estimated using this model, and a new estimator of the total expenditure is obtained. In the case of simple random sampling and known model parameters, it is shown by simulation that the empirical mean square error of such a model-based estimator is lower than that of the design-based one. Due to estimation of the unknown model parameters the mean square error of the estimator is higher.

Key words: finite population, model-based estimator of a total, tobit model

1. Introduction

The variable of expenditure on environmental protection is quite irregular: some of the enterprises have high expenditure and some of them have none at all. Thus, distribution of such a variable is skewed with the peak at zero and has a high population variance (Fig. 1). For this reason the design-based Horvitz-Thompson estimator [8] of the total of such a variable has a high variance as well. We are looking for another way of estimating the total.

One of the possible ways to solve this problem is to use the model-based approach of a study variable. In this case the values of variables of the elements of a finite population $\mathcal{U}$ are assumed to be generated according to some super-population model. The non-sampled values of the study variable are

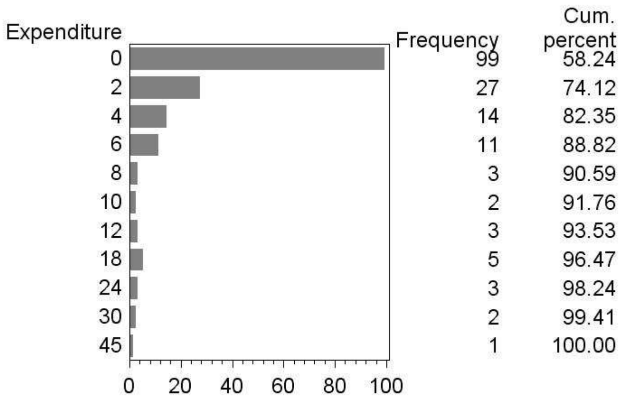

Figure 1. Expenditure to the environment protection.

predicted by this model and used for the estimation of the total. If we can find a distribution model that closely resembles the distribution of the study variable, a model based-estimator may have a smaller mean square error than a design-based one. This way of estimation is presented in Valliant et. al. [10]. It was used to estimate the total of the non-negative variable with zero values by Karlberg [5]. The study variable is expressed there as a product of two variables: nonzero and binary.

We shall use a new approach and apply a model which describes positive and zero values of logarithmic transformation of the study variable–censored regression or a tobit model. The tobit model has been introduced by Tobin [9] in 1958, developed by Amemiya [1], Green [3] and widely discussed in Amemiya [2], Green [4], Madalla [6], Magnus et. al. [7]. We suggest some application of the tobit model in the estimation of the finite population total of a non-negative study variable.

2. Design-Based Estimator of a Total

Let us denote by $\mathcal{U} = \{1, 2, \ldots, N\}$ the finite population, consisting of N elements, and by y the study variable with the values y_k, $k = 1, \ldots, N$, defined for the elements of the population, correspondingly; the total of this variable is defined by $t_y = y_1 + \cdots + y_N$. The n size sampling design is said to be simple random sampling if any collection of n distinct elements from the population $\mathbf{i}$, $\mathbf{i} \subset \mathcal{U}$, has the same probability to be selected. The simple random sampling design can be obtained when n elements from the finite population are taken with equal selection probabilities without replacement.

The Horvitz-Thompson estimator of the population total

$$\hat{t}_y^{(SI)} = \frac{N}{n} \sum_{k \in \mathbf{i}} y_k$$

is unbiased under simple random sampling [8]. The unbiased estimator of the variance of this estimator is used:

$$\widehat{Var}(\hat{t}_y) = N^2 \Big(1 - \frac{n}{N}\Big) \frac{\hat{s}^2}{n},$$

$$\hat{s}^2 = \frac{1}{N-1} \sum_{k \in \mathbf{i}} (y_k - \bar{y})^2, \quad \bar{y} = \frac{1}{n} \sum_{k \in \mathbf{i}} y_k.$$

We restrict ourselves to the simple random sampling in the paper.

3. Model-Based Estimator of a Total

Let x be an auxiliary variable, and its values are denoted by x_k, $k = 1, \dots, N$, they are known for all the population elements. Let $\mathbf{i}$, $\mathbf{i} \subset \mathcal{U}$, be a probability sample of size n and $\bar{\mathbf{i}} = \mathcal{U} \setminus \mathbf{i}$ be a subset of the non-sampled elements of population. The values y_k are known only for the elements k that belong to the sample: $k \in \mathbf{i}$.

The parameter of interest is the population total $t_y = \sum_{k=1}^N y_k$. It can be expressed as a sum $t_y^{(\mathbf{i})}$ of the sampled elements and a sum $t_y^{(\bar{\mathbf{i}})}$ of the non-sampled ones:

$$t_y = t_y^{(\mathbf{i})} + t_y^{(\bar{\mathbf{i}})} = \sum_{k \in \mathbf{i}} y_k + \sum_{k \in \bar{\mathbf{i}}} y_k.$$

Here the value of $t_y^{(\mathbf{i})}$ is known from the sampled data. A super-population model for y_k, $k \in \mathcal{U}$ has been built. This model is used to construct the estimators $\hat{y}_k$, $k \in \bar{\mathbf{i}}$. Then $t_y^{(\bar{\mathbf{i}})}$ is estimated by $\hat{t}_y^{(\bar{\mathbf{i}})} = \sum_{k \in \bar{\mathbf{i}}} \hat{y}_k$ and the population total t_y is estimated by

$$\hat{t}_y^{(model)} = t_y^{(\mathbf{i})} + \hat{t}_y^{(\bar{\mathbf{i}})}. \tag{3.1}$$

Some special model for the variable y with zero and positive values is used.

3.1. Tobit model of a study variable

Let us denote by $\mathbf{x}_k = (1, x_k)'$, values of the known non-random auxiliary vector $\mathbf{x}$ and by z_k^*, $k = 1, \dots, N$, values of some unobserved random variable z^*. Suppose the values z_k^*, can be written in the form

$$z_k^* = \mathbf{x}_k' \boldsymbol{\beta} + \varepsilon_k, \tag{3.2}$$

i.e. it depends on the unknown regression model parameter $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ and error terms $\varepsilon_k \sim \mathcal{N}(0, \sigma^2)$, which are independent. Let z_k be expressed through the values z_k^* of an unobserved variable z^* by

$$z_k = \begin{cases} z_k^*, & \text{if } z_k^* \geq 0, \\ 0, & \text{if } z_k^* < 0, \end{cases} \qquad k = 1, \ldots, N. \tag{3.3}$$

The model given by equations (3.2) and (3.3) and associating z_k with z_k^* is called a censored regression model or a *tobit* model of the variable z.

The conditional mean of z_k is equal to (see [4]):

$$\mathbf{E}(z_k|\mathbf{x}_k) = \mathbf{x}_k'\boldsymbol{\beta}\varPhi\Big(\frac{\mathbf{x}_k'\boldsymbol{\beta}}{\sigma}\Big) + \sigma\phi\Big(\frac{\mathbf{x}_k'\boldsymbol{\beta}}{\sigma}\Big), \quad k = 1, \ldots, N. \tag{3.4}$$

Here $\varPhi(\cdot)$ is a standard normal distribution function and $\phi(\cdot)$ is its density function.

Let us consider a study variable that acquires non-negative values y_k, such that $z_k = \ln(y_k + 1),\ k = 1, \ldots, N$.

Proposition 1. *The conditional mean of a non-negative variable y with the values y_k, $k = 1, \ldots, N$, in a finite population, defined by*

$$z_k = \ln(y_k + 1) \tag{3.5}$$

for z satisfying (3.2), (3.3), *is equal to*

$$\mathbf{E}(y_k|\mathbf{x}_k) = \mathrm{e}^{\mathbf{x}_k'\boldsymbol{\beta}+\sigma^2/2}\,\varPhi\Big(\frac{\mathbf{x}_k'\boldsymbol{\beta}}{\sigma} + \sigma\Big) - \varPhi\Big(\frac{\mathbf{x}_k'\boldsymbol{\beta}}{\sigma}\Big). \tag{3.6}$$

Proof. Let us express

$$\mathbf{E}(y_k|\mathbf{x}_k) = \mathbf{E}\big(\mathrm{e}^{z_k^*} - 1|z_k^* > 0\big)\,\mathbf{P}\big(z_k^* > 0\big) + \mathbf{E}\big(0|z_k^* \leq 0\big)\,\mathbf{P}\big(z_k^* \leq 0\big).$$

By virtue of (3.2) and (3.3)

$$\mathbf{E}(y_k|\mathbf{x}_k) = \frac{1}{\sigma\sqrt{2\pi}}\int_0^\infty (\mathrm{e}^u - 1)\mathrm{e}^{-\frac{(u-\mathbf{x}_k'\boldsymbol{\beta})^2}{2\sigma^2}}\,\mathrm{d}u.$$

By integrating we obtain

$$\begin{aligned} \mathbf{E}(y_k|\mathbf{x}_k) &= \frac{1}{\sqrt{2\pi}}\int_{-\frac{\mathbf{x}_k'\boldsymbol{\beta}}{\sigma}}^\infty \Big(\mathrm{e}^{\sigma a+\mathbf{x}_k'\boldsymbol{\beta}} - 1\Big)\mathrm{e}^{-\frac{a^2}{2}}\,\mathrm{d}a \\ &= \frac{1}{\sqrt{2\pi}}\mathrm{e}^{\mathbf{x}_k'\boldsymbol{\beta}}\int_{-\frac{\mathbf{x}_k'\boldsymbol{\beta}}{\sigma}}^\infty \mathrm{e}^{\sigma a-\frac{a^2}{2}}\,\mathrm{d}a - \frac{1}{\sqrt{2\pi}}\int_{-\frac{\mathbf{x}_k'\boldsymbol{\beta}}{\sigma}}^\infty \mathrm{e}^{-\frac{a^2}{2}}\,\mathrm{d}a \\ &= \mathrm{e}^{\mathbf{x}_k'\boldsymbol{\beta}+\frac{\sigma^2}{2}}\varPhi\Big(\frac{\mathbf{x}_k'\boldsymbol{\beta}}{\sigma^2/2} + \frac{\sigma^2}{2}\Big) - \varPhi\Big(\frac{\mathbf{x}_k'\boldsymbol{\beta}}{\sigma^2/2}\Big). \end{aligned}$$

■

4. Estimation of the Total

For the known parameters $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ and σ^2 of the model (3.2), (3.3), (3.5) let us estimate the values of y of non-sampled elements from (3.6) by

$$\hat{y}_k^{(ltobit)} = \mathrm{e}^{\mathbf{x}_k'\boldsymbol{\beta}+\frac{\sigma^2}{2}}\,\varPhi\Big(\frac{\mathbf{x}_k'\boldsymbol{\beta}}{\sigma^2/2}+\frac{\sigma^2}{2}\Big) - \varPhi\Big(\frac{\mathbf{x}_k'\boldsymbol{\beta}}{\sigma^2/2}\Big).$$

Replacing $\hat{y}_k$ by $\hat{y}_k^{(ltobit)}$ in (3.1), $k \in \bar{\mathbf{i}}$, we get the new unbiased estimator of the population total:

$$\hat{t}_y^{(ltobit)} = t_y^{(\mathbf{i})} + \hat{t}_y^{(\bar{\mathbf{i}})} = \sum_{k\in\mathbf{i}} y_k + \sum_{k\in\bar{\mathbf{i}}} \hat{y}_k^{(ltobit)}. \tag{4.1}$$

In the case where the model parameters $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ and σ^2 are unknown they are estimated from the sample data for the tobit model (3.2), (3.3), using the maximum likelihood method. Their estimators $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)'$ and $\hat{\sigma}^2$ are used to estimate the values y_k of non-sampled elements:

$$\hat{\hat{y}}_k^{ltobit} = \mathrm{e}^{\mathbf{x}_k'\hat{\boldsymbol{\beta}}+\hat{\sigma}^2/2}\varPhi\Big(\frac{\mathbf{x}_k'\hat{\boldsymbol{\beta}}}{\hat{\sigma}}+\hat{\sigma}\Big) - \varPhi\Big(\frac{\mathbf{x}_k'\hat{\boldsymbol{\beta}}}{\hat{\sigma}}\Big).$$

Replacing $\hat{y}_k$ by $\hat{\hat{y}}_k^{(ltobit)}$ in (3.1), $k \in \bar{\mathbf{i}}$, we get the estimator of the population total

$$\hat{\hat{t}}_y^{(ltobit)} = t_y^{(\mathbf{i})} + \hat{\hat{t}}_y^{(\bar{\mathbf{i}})} = \sum_{k\in\mathbf{i}} y_k + \sum_{k\in\bar{\mathbf{i}}} \hat{\hat{y}}_k^{(ltobit)}. \tag{4.2}$$

This estimator is not unbiased. The properties of the estimators proposed will be investigated by simulation in the next section.

5. Simulation Study

The simulation was performed using two populations: artificial and real.

The size of the artificial population is $N = 1\,000$. The values of the variable x were simulated by the normal distribution: $x_k \sim \mathcal{N}(3.5,\ 1.6)$, the error term ε by $\varepsilon_k \sim \mathcal{N}(0,\ 0.8)$, $k = 1, \ldots, N$. The constant vector is chosen $\boldsymbol{\beta} = (\beta_0, \beta_1)' = (-3.45, 1.1)'$. Then the values of z_k^* and z_k are constructed by (3.1) and (3.2). The values of y_k are obtained by $y_k = \mathrm{e}^{z_k} - 1$, $k = 1, \ldots, N$. Such a variable y has 41% of values equal to 0 in the population. The correlation coefficient between z and x equals 0.75 for $z_k > 0$.

The population of enterprises of the clothing manufacture activity of $N = 187$ elements is studied as a real population. The study variable y (Fig. 1) is expenditure on environmental protection. It has 53% of zeroes in the population. It is transformed using $z = \ln(y + 1)$. The logarithm of the income is taken as an auxiliary variable x with the values x_k. The correlation coefficient between z and x equals 0.58 for $z_k > 0$. The distribution of z in both cases is given in Fig. 2.

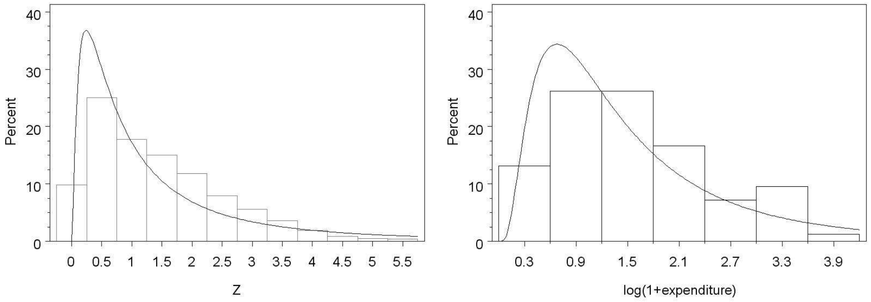

Figure 2. Distribution of logarithms of the positive values of the study variable in the artificial population (left) and the real population (right).

1 000 simple random samples of size $n = 20, 40, 100$ are drawn from both populations. The procedure *lifereg* of a statistical program package SAS has been used for estimation of the tobit model parameters by maximum likelihood method. The estimates $\hat{t}_y^{(SI)}$, $\hat{t}_y^{(ltobit)}$ and $\hat{\hat{t}}_y^{(ltobit)}$ have been calculated in each case. For all the estimators $\hat{t}_y = \hat{t}_y^{(SI)}$, $\hat{t}_y^{(ltobit)}$ and $\hat{\hat{t}}_y^{(ltobit)}$, the average of the estimates $\bar{\hat{t}}_y = (\sum_{i=1}^{1\,000} \hat{t}_{yi})/1\,000$ (replicates of $\hat{t}_y$ are denoted by $\hat{t}_{yi}$), empirical variance of the estimates

$$\widehat{Var}(\hat{t}_y) = \frac{1}{1\,000} \sum_{i=1}^{1\,000} (\hat{t}_{yi} - \bar{\hat{t}}_y)^2,$$

empirical standard deviation $Std(\hat{t}_y) = \sqrt{\widehat{Var}(\hat{t}_y)}$, empirical bias $\widehat{Bias}(\hat{t}_y) = \bar{\hat{t}}_y - t_y$, relative mean square error

$$RMSE(\hat{t}_y) = \frac{\sqrt{\widehat{Var}(\hat{t}_y) + \widehat{Bias}^2(\hat{t}_y)}}{\bar{\hat{t}}_y},$$

and skewness

$$Skew = \frac{M_3(\hat{t}_y)}{Std^3(\hat{t}_y)}, \quad M_3(\hat{t}_y) = \frac{1}{999} \sum_{i=1}^{1\,000} (\hat{t}_y - \bar{\hat{t}}_y)^3$$

have been calculated. The results of simulation are presented in Table 1.

Comments on the Simulation Results

1. Artificial population.

1.1. Comparison of $\hat{t}_y^{(ltobit)}$ and $\hat{t}_y^{(SI)}$. The empirical bias of $\hat{t}_y^{(ltobit)}$ is larger than $\hat{t}_y^{(SI)}$, while the empirical standard deviation is lower, and the estimate of the relative mean square error is less. This effect is stronger for the small

Table 1. Simulation results.

Case	Population	Sample size n	t_y	$\bar{\hat{t}}_y$	$\widehat{Bias}(\hat{t}_y)$	$Std(\hat{t}_y)$	$RMSE(\hat{t}_y)$	Skewness
Model parameters known $\hat{t}_y = \hat{t}_y^{(ltobit)}$	Artificial	20	4 695	4 791	96	55	0.02	0.81
		40		4 792	97	80	0.03	0.70
		100		4 782	87	122	0.03	0.25
	Real	20	521	622	102	24	0.17	0.67
		40		605	84	28	0.15	0.37
		100		567	47	36	0.10	−0.08
Model parameters unknown $\hat{t}_y = \hat{\hat{t}}_y^{(ltobit)}$	Artificial	20	4 695	5 743	1 048	4546	0.99	4.44
		40		5 095	400	2043	0.44	2.11
		100		4 908	213	1 137	0.25	0.93
	Real	20	521	690	169	594	1.19	12.01
		40		615	95	194	0.41	0.96
		100		567	47	74	0.17	−0.10
Simple random sampling $\hat{t}_y = \hat{t}_y^{(SI)}$	Artificial	20	4 695	4 623	− 72	3 518	0.76	1.83
		40		4 757	62	2 420	0.51	1.38
		100		4 665	− 30	1 519	0.33	0.73
	Real	20	521	544	24	249	0.46	0.69
		40		511	− 10	146	0.29	0.32
		100		520	0	73	0.14	−0.04

sample size. The skewness of the distribution of $\hat{t}_y^{(ltobit)}$ is less than that of $\hat{t}_y^{(SI)}$, especially for the small sample size.

1.2. Comparison of $\hat{t}_y^{(ltobit)}$ and $\hat{\hat{t}}_y^{(ltobit)}$. If the model parameters are not known and they are estimated from the sample, the empirical bias, empirical standard deviation, relative mean square error and skewness are much higher than in the case where the model parameters are known.

1.3. Comparison of $\hat{\hat{t}}_y^{(ltobit)}$ and $\hat{t}_y^{(SI)}$. Despite that the empirical bias of $\hat{\hat{t}}_y^{(ltobit)}$ is higher, its empirical standard deviation and estimate of the relative mean square error become less than that of $\hat{t}_y^{(SI)}$ for an increasing sample size.

2. Real population. We note that the distribution of z_k for $z_k > 0$ cannot be approximated so well by the log-normal distribution (Fig. 2), and the correlation coefficient between z and x is not so high as in the case of the artificial population.

2.1. Comparison of $\hat{t}_y^{(ltobit)}$ and $\hat{t}_y^{(SI)}$ (see Fig. 3, 4). The empirical bias of $\hat{t}_y^{(ltobit)}$ is not much higher than that of $\hat{t}_y^{(SI)}$, but the standard deviation is significantly lower. For this reason, the relative mean square error is significantly lower than that of $\hat{t}_y^{(SI)}$. The skewness of both estimators is similar.

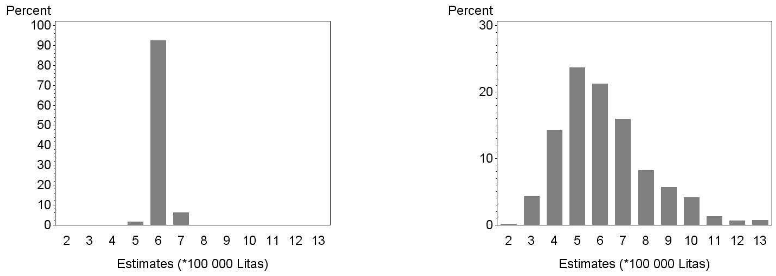

Figure 3. 1 000 estimates of size $n = 40$ from the real population with the known model parameters (left) and the estimated ones (right).

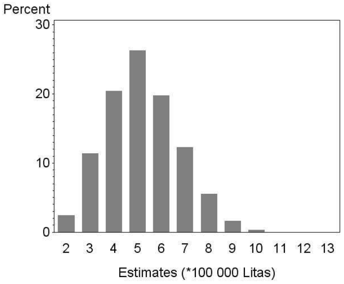

Figure 4. 1 000 simple random design-based estimates from the real population of size $n = 40$.

2.2. Comparison of $\hat{t}_y^{(ltobit)}$ and $\hat{\hat{t}}_y^{(ltobit)}$ (see Fig. 3). Due to the estimation of the model parameters the empirical bias of $\hat{\hat{t}}_y^{(ltobit)}$ is larger and its relative mean square error is much higher than in the case where the model parameters are known.

2.3. Comparison of $\hat{\hat{t}}_y^{(ltobit)}$ and $\hat{t}_y^{(SI)}$. Because of the estimation of the model parameters the empirical bias becomes a little bit higher, while the empirical standard deviation becomes significantly higher than that of $\hat{t}_y^{(SI)}$.

6. Conclusions

1. If the distribution of the study variable satisfies the assumptions of the model (3.2), (3.3), (3.5) and the parameters of this model are known, then the tobit model-based estimator has a lower relative mean square error than in the case of a simple random design-based estimator. If the model parameters are not known, then the tobit model-based estimator

is not unbiased and has a high mean square error. Some bias adjustment is needed. This is the case of most real surveys.

2. The tobit model-based estimator of a total is sensitive to the model assumptions. The artificial population is generated by the model, and the relative mean square error of the model-based estimator of the total with the known population parameters is lower than that for real data.
3. The proposed model (3.2), (3.3), (3.5) can be applied to estimate the total of expenditure on environmental protection in the case of the known model parameters.

Acknowledgment

The author is thankful to the student of Gediminas Technical University L. Ramanauskaitė for her input making simulations and to Statistics Lithuania for the opportunity to analyze the survey data.

References

[1] T. Amemiya. Regression analysis when the dependent variable is truncated normal. *Econometrica*, **41**(6), 997–1016, 1973.
[2] T. Amemiya. Tobit models: a survey. *Journal of Econometrics*, **24**, 3–61, 1984.
[3] W. H. Greene. On the asymptotic bias of the ordinary least squares estimator of the tobit model. *Econometrica*, **49**(2), 505–503, 1981.
[4] W. H. Greene. *Econometric Analysis.* Prentice Hall, Upper Saddle River, 2003.
[5] F. Karlberg. Survey estimation for highly skewed populations in the presence of zeroes. *Journal of Official Statistics*, **3**(2), 229–241, 2000.
[6] G. S. Maddala. *Limited dependent and qualitative variables in econometrics.* Cambridge University Press, Cambridge, 1983.
[7] Я. П. Магнус and П. К. Катышев and А. А. Пересецкий. *Эконометрика. Начальный курс.* Дело, Москва, 2000.
[8] C.-E. Särndal, B. Swensson and J. Wretman. *Model Assisted Survey Sampling.* Springer-Verlag, New York, 1992.
[9] J. Tobin. Estimation of relationships for limited dependent variables. *Econometrica*, **26**, 24–36, 1958.
[10] R. Valliant, A. H. Dorfman and R. M. Royall. *Finite Population Sampling and Inference.* A Prediction Approach. John Wiley & Sons, New York, 2000.