ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 225, 2009

*Krystyna Pruska**

ESTIMATION OF VARIANCE OF LOGISTIC REGRESSION PREDICTORS FOR SMALL AREAS

Abstract

Logistic regression models can be applied for analysis of Bernoulli variables in studies of small areas. In the paper the logistic regression predictors for parameter of the Bernoulli distribution and an estimation of variance and MSE for these predictors are considered for small areas. The results of simulation experiments conducted for analysis of properties of estimators of variance and mean square error for the predictors are presented.

Key words: logistic regression model, jackknife method, small area, simulation experiments.

1. Introduction

Frequently realizations of random samples are data sets, whose elements are zero's and one's. In conformity with the conception of superpopulation we can assume that the values are realizations of random variables which have the Bernoulli distribution. Logistic regression models can be applied for analysis of such data sets and distributions. In this case different problems appear, for example the estimation of predictor variance for different sample designs.

2. Logistic regression model for small areas

We consider a population which is divided into G strata and M small areas.

Let Y_i be a random variable investigated in i-th small area and let a distribution function of Y_i have the following form:

* Ph.D., Associate Professor, Chair of Statistical Methods, University of Łódź.

$$P(Y_i=1) = \theta_i, \tag{1}$$

$$P(Y_i=0) = 1 - \theta_i \tag{2}$$

where $0 < \theta_i < 1$ and $i = 1, ..., M$.

The parameter θ_i is called the proportion for i-th small area.

We consider some auxiliary variables, too. Let vector x_i denote these variables or their values for i-th small area.

We can construct the following logistic regression model (see Wiśniewski, 1986; Jajuga, 1990):

$$L^{-1}(p_i) = x_i'\alpha + \varepsilon_i \text{ for } i = 1, ..., M \tag{3}$$

where

$$L^{-1}(p_i) = ln \frac{p_i}{1-p_i} \tag{4}$$

and p_i is an estimator of parameter θ_i, α is the model parameter, ε_i is a random error, $E(\varepsilon_i) = 0$.

If we know estimates p_i for $i = 1, ..., m$ (for m small areas which are random sample drawn from a whole set of small areas) and x_i for all small areas, then we can estimate the parameter vector of model (3) and unknown probabilities θ_i for small areas which are undrawn to sample. The estimate of parameter θ_i obtained on the basis of model (3) is value of estimator which is called predictor. We may consider an estimation of variance of this predictor for different sampling designs.

3. Jackknife method

The jackknife method gives a possibility of estimation of bias and variance of parameter estimator (see Shao, Tu, 1996).

Let $X_1, ..., X_n$ be a random sample drawn from population. Let T_n be an estimator of population parameter θ and $T_{n-1,i}$ $(i = 1, ..., n)$ be analogous estimator of parameter θ determined on the basis of sequence $X_1, ..., X_{i-1}, X_{i+1}, ..., X_n$.

Let

$$\overline{T_n} = \frac{1}{n}\sum_{i=1}^{n} \overline{T}_{n-1,i} \tag{5}$$

$$b_{JACK} = (n-1)(\overline{T_n} - \overline{T_n}) \tag{6}$$

$$T_{JACK} = T_n - b_{JACK} = nT_n - (n-1)\overline{T_n} \tag{7}$$

$$v_{JACK} = \frac{n-1}{n}\sum_{i=1}^{n}(T_{n-1,i} - \overline{T_n})^2 \tag{8}$$

The estimator b_{JACK} is the estimator of bias of estimator T_n, it means $E(T_n) - \theta$. The estimator v_{JACK} is the estimator of variance of estimator T_n.

4. Simulation analysis of properties of variance estimators of logistic regression predictors for small areas

In order to investigate properties of variance of proportion predictor which are determined for small areas on the basis of logistic regression models a simulation analysis was conducted. Three populations A, B, C are created. Each population is a set of 100 000 elements and is devided into10 strata and 100 small areas. These populations are sets of points $(y_{igk}, x_{1igk}, x_{2igk})$ where $i = 1, ..., 100$, $g = 1, ...,10$, $k = 1, ..., 100$, and where i denotes the number of small area, g – the number of stratum, k – the number of element in i-th small area and g-th stratum, y_{igk} is value 0 or 1 and x_{1igk}, x_{2igk} are values of transformation of random numbers generated from normal distribution. In experiments the auxiliary variables are transformations of random variables which are normal distributed because the normal distribution appears frequently as distribution of statistical variables. The parameters of auxiliary variables' distributions are taken in such way that these distributions differ more or less in strata and small areas.

The values y_{igk}, x_{1igk}, x_{2igk} are realizations of random variables Y_{ig}, X_{1ig}, X_{2ig} respectively. These variables are determined separately for each small area and each stratum as follows:

$$Y_{ig} = 1, \text{if } Z_{ig} < c_{ig} \tag{9}$$

$$Y_{ig} = 0, \text{ if } Z_{ig} \geq c_{ig} \tag{10}$$

and

$$X_{1ig} = U_1 + \xi_{1ig} \tag{11}$$

$$X_{2ig} = U_2 + \xi_{2ig} \tag{12}$$

$$Z_{ig} = 2X_{1ig} + 3X_{ig} + \varepsilon_i \tag{13}$$

where $U_1 \sim N(10; 2)$, $\xi_{1ig} \sim N(0; (i+j)/500)$, $U_2 \sim N(5; 1)$, $\xi_{2ig} \sim N(0; (i+j)/1000)$, $\varepsilon_i \sim N(i/b; i/10)$, $b = 30$ for populations A, C and $b = 100$ for population B; the

random variables U_1, ξ_{1ig}, U_2, ξ_{2ig}, ε_i are independent; the value c_{ig} is the 10-th centile of distribution of random variable W_{ig} which has the form:

$$W_{ig} = (Z_{ig} - 35 - i/30) / [16 + 4((i+j)/500)^2 + 9 + 9((i+j)/1000)^2 + 0.01]$$

for population A, (14)

$$W_{ig} = (Z_{ig} - 35) / [16 + 4((i+j)/500)^2 + 9 + 9((i+j)/100)^2 + 0.01]$$

for population B and C. (15)

Some parameters of small areas for populations A, B, C are presented in Table 1. We can see that these parameters are the least different in the population A and the most different in the population C. The same populations were analysed with respect to errors of estimation of proportion on the basis of logistic regression model for small areas in the paper P r u s k a (2005).

T a b l e 1

Some parameters of small areas [a] for populations A, B, C

Population	$\min_{1 \le i \le 100} \theta_i$	$\max_{1 \le i \le 100} \theta_i$
A	0.513	0.578
B	0.479	0.563
C	0.361	0.561

[a] Parameter θ_i is the proportion for i-th small area.
S o u r c e: own calculations.

Next, from the set of small areas for each population A, B, C fifteen small areas were drawn (sampling with replacement). From each of these small areas and from the whole population random samples were drawn. Four sampling designs for sampling with replacement and sampling without replacement were applied:

– sampling from the whole population and poststratification,
– stratified sampling from the whole population,
– sampling from each drawn small area and poststratification,
– stratified sampling from each drawn small area.

In case of drawing from the whole population two sizes of population sample are considered: 4000 and 10 000. Next, small area samples are created. In this case small area sample consists of these elements of population sample which belong to the small area.

Moreover, samples were drawn from each small area from among fifteen drawn small areas. In this case two sizes of small area sample are also taken into consideration: 40 and 100.

The samples were drawn 1000 times for each sampling design and each population.

The estimates of proportion of one's for i-th small area, which is denoted by θ_i, were determined on the basis of these samples and on the basis of logistic regression model.

At first, the following values were calculated for each drawn small area and for each population sample, and for each sampling design:

$$\hat{\theta}_{ij} = \bar{y}_{ij} = \frac{1}{NM_i} \sum_{g=1}^{G} \left(\frac{N_{ig}}{n_{igj}} \sum_{l=1}^{n_{igj}} y_{igjl} \right) \tag{16}$$

$$\bar{x}_{1ij} = \frac{1}{NM_i} \sum_{g=1}^{G} \left(\frac{N_{ig}}{n_{igj}} \sum_{l=1}^{n_{igj}} x_{1igjl} \right) \tag{17}$$

$$\bar{x}_{2ij} = \frac{1}{NM_i} \sum_{g=1}^{G} \left(\frac{N_{ig}}{n_{igj}} \sum_{l=1}^{n_{igj}} x_{2igjl} \right) \tag{18}$$

where:

i – number of drawn small area (i = 1, ..., 15),

g – number of stratum (g = 1, ..., 10),

j – number of repetition (j = 1, ..., 1000),

NM_i – number of elements in i-th small area,

N_{ig} – number of elements in i-th small area and g-th stratum,

n_{igj} – number of sample elements which belong to i-th small area and g-th stratum for j-th repetition,

y_{igjl} – value of random variable $Y_i(Y_{ig})$ for l-th sample element from i-th small area and g-th stratum for j-th repetition,

x_{1igjl} – value of random variable $X_{1i}(X_{1ig})$ for l-th sample element from i-th small area and g-th stratum for j-th repetition,

x_{2igjl} – value of random variable $X_{2i}(X_{2ig})$ for l-th sample element from i-th small area and g-th stratum for j-th repetition.

Next, the following model was considered:

$$L^{-1}(\hat{\theta}_{ij}) = \alpha_0 + \alpha_1 \bar{x}_{1ij} + \alpha_2 \bar{x}_{2ij} + \varepsilon_{ij} \tag{19}$$

where

$$L^{-1}(\hat{\theta}_{ij}) = ln \frac{\hat{\theta}_{ij}}{1-\hat{\theta}_{ij}} \tag{20}$$

for each sampling design and for each j-th repetition, separately.

The parameters of model (19) were estimated by GLS method on the basis of information about 15 drawn small areas for each repetition and for each population, separately. The estimates of these parameters are denoted by a_{0j}, a_{1j}, a_{2j} ($j = 1, ..., 1000$) respectively.

Then the following values were calculated:

$$l_{ij} = a_{0j} + a_{1j}\overline{X}_{1i} + a_{2j}\overline{X}_{2i} \tag{21}$$

and

$$\hat{\theta}_{ij}^{L} = \frac{\exp(l_{ij})}{1+\exp(l_{ij})} \tag{22}$$

for $i = 1, ..., 100$ and $j = 1, ..., 1000$ where $\overline{X}_{1i}$ and $\overline{X}_{2i}$ are means of variables X_{1i} and X_{2i} for i-th small area. The $\hat{\theta}_{ij}$ and $\hat{\theta}_{ij}^{L}$ are values of different estimators of parameter θ_i. The values $\hat{\theta}_{ij}^{L}$ are determined on the basis of values $\hat{\theta}_{ij}$ ($\hat{\theta}_{ij}$ are used in estimation of parameters of model (19)).

Next, values of the following measures were calculated:

$$SO_i^2 = \frac{1}{1000}\sum_{j=1}^{1000}(\hat{\theta}_{ij}^{L} - \theta_i)^2 \quad \text{for } i = 1, ..., 100, \tag{23}$$

$$S_i^2 = \frac{1}{1000}\sum_{j=1}^{1000}(\hat{\theta}_{ij}^{L} - \overline{\theta}_{ij}^{L})^2 \quad \text{for } i = 1, ..., 100 \tag{24}$$

where

$$\overline{\theta}_{ij}^{L} = \sum_{j=1}^{1000}\hat{\theta}_{ij}^{L} \tag{25}$$

We can take that the formulas (23) and (24) defined, respectively, the value of estimator of MSE (mean square error) and the value of estimator of variance for predictor which is obtained on the basis of model (19). Another estimator of this predictor variance is estimator determined by jackknife method. It is denoted by v_{JACK} and is given by formula (8). The value of v_{JACK} for θ_i (for i-th small area) and j-th repetition is denoted by v_{ij}. In this case in formula (8) the value of T_n is determined by formula (16). The value of v_{JACK} can be calculated in empirical investigations but the values SO_i^2 and S_i^2 can be determined only in simulation experiments and were calculated for comparisons in this analysis.

Table 2

Some results of simulation experiments for population A for sampling from population[a)]

Set T	Measure	Size of population sample and sampling design			
		4 000		10 000	
		Poststratification	Stratified sampling	Poststratification	Stratified sampling
Sampling with replacement					
Set of indices of drawn (to sample) small areas	$\min_{i\in T \atop 1\le s\le 1000} v_{ij}$	0.000005*	0.000014	0.000004	0.000001
	$\max_{i\in T \atop 1\le s\le 1000} v_{ij}$	0.011877*	0.011359	0.004485	0.002105
	$\min_{i\in T} S0_i^2$	0.0006	0.0005	0.0001	0.0001
	$\max_{i\in T} S0_i^2$	0.0018*	0.0017	0.0006	0.0006
	$\min_{i\in T} S_i^2$	0.0005	0.0005	0.0001	0.0001
	$\max_{i\in T} S_i^2$	0.0006	0.0005	0.0001	0.0001
Set of indices of undrawn (to sample) small areas	$\min_{i\in T \atop 1\le s\le 1000} v_{ij}$	0.000005*	0.000022	0.000003	0.000001
	$\max_{i\in T \atop 1\le s\le 1000} v_{ij}$	0.013423	0.013044	0.007093*	0.002683
	$\min_{i\in T} S0_i^2$	0.0005	0.0005	0.0001	0.0001
	$\max_{i\in T} S0_i^2$	0.0017*	0.0014	0.0005*	0.0006*
	$\min_{i\in T} S_i^2$	0.0005	0.0005	0.0001	0.0001
	$\max_{i\in T} S_i^2$	0.0007	0.0006	0.0002	0.0002*
Sampling without replacement					
Set of indices of drawn (to sample) small areas	$\min_{i\in T \atop 1\le s\le 1000} v_{ij}$	0.000016	0.000020	0.000002	0.000001
	$\max_{i\in T \atop 1\le s\le 1000} v_{ij}$	0.008297	0.009516	0.002989	0.003303
	$\min_{i\in T} S0_i^2$	0.0006	0.0005	0.0001	0.0001
	$\max_{i\in T} S0_i^2$	0.0017	0.0016	0.0006*	0.0005
	$\min_{i\in T} S_i^2$	0.0005	0.0005	0.0001	0.0001
	$\max_{i\in T} S_i^2$	0.0006	0.0006	0.0001	0.0001
Set of indices of undrawn (to sample) small areas	$\min_{i\in T \atop 1\le s\le 1000} v_{ij}$	0.000011	0.000017	0.000002	0.000001
	$\max_{i\in T \atop 1\le s\le 1000} v_{ij}$	0.011844*	0.011321	0.007232	0.006343
	$\min_{i\in T} S0_i^2$	0.0005	0.0005	0.0001	0.0001
	$\max_{i\in T} S0_i^2$	0.0017	0.0016	0.0006*	0.0006*
	$\min_{i\in T} S_i^2$	0.0005	0.0005	0.0001	0.0001
	$\max_{i\in T} S_i^2$	0.0007	0.0007	0.0002*	0.0002*

[a)] The symbol "*" denotes that the number appears rarely.

Source: own calculations.

Table 3

Some results of simulation experiments for population A for sampling from each drawn small areas[a])

Set T	Measure	Size of small area sample and sampling design			
		40		100	
		Poststratification	Stratified sampling	Poststratification	Stratified sampling
Sampling with replacement					
Set of indices of drawn (to sample) small areas	$\min_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.000019	0.000008	0.000003	0.000002
	$\max_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.014829	0.005693	0.002300	0.003583
	$\min_{i\in T} S0_i^2$	0.0005	0.0002	0.0001	0.0001
	$\max_{i\in T} S0_i^2$	0.0013	0.0007	0.0006*	0.0006*
	$\min_{i\in T} S_i^2$	0.0005	0.0002	0.0001	0.0001
	$\max_{i\in T} S_i^2$	0.0005	0.0002	0.0001	0.0001
Set of indices of undrawn (to sample) small areas	$\min_{i\notin T,\ 1\le j\le 1000} v_{ij}$	0.000019	0.000005	0.000003	0.000002
	$\max_{i\notin T,\ 1\le j\le 1000} v_{ij}$	0.021409*	0.007172	0.003631	0.003947
	$\min_{i\notin T} S0_i^2$	0.0005	0.0002	0.0001	0.0001
	$\max_{i\notin T} S0_i^2$	0.0015	0.0007	0.0006*	0.0001
	$\min_{i\notin T} S_i^2$	0.0004	0.0002	0.0001	0.0001
	$\max_{i\notin T} S_i^2$	0.0007	0.0003	0.0002*	0.0002*
Sampling without replacement					
Set of indices of drawn (to sample) small areas	$\min_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.000016*	0.000007	0.000002	0.000002
	$\max_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.007907	0.004766	0.002696*	0.003059*
	$\min_{i\in T} S0_i^2$	0.0005	0.0002	0.0001	0.0001
	$\max_{i\in T} S0_i^2$	0.0015	0.0007	0.0006*	0.0005*
	$\min_{i\in T} S_i^2$	0.0004	0.0002	0.0001	0.0001
	$\max_{i\in T} S_i^2$	0.0005	0.0002	0.0001	0.0001
Set of indices of undrawn (to sample) small areas	$\min_{i\notin T,\ 1\le j\le 1000} v_{ij}$	0.000015*	0.000007	0.000002	0.000002
	$\max_{i\notin T,\ 1\le j\le 1000} v_{ij}$	0.12663*	0.008734*	0.003900*	0.004426*
	$\min_{i\notin T} S0_i^2$	0.0004	0.0002	0.0001	0.0001
	$\max_{i\notin T} S0_i^2$	0.0013	0.0008	0.0006*	0.0006*
	$\min_{i\notin T} S_i^2$	0.0004	0.0002	0.0001	0.0001
	$\max_{i\notin T} S_i^2$	0.0006	0.0003	0.0002*	0.0002*

[a]) The star "*" denotes that the number appears rarely.

S o u r c e: own calculations.

Table 4

Some results of simulation experiments for population B for sampling from population[a)]

Set T	Measure	Size of population sample and sampling design			
		4000		10000	
		Poststratification	Stratified sampling	Poststratification	Stratified sampling
Sampling with replacement					
Set of indices of drawn (to sample) small areas	$\min_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.000011	0.000015	0.000004	0.000001
	$\max_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.011089	0.014596	0.004485	0.003304
	$\min_{i\in T} S0_i^2$	0.0006	0.0005	0.0001	0.0001
	$\max_{i\in T} S0_i^2$	0.0016	0.0015	0.0014*	0.0014*
	$\min_{i\in T} S_i^2$	0.0005	0.0004	0.0001	0.0001
	$\max_{i\in T} S_i^2$	0.0006	0.0005	0.0002*	0.0001
Set of indices of undrawn (to sample) small areas	$\min_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.000010	0.000015	0.000004	0.000001
	$\max_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.013159	0.013633	0.005213	0.005702*
	$\min_{i\in T} S0_i^2$	0.0005	0.0004	0.0001	0.0001
	$\max_{i\in T} S0_i^2$	0.0028	0.0027	0.0014*	0.0015*
	$\min_{i\in T} S_i^2$	0.0005	0.0004	0.0001	0.0001
	$\max_{i\in T} S_i^2$	0.0007	0.0005	0.0002*	0.0002*
Sampling without replacement					
Set of indices of drawn (to sample) small areas	$\min_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.000010	0.000016	0.000006	0.000003
	$\max_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.008195	0.013593*	0.002833*	0.004666
	$\min_{i\in T} S0_i^2$	0.0005	0.0005	0.0001	0.0001
	$\max_{i\in T} S0_i^2$	0.0016	0.0017	0.0014*	0.0013*
	$\min_{i\in T} S_i^2$	0.0005	0.0005	0.0001	0.0001
	$\max_{i\in T} S_i^2$	0.0006*	0.0006*	0.0001	0.0001
Set of indices of undrawn (to sample) small areas	$\min_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.000007	0.000012	0.000003*	0.000003
	$\max_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.010368	0.012221*	0.005940*	0.005559
	$\min_{i\in T} S0_i^2$	0.0005	0.0005	0.0001	0.0001
	$\max_{i\in T} S0_i^2$	0.0027*	0.0025*	0.0014*	0.0014*
	$\min_{i\in T} S_i^2$	0.0005	0.0005	0.0001	0.0001
	$\max_{i\in T} S_i^2$	0.0007*	0.0007*	0.0002*	0.0002*

[a)] The star "*" denotes that the number appears rarely.

Source: own calculations.

Table 5

Some results of simulation experiments for population B for sampling from each drawn small areas[a)]

Set T	Measure	Size of small area sample and sampling design			
		40		100	
		Poststratification	Stratified sampling	Poststratification	Stratified sampling
Sampling with replacement					
Set of indices of drawn (to sample) small areas	$\min_{i\in T,\ 15\le s\le 1000} v_{ij}$	0.000014	0.000012	0.000003*	0.000003
	$\max_{i\in T,\ 15\le s\le 1000} v_{ij}$	0.012165	0.004583	0.003412	0.004067*
	$\min_{i\in T} S0_i^2$	0.0005	0.0002	0.0001	0.0001
	$\max_{i\in T} S0_i^2$	0.0016	0.0015*	0.0014*	0.0014*
	$\min_{i\in T} S_i^2$	0.0005	0.0002	0.0001	0.0001
	$\max_{i\in T} S_i^2$	0.0005	0.0002	0.0001	0.0001
Set of indices of undrawn (to sample) small areas	$\min_{i\in T,\ 15\le s\le 1000} v_{ij}$	0.000010	0.000009	0.000003*	0.000002
	$\max_{i\in T,\ 15\le s\le 1000} v_{ij}$	0.015466	0.005182	0.005186	0.004846
	$\min_{i\in T} S0_i^2$	0.0005	0.0002	0.0001	0.0001
	$\max_{i\in T} S0_i^2$	0.0025*	0.0016*	0.0014*	0.0014*
	$\min_{i\in T} S_i^2$	0.0004	0.0002	0.0001	0.0001
	$\max_{i\in T} S_i^2$	0.0006	0.0003*	0.0002*	0.0002*
Sampling without replacement					
Set of indices of drawn (to sample) small areas	$\min_{i\in T,\ 15\le s\le 1000} v_{ij}$	0.000024	0.000007	0.000003	0.000003
	$\max_{i\in T,\ 15\le s\le 1000} v_{ij}$	0.012840	0.005639*	0.002985	0.005131
	$\min_{i\in T} S0_i^2$	0.0005	0.0002	0.0001	0.0001
	$\max_{i\in T} S0_i^2$	0.0015	0.0015*	0.0014*	0.0014*
	$\min_{i\in T} S_i^2$	0.0004	0.0002	0.0001	0.0001
	$\max_{i\in T} S_i^2$	0.0005	0.0002	0.0001	0.0001
Set of indices of undrawn (to sample) small areas	$\min_{i\in T,\ 15\le s\le 1000} v_{ij}$	0.000019	0.000003	0.000003	0.000001*
	$\max_{i\in T,\ 15\le s\le 1000} v_{ij}$	0.016199	0.009569	0.005547	0.005660
	$\min_{i\in T} S0_i^2$	0.0004	0.0002	0.0001	0.0001
	$\max_{i\in T} S0_i^2$	0.0024	0.0016*	0.0014*	0.0014*
	$\min_{i\in T} S_i^2$	0.0004	0.0002	0.0001	0.0001
	$\max_{i\in T} S_i^2$	0.0006*	0.0003*	0.0002*	0.0002*

[a)] The star "*" denotes that the number appears rarely.

S o u r c e: own calculations.

Table 6

Some results of simulation experiments for population C for sampling from population [a)]

Set T	Measure	Size of population sample and sampling design			
		4000		10000	
		Poststratification	Stratified sampling	Poststratification	Stratified sampling
Sampling with replacement					
Set of indices of drawn (to sample) small areas	$\min_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.000010	0.000038	0.000008	0.000004*
	$\max_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.016132	0.019004	0.011012*	0.010948
	$\min_{i\in T} S0_i^2$	0.0007	0.0006	0.0002	0.0002
	$\max_{i\in T} S0_i^2$	0.0077*	0.0075*	0.0064	0.0065
	$\min_{i\in T} S_i^2$	0.0006	0.0005	0.0002	0.0001*
	$\max_{i\in T} S_i^2$	0.0006	0.0006*	0.0003*	0.0003*
Set of indices of undrawn (to sample) small areas	$\min_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.000005	0.000037	0.000006	0.000005
	$\max_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.016165	0.019915	0.009805	0.011165
	$\min_{i\in T} S0_i^2$	0.0006	0.0005	0.0002	0.0002
	$\max_{i\in T} S0_i^2$	0.0120*	0.0117*	0.0095	0.0097*
	$\min_{i\in T} S_i^2$	0.0006	0.0005	0.0002	0.0001
	$\max_{i\in T} S_i^2$	0.0008*	0.0008*	0.0006*	0.0005*
Sampling without replacement					
Set of indices of drawn (to sample) small areas	$\min_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.000011	0.000028	0.000016	0.000004
	$\max_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.014000	0.016738*	0.008700*	0.016058*
	$\min_{i\in T} S0_i^2$	0.0006*	0.0006*	0.0002*	0.0002
	$\max_{i\in T} S0_i^2$	0.0073*	0.0071*	0.0064	0.0063
	$\min_{i\in T} S_i^2$	0.0005	0.0005*	0.0001	0.0001
	$\max_{i\in T} S_i^2$	0.0006	0.0007*	0.0003	0.0003*
Set of indices of undrawn (to sample) small areas	$\min_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.000007	0.000022	0.000009	0.000004
	$\max_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.018680	0.019316	0.017937*	0.017850
	$\min_{i\in T} S0_i^2$	0.0006	0.0006	0.0002*	0.0002*
	$\max_{i\in T} S0_i^2$	0.0115*	0.0112*	0.0095*	0.0094*
	$\min_{i\in T} S_i^2$	0.0005	0.0005	0.0001	0.0001
	$\max_{i\in T} S_i^2$	0.0008*	0.0008*	0.0005*	0.0005*

[a)] The star "*" denotes that the number appears rarely.

Source: own calculations.

Table 7

Some results of simulation experiments for population C for sampling from each drawn small areas[a)]

Set T	Measure	Size of small area sample and sampling design			
		40		100	
		Poststratification	Stratified sampling	Poststratification	Stratified sampling
Sampling with replacement					
Set of indices of drawn (to sample) small areas	$\min_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.000024	0.000022	0.000007	0.000012
	$\max_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.016722	0.009422*	0.009707	0.007922
	$\min_{i\in T} S0_i^2$	0.0006*	0.0003*	0.0002*	0.0001*
	$\max_{i\in T} S0_i^2$	0.0074*	0.0068	0.0064	0.0062
	$\min_{i\in T} S_i^2$	0.0005	0.0002*	0.0002	0.0001*
	$\max_{i\in T} S_i^2$	0.0006*	0.0004*	0.0003*	0.0003*
Set of indices of undrawn (to sample) small areas	$\min_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.000019	0.000012*	0.000007	0.000012
	$\max_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.022329	0.015978*	0.010777	0.011389
	$\min_{i\in T} S0_i^2$	0.0005	0.0003	0.0002	0.0002*
	$\max_{i\in T} S0_i^2$	0.0115*	0.0096*	0.0094	0.0094
	$\min_{i\in T} S_i^2$	0.0005	0.0002	0.0001	0.0001
	$\max_{i\in T} S_i^2$	0.0008*	0.0005*	0.0005*	0.0005*
Sampling without replacement					
Set of indices of drawn (to sample) small areas	$\min_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.000009	0.000025	0.000011	0.000011
	$\max_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.010308	0.010170	0.015993	0.006652
	$\min_{i\in T} S0_i^2$	0.0005*	0.0003*	0.0002*	0.0001*
	$\max_{i\in T} S0_i^2$	0.0072*	0.0067*	0.0063	0.0064*
	$\min_{i\in T} S_i^2$	0.0005	0.0003	0.0001*	0.0001
	$\max_{i\in T} S_i^2$	0.0006*	0.0004*	0.0003*	0.0003*
Set of indices of undrawn (to sample) small areas	$\min_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.000008	0.000020	0.000007	0.000002*
	$\max_{i\in T,\ 1\le j\le 1000} v_{ij}$	0.013104	0.013843	0.019916	0.015537*
	$\min_{i\in T} S0_i^2$	0.0005	0.0003	0.0002*	0.0002*
	$\max_{i\in T} S0_i^2$	0.0112*	0.0094*	0.0094*	0.0094*
	$\min_{i\in T} S_i^2$	0.0005	0.0003	0.0001	0.0001
	$\max_{i\in T} S_i^2$	0.0007*	0.0005*	0.0005*	0.0005*

[a)] The star "*" denotes that the number appears rarely.

Source: own calculations.

Some results of simulation experiments are presented in Tables 2–7. We can see that the values of S_i^2 are the least differentiated and the values of v_{ij} are the most differentiated. We can take that the estimator SO_i^2 is the most precise with respect to its definitions. In confirmity with the known facts we observe smaller estimates of variance for the estimator of θ_i for larger sample size; in general, we observe smaller estimation errors for the estimator of θ_i for dependent sampling and stratified sampling than for other cases of sampling but differences are small (it is expected result). In general, differences between estimates of variance of estimator of θ_i for drawn (to sample) small areas and undrawn small areas are small yet they are slightly bigger for undrawn small areas than drawn small areas. Generally, values of estimator SO_i^2 are greater than values of estimator S_i^2. The differentiation of values v_{ij} means that quite big estimation errors can appear when we apply the jackknife method to estimation of variance of proportion predictor.

4. Final refmark

We can use different forms of logistic regression models and different methods of estimation of their parameters for small areas (for comparison see Rao (2001)). The properties of estimators of variance of predictor obtained on the basis of logistic regression models depend on many factors (sampling designs, sample size, auxiliary information, construction of estimator). So there is a need to continue the conducted analysis.

Reference

Jajuga K. (1990), *Modele z dyskretną zmienną objaśnianą*, [in:] Bartosiewicz S., *Estymacja modeli ekonometrycznych*, PWE, Warszawa.

Pruska K. (2005), *Logistic regression models in small area investigations*, paper presented on the SAE2005 Conference, 28–31 August, Jyväskylä, Finland.

Rao J. N. K. (2003), *Small Area Estimation*, John Wiley &Sons, New Jersey.

Shao J., Tu D. (1996), *The Jackknife and Bootstrap*, Springer, New York.

Wiśniewski J. (1986), *Ekonometryczne badanie zjawisk jakościowych, Studium metodologiczne*, Uniwersytet Mikołaja Kopernika, Toruń.

Krystyna Pruska

Estymacja wariancji logistyczno-regresyjnych predyktorów dla małych obszarów

Modele regresji logistycznej mogą być wykorzystywane do analizy zmiennych o rozkładzie zero-jedynkowym dla małych obszarów.

W pracy rozpatrywane są logistyczno-regresyjne predyktory parametru rozkładu zero-jedynkowego oraz estymacja MSE i wariancji tych predyktorów w przypadku małych obszarów. Przedstawione są wyniki eksperymentów symulacyjnych, których celem jest analiza własności estymatorów wariancji i błędu średniokwadratowego logistyczno-regresyjnych predyktorów.