УДК [311.2:167.22]:519.233.5 **Doi: 10.31767/su.3(94)2021.03.01**
JEL Classification: C15, C51, C53

L. A. Soshnikova,
DSc in Economics, Professor,
Professor of the Department,
Educational Establishment "Belarus State Economic University",
E-mail: ludmila_sosh@mail.ru
ResearcherID: AAX-4737-2021,
ORCID: https://orcid.org/0000-0002-1402-5490

Using the Logistic Regression in Analysis of Results from Statistical Observations

The article is focused on investigating the problems arising in the statistical analysis with use of logistic models of the ordered multiple choice, which are constructed by the results of statistical observations involving the existence of a categorical dependent variable. This group of models should be used when a discrete dependent variable takes several alternative values. The examples include assessment of student performance (perfect, good, satisfactory, unsatisfactory). These models' parameters are estimated using the algorithms based on elements of the probability theory. The purpose of constructing the multiple choice model is to determine the factors with impact on the probability of the occurrence of a particular event and the choice of an alternative, as well as the strength of this impact. A detailed description of the algorithms for estimating logit models of binary and multiple choice is given, with demonstrating the model application in solving a particular problem (statistical analysis of the results of self-assessment of health status by household members) by use of SPSS package. It should be noted that statistical packages like SPSS, STATISTICA or STATA contain the modules for constructing logit and probit models.

The assessment of population's health status includes the objective assessment of their health status by the official statistics data on the prevalence of deceases and the cumulative subjective assessment of the individual health status by the results of sociological studies It is important to know to what extent the objective assessment of the population's health status complies with the subjective perception of the health status by individuals. Because the primary files of the sample survey of households are confidential, the multiple choice model was constructed by the author using the proxy data with characteristics close to actual values. Variables such as residence place, gender, age, assessment of health status, sports practicing, smoking and income were reported in the process of the sample survey. In constructing the model, the variable "health" was used as a dependent variable; "gender" and "education" were used as categorical variables; "age" and "income" were used as covariates. Once the model was constructed and its identification capacity (i. e. the correctness of the predicted dependent variable) estimated, its specification was saved in a special file for the subsequent rebuilding.

Key words: *logit model, binary choice, multiple choice, chance logarithm, maximum likelihood method, self-assessment of health status.*

Introduction. The present-day practice of statistical analysis features the continuously extending set of statistical and econometric methods. They have been commonly and quite effectively used in correlation analyses, consumer behavior predictions, assessments of financial sustainability of businesses, creditworthiness of legal entities and physical persons, etc. Also, applications of these techniques helps increase the effectiveness of analytical and predictive efforts focused on assessment of financial condition of business enterprises and credit risks for various categories of economic activities.

Various types of regression models can also be useful for comprehensive analyses of results from sample observations. Thus, in the sample survey of households, designed to assess the employment, three categories of persons are involved: unemployed persons, partially employed persons, persons with full day employment; in self-assessment of the population's health status, the participants are also proposed to choose one of the several responses: good, satisfactory, bad. The classical linear regression is not applicable for situations where there is no natural ordering of values of a dependent variable. The best alternative in these cases can be the polynomial logistic regression. A dependent variable that takes several alternative value is called "discrete variable". Accordingly, the regression models with a discrete variable selected as a dependent one are called "discrete choice models".

The class of econometric models explored in this article arouses wide interest of statisticians, sociologists, analysts working in the bank sector, ecologists, market analysts and others. A lot of recent scientific research of domestic and foreign authors has been conducted in the field of applied statistics and econometrics (S. Anatolyev [1], E. Sedova [11], S. Zsolt [15]), assessment of bank sustainability, analysis of credit risks (Ya. Magnus [7], A. Peresetskiy [9]). These works deal with various dimensions of constructing binary and multiple choice models.

This article's objective is the author's desire to demonstrate wide-scale analytical capacities of discrete choice models (logistic models of multiple choice) for analyses of results from sample statistical observations, and to show how their construction algorithms can be implemented in the very popular statistical package SPSS.

Methods for analyzing discrete dependent variables. Models for binary and multiple choice are distinguished depending on the number of alternatives. Binary choice models tend to be used when a dependent variable can take only two values: 0 and 1. The choice of function determines the type of binary model. When the standard normal distribution function is used, the binary choice model will be called "probit model" [5, p. 223]:

$$F(AX) = P\left(y = 1/X\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(-\frac{z^2}{2}\right) dz, \quad (1)$$

where $z = a_1 x_1 + a_2 x_2 + ... + a_m x_m$.

When the logistic distribution function is used, the binary choice model will be called "logit model":

$$F(AX) = P\left(y = 1/X\right) = \frac{1}{1 + \exp(-z)},$$

$$P\left(y = 0/X\right) = 1 - \frac{1}{1 + \exp(-z)} = \frac{1}{1 + \exp(z)}.$$

Then the ratio of the probability of the occurrence of the event P to the probability of the non-occurrence of the event $1 - P$ will be presented as follows:

$$\frac{P}{1-P} = \frac{1 + \exp(z)}{1 + \exp(-z)} = \exp(z).$$

When the logarithm is used, the expression will take the following form:

$$\ln \frac{P}{1-P} = z = a_1 x_1 + a_2 x_2 + ... + a_m x_m. \quad (2)$$

The expression (2) is called "chance logarithm" (logit). Mathematically, the logistic regression model shows its dependence on the linear combination of independent variables x_i, $i = \overline{1, m}$.

To derive the vector of estimates $A = (a_1, a_2, ..., a_m)$ in discrete choice models, the maximum likelihood method is most commonly used [6, p. 621; 8, p. 243]. The essence of this method is as follows. At the first phase the likelihood function is constructed, which needs to be maximized on a certain set of the observable input variables x_i used as factors or covariates for the studied categorical targeted variable Y:

$$L(A) = \prod_{i=1}^{N} P\left(Y_i = 1/X_i; A\right)^{Y_i} \cdot P\left(Y_i = 0/X_i; A\right)^{1-Y_i}, \quad (3)$$

where i is the number of observation, $i = \overline{1, N}$.

After that, the likelihood function will be logarithmed for the convenience of subsequent calculations:

$$\ln L(A) = \sum_{i=1}^{N} Y_i \ln F\left(X_i^T A\right) + \\ + \sum_{i=1}^{N} (1 - Y_i) \ln\left(1 - F\left(X_i^T A\right)\right). \quad (4)$$

The estimates of parameters $a_1, a_2, ..., a_m$, derived by this method, will be valid, asymptotically effective and asymptotically normal.

The chance logarithm shows the linear dependence of the probability of the occurrence of a particular event on the values of independent variables. The constant in the model shows the natural level of the occurrence of the modelled event (such as bank default) given that all the independent variables equal to zero. The values of coefficients at independent categorical variables showing their impact on the chance of the event occurrence are measured in the logarithmic scale. The coefficients of the logistic regression model are usually interpreted using the exponential form of its presentation [4]:

$$P\left(y = 1/X\right) = \frac{1}{1 + \exp(-(\alpha_0 + \alpha_1 x_1 + ... + \alpha_m x_m + \varepsilon))}. \quad (5)$$

The coefficients presented in this form show the average extent of change in the chances of the event occurrence when the independent variable is changed by one unit of its measurement and the other variables remain unchanged. When the regression coefficient is positive, its exponent will be higher than one, and the chances will be increasing, but when the coefficient is negative, its exponent will be lower than one, and the chances for realization of the modelled event will be declining.

When the categorical independent variable is included in the model (such as bank size: large, medium or small banks), the regression coefficient in the exponential form will show the proportion of chances given the existence of the factor reflected in this independent variable compared with its absence.

The hypothesis about the significance of the coefficients in binary choice models is checked using several techniques [6; 7; 14]:

- Wald test: the quarter of the ratio of the coefficient to its standard error;
- Lagrange multiplier (LM) test;
- Likelihood ratio (LR) test.

The next type of model is *binary choice model* that is a synthesis of logit binary choice model. With narrow applications in the statistical analysis today, it can become more popular with the development of statistical toolkits and the expanding scope of specific problems which solutions are focused on estimating the probabilities of the occurrence of a particular event.

In multiple choice models [2; 7; 12], the existence of k non-observable continuous variables $z_1, z_2, ..., z_k$ is assumed for the dependent variable y with k categories, and each of these variables can be considered as "a propensity to the respective category". The correlation between z and the probability of a particular value of the effective variable can be described mathematically by the formula (6):

$$P\left(y_{ik} = 1 / x_i\right) = \frac{\exp(z_{ik})}{\exp(z_{i1}) + \exp(z_{i2}) + ... + \exp(z_{ik})}, \quad (6)$$

where $P\left(y_{ik} = 1 / x_i\right)$ is the probability of falling of i observation into the category k; z_{ik} is the value k of non-observable continuous variable for i observation; z_{ik} is also considered as linearly related with predictors:

$$z_{ik} = a_{k0} + a_{k1}x_{i1} + a_{k2}x_{i2} + ... + a_{km}x_{im}, \quad (7)$$

where x_{ij} is the value of j predictor for i observation; a_{kj} is the coefficient of j predictor for k non-observable variable.

If z_k were a non-observable variable, the multiple linear regression would be easily applied for the assessment of a_{ki} parameters. But because it is non-observable, it will be necessary to relate the predictors with the probability by replacing z_k with a combination of predictors:

$$P(y_{ik} = 1) = \frac{\exp(a_{k0} + a_{k1}x_{i1} + ... + a_{km}x_{im})}{\exp(a_{10} + a_{11}x_{i1} + ... + a_{1m}x_{im}) + ... + \exp(a_{k0} + a_{k1}x_{i1} + ... + a_{km}x_{im})}. \quad (8)$$

The problem of model identifiability (7) is solved by assuming one of the non-observable variables as equal to zero ($z_k = 0$). Then the k category will be called "reference category", because all the parameters in the model are interpreted with reference to it. When the constant c is added to each z, the probability of the event occurrence will be unchanged:

$$P\left(y_{ik} = 1\right) = \frac{\exp(z_{ik} + c)}{\exp(z_{i1} + c) + \exp(z_{i2} + c) + ... + \exp(z_{ik} + c)} =$$

$$= \frac{\exp(z_{ik}) \cdot \exp(c)}{\exp(z_{i1}) \cdot \exp(c) + \exp(z_{i2}) \cdot \exp(c) + ... + \exp(z_{ik}) \cdot \exp(c)} =$$

$$= \frac{\exp(z_{ik})}{\exp(z_{i1}) + \exp(z_{i2}) + ... + \exp(z_{ik})} = P\left(y_{ik} = 1\right).$$

Coefficients a_{kj} are estimated by the iterative maximal likelihood method [13, c. 309; 14, c. 216], i. e. the procedure of estimating the regression coefficients is confined to the maximization of the probability of the occurrence of a particular value of the dependent variable (with the preset observable values).

This article describes the experience of applications of the multiple choice model in solving the specific problem: to analyze the results from the sample survey of households, for producing data on self-assessment of one's health status. The assessment of population's health status covers both the objective assessment of their health status by the official statistics data and the cumulative subjective assessment of the individual health status by the results of sample statistical observations. It is important to know to what extent the objective assessment of the population's health status complies with the subjective perception of the health status by individuals [8; 10; 12].

Multiple choice models in analyzing the results of self-assessment of health status by the population in SPSS package. The algorithm for constructing the multiple choice model (multiple logit regression) and the results of calculations will be demonstrated by the data from the regional sample survey of households, used to produce information on self-assessment of health status by the population.

The multiple choice model was constructed using the logistic regression with three categories of the dependent variable (the health status is good, satisfactory or bad). The analysis of polynomial logistic regression was started by setting the following procedure in the SPSS package menu: Analyze > Regression > Polynomial logistic regression [3]. The dependent variable "Health" and the factors (age, income, gender, and education) were selected in the opened window (Figure 1).

Because in the multiple choice logit model only categorical variables (gender and education in our case) can be used as factors with impact on the effective variable, quantitative factors (age and income) were assumed as covariates (predictors) that also had impact on the effective variable and were included in the model, but as contributing variables (see Figure 1).

The following regression equation was used (8):

$$P(z_k) = \frac{\exp(z_k)}{\exp(z_1) + \exp(z_2) + \exp(z_3)}, \quad (9)$$

where independent variables z_1, z_2, z_3 were estimated by use of formula (7) as, respectively, good (k = 1), satisfactory (k = 2) and bad (k = 3) health status. Age (x_2), income (x_2), gender (x_3), and education (x_4) were selected as the factors with impact on the effective variable.

Once the factors were selected and included in the model, it was necessary to set the stored variables that would be added in the file of input data after calculations. The two variables were selected: the predicted category and the probability of inclusion in the predicted category (Figure 2). When the model was constructed, its specifications were saved in Health.xml file containing all the information required for the model rebuilding. If necessary, the module "Master of scoring" can be used and applied with the already estimated parameters for other data sets (some models produce XML file of the model, others produce ZIP file of archive)

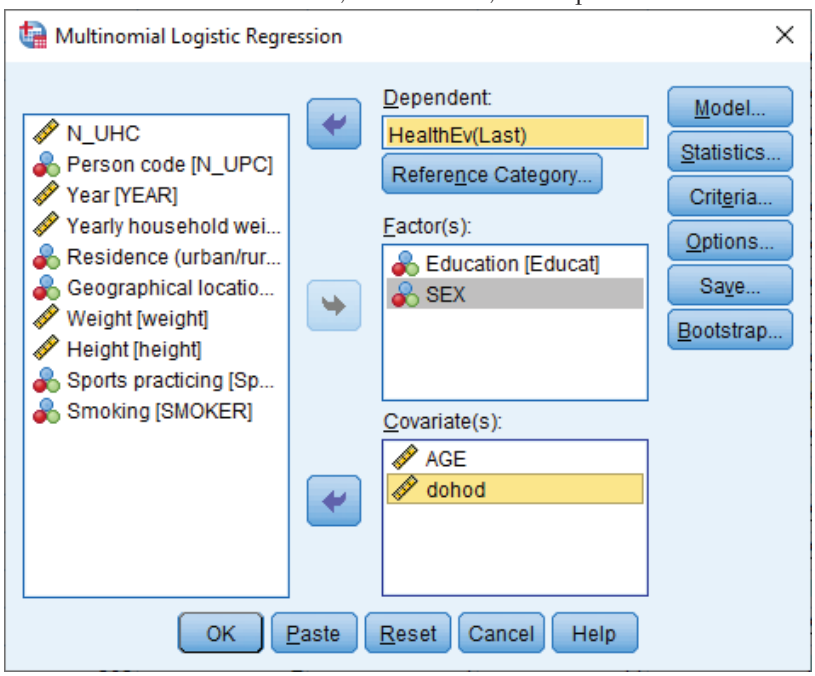

Figure 1. Selection of factors for the multiple choice logit model

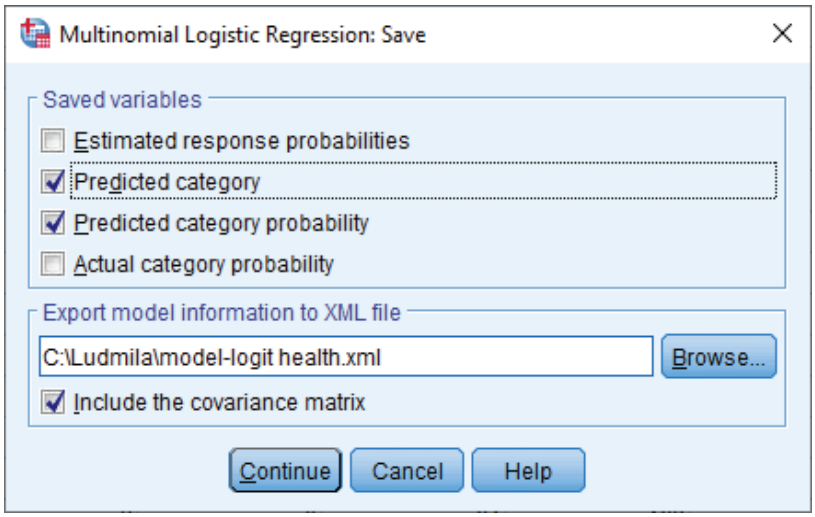

Figure 2. The choice of the stored variables and the exported model

The analysis demonstrates that the model is statistically reliable, because the proportion of identification is rather high, as shown in Table 1 (author's development). The results allow for the conclusion that of the total number of people assessing their health status as good (226 persons), the test could confirm this status for only 153 (67.7%). The other 73 persons were recognized by the test as ones with satisfactory (72) and bad (1) health status, although they considered themselves as perfectly healthy.

Of the total number of persons with satisfactory health status (309), the test could confirm this status for 231, with another 76 persons classified as ones with good health and two persons – as ones with bad health. Of the total number of persons with bad health status according to self-assessment, only 4 persons of 50 were identified as such, whereas good health status was confirmed by the test for 2 persons, and satisfactory health status – for 44. Basically, 388 cases of 585 (or 66.3%) could be correctly identified in all the health assessment categories. The estimates of model parameters and their significance are given in Tables 2 and 3 (author's development).

Table 1

The degree of identification of self-assessment of health status by the population of a region

Observable values		Predicted values: good	Predicted values: satisfactory	Predicted values: bad	Proportion of correct identifications
Assessment of health status:	good	153	72	1	67.7%
	satisfactory	76	231	2	74.8%
	bad	2	44	4	8.0%
Total proportion of identification		39.5%	59.3%	1.2%	66.3%

Table 2

The estimates of multiple choice logit regression model for good health status

Factors for assessment of the probability of good health status		Regression coefficients	Wald test
Incomes		0.006	0.000
Age		–0.151	0.000
Gender	Male	0.045	0,908
	Female	0.000	0.000
Education	Higher education	–15.214	0.000
	Secondary specialized education	–15.955	0.000
	Technical and vocational	–15.221	0.000
	Secondary education	–15.934	0.000
	Basic education	–14.758	0.000
	Primary education or no education	–16.279	0.000
Constant		22.759	0.000

The significance of model parameters assessed by Wald test $p < 0,05$ shows that the estimates of model parameters are significant for variables like age, income or education, and insignificant for gender variable [2, p. 298].

Table 3

The estimates of multiple choice logit regression model for satisfactory health status

Factors for assessment of the probability of satisfactory health status		Regression coefficients	Wald test
Incomes		0.006	0.000
Age		–0.072	0.000
Gender	Male	–0.585	0.089
	Female	0.000	0.000
Education	Higher education	–15.509	0.000
	Secondary specialized education	–16.092	0.000
	Technical and vocational	–15.996	0.000
	Secondary education	–16.155	0.000
	Basic education	–16.201	0.000
	Primary education or no education	–16.585	0.000
Constant		20.459	0.000

The probabilities of self-assessment of males' health status can be determined in the following way. The covariates (age and income) in the subsequent calculation are to be fixed at the same level equal to one. Three independent variables given $x_4 = 1$ (male) and $x_5 = 1$ (higher education) are to be estimated as:

$$z_1 = 22,759 + 0,006 - 0,151 + 0,045 \cdot 1 - 15,214 \cdot 1 = 7,445;$$

$$z_2 = 20,459 + 0,006 - 0,072 - 0,585 \cdot 1 - 15,509 \cdot 1 = 4,299;$$

$$z_3 = 0;$$

$$P(good) = \frac{\exp(z_1)}{\exp(z_1) + \exp(z_2) + \exp(z_3)} = \frac{\exp(7.445)}{\exp(7.445) + \exp(4.299) + \exp(0)} =$$
$$= \frac{1711.286}{1711.86 + 73.626 + 1} = \frac{1711.286}{1785.912} = 0.9582;$$

$$P(satisfactory) = \frac{\exp(z_2)}{\exp(z_1) + \exp(z_2) + \exp(z_3)} = \frac{\exp(4.299)}{\exp(7.445) + \exp(4.299) + \exp(0)} =$$
$$= \frac{73.626}{1785.912} = 0.0412;$$

$$P(bad) = \frac{\exp(z_3)}{\exp(z_1) + \exp(z_2) + \exp(z_3)} = \frac{\exp(0)}{\exp(7.445) + \exp(4.299) + \exp(0)} =$$
$$= \frac{1}{1785.912} = 0.0006.$$

So, the probabilities of a male with higher education falling into the group with good, satisfactory, and bad health status are equal to 95.8%, 4.1%, and 0.06%, respectively.

The probabilities for females with higher education are estimated as follows:

$$z_1 = 22{,}759 + 0{,}006 - 0{,}151 + 0 \cdot 2 - 15{,}214 \cdot 1 = 7{,}4;$$

$$z_2 = 20{,}459 + 0{,}006 - 0{,}072 + 0 \cdot 2 - 15{,}509 \cdot 1 = 4{,}884;$$

$$z_3 = 0;$$

$$P(good) = \frac{\exp(7.400)}{\exp(7.400) + \exp(4.884) + \exp(0)} =$$
$$= \frac{1635.984}{1635.984 + 132.158 + 1} = \frac{1635.984}{1769.142} = 0.9247;$$

$$P(satisfactory) = \frac{\exp(4.884)}{\exp(7.400) + \exp(4.884) + \exp(0)} = \frac{132.158}{1769.142} = 0.0747;$$

$$P(bad) = \frac{\exp(0)}{\exp(7.400) + \exp(4.884) + \exp(0)} = \frac{1}{1769.142} = 0.0006.$$

The estimated probabilities for all the education levels of males and females are shown in Table 4.

Table 4

The probability of self-assessment of health status for various categories of surveyed persons

(%)

Education	Probability of self-assessment of health status					
	good		satisfactory		bad	
	males	females	males	females	males	females
Higher, postgraduate	95.8	92.5	4.1	7.4	0.06	0.06
Specialized secondary	91.3	92.1	8.6	8.6	0.13	0.12
Technical and vocational	88.7	95.2	11.1	4.8	0.22	0.06
Secondary	89.4	85.6	10.3	14.1	0.22	0.25
Basic	82.8	77.2	16.5	22.0	0.68	0.72
Primary	84.8	77.9	14.2	19.1	0.96	1.04

A comparison of males' and females' estimates of self-assessment of health status can show that while the category "good health" has the highest proportion for both males and females irrespective of the education level, the category "bad health" has the lowest proportion. Also, the females with higher, secondary or basic and primary education tend to have a lower proportion within the category "good health" and a higher proportion within the categories "satisfactory health" and "bad health".

For a more detailed analysis of the distribution of self-assessment of health status, analogous models accounting for residence place (large city, small and medium town, rural area) should be built. They will allow to reveal the dependence of self-assessments of health status on not only the gender or education of household members, but on the type of their residence area.

Conclusion. Sample statistical observations need to be finished by a comprehensive econometric analysis of their results, intended to determine the correlation between the dependent variable and its factors, to estimate the impact of exogenous variables and test the character of this correlation. This is demonstrated by the data obtained from the sample survey of self-assessment of health status of household members, to show that they will have a practical significance only when a multiple choice logit model is constructed and estimated. It allows to find out if the self-assessment of health status depends on categorical variables (gender, education, etc.). Also, the probabilities of self-assessment of health status for each individual can be estimated when his/her characteristics are included in the model. Gender and education were considered as categorical variables in this study in order to highlight gender and education status of respondents in self-assessment of health status. The scope of this analysis can be extended by including residence place, income (preselect and mark groups), sports practicing, etc. in the set of categorical variables. It should be noted that statistical packages like SPSS, Statistica, STATA contain modules for constructing logit and probit binary and multiple choice models.

Multiple choice models (multiple logit regression models) have similarities with the discriminant analysis, they can be effectively used in financial analyses, marketing studies or other fields for scoring. The author believes that these models have good perspectives in the practice of statistical analysis.

References

1. Anatolyev, S. (2009). Neparametricheskaya regressiya [Nonparametric regression]. K*vantil − Quantile, 7,* 37–52. Retrieved from http://www.quantile.ru/07/07-SA.pdf [in Russian].

2. Ayvazyan, S. A., & Mkhitaryan, V. S. (1998). *Prikladnaya statistika i osnovy ekonometriki [Applied statistics and foundations of econometrics].* Moscow: Yuniti. [in Russian].

3. Buyul, A. & Zöfel P. (2002). *SPSS: iskusstvo obrabotki informatsii. Analiz statisticheskikh dannykh i vosstanovlenie skrytykh zakonomernostey [SPSS: The Art of Information Processing. Analysis of statistical data and restoration of hidden patterns].* V. E. Momot (Ed.). St. Petersburg: Dia Soft UP. [in Russian].

4. Dougherty, K. (2009). *Vvedenie v ekonometriku [Introduction to Econometrics].* (3d ed.). Moscow: INFRA-M. [in Russian].

5. Eliseeva, I. I., Kurysheva, S. V., Kosteeva, T. V. (2007). *Ekonometrika [Econometrics].* Moscow: Financy i statistika. [in Russian].

6. Green, W. G. (2016). *Ekonometricheskiy analiz [Econometric Analysis].* S. S. Sinelnikov, M. Yu. Turuntseva (Eds.). Book 1. Moscow: ID "Delo" RANKhiGS. [in Russian].

7. Magnus, Ya. R., Katyshev, P. K., & Peresetskiy, A. A. (2004). *Ekonometrika. Nachalnyy kurs [Econometrics. Elementary course].* (6th ed.). Moscow: ID "Delo". [in Russian].

8. Pautova, N. I., & Pautov, I. S. (2015). Gendernye osobennosti samootsenki zdorovya i ego vospriyatiya kak sotsiokulturnoy tsennosti (po dannym 21-y volny RLMS-HSE) [Gender characteristics of health self-assessment and perception as a socio-cultural value (Based on the data of the 21st round of RLMS-HSE)]. *Zhenschina v rossiyskom obschestve − Woman in Russian Society, 2 (75),* 60–75. Retrieved from https://womaninrussiansociety.ru/wp-content/uploads/2015/06/Pautova-Pautov_64_80.pdf [in Russian].

9. Peresetskiy, A. A. (2021). *Ekonometricheskie metody v distantsionnom analize deyatelnosti rossiyskikh bankov [Econometric methods in remote analysis of the activities of Russian banks].* Moscow: NIU "Vysshaya shkola ekonomiki". [in Russian].

10. Perova, M. B. (2016). Obektivnaya i subektivnaya otsenka sostoyaniya zdorovya naseleniya Rossii [Objective and subjective assessment of population health status in Russia]. *Sistemnoe upravlenie − System management,* 1 (30). Retrieved from http://sisupr.mrsu.ru/2016-1/PDF/Perova_2016-1.pdf [in Russian].

11. Sedova, E. N. (2008). Modeli mnozhestvennogo vybora v zadachakh otsenki i upravleniya ekologo-ekonomicheskimi riskami [Multiple choice models in the tasks of assessing and managing environmental and economic risks]. *Vestnik OGU − Bulletin of OSU, 10,* 96–102. Retrieved from https://cyberleninka.ru/article/n/modeli-mnozhestvennogo-vybora-v-zadachah-otsenki-i-upravleniya-ekologo-ekonomicheskimi-riskami/viewer [in Russian].

12. Tsypina, N. Yu. (2010). Statisticheskiy analiz faktorov, okazyvayuschikh vliyanie na kachestvo predostavlyaemykh uslug naseleniyu Orenburgskoy oblasti [Statistical analysis of factors influencing the quality of services provided to the population of the Orenburg region]. *Ekonomicheskie nauki – Economic Sciences, 9 (70),* 227–231. Retrieved from https://ecsn.ru/files/pdf/201009/201009_227.pdf [in Russian].

13. Verbik, M. (2008). *Putevoditel po sovremennoy ekonometrike [A guide to modern econometrics].* S. A. Ayvazyan (Ed.). Moscow: Nauchnaya kniga. [in Russian].

14. Voischeva, O. S. (2006). Ekonometricheskie modeli kachestvennykh peremennykh v prognoznykh zadachakh marketinga [Econometric models of qualitative variables in predictive marketing problems]. *Vestnik VGU, Seriya: Ekonomika i upravleniye – Proceedings of VSU, Series: Economics and Management, 2,* 261–268. Retrieved from http://www.vestnik.vsu.ru/pdf/econ/2006/02/2006-02-42.pdf [in Russian].

15. Zsolt S. (2009). Multinomialnye modeli diskretnogo vybora [Multinomial discrete choice models]. K*vantil – Quantile, 7*, 9–19. Retrieved from http://quantile.ru/07/07-ZS.pdf [in Russian].

Л. А. Сошникова,
доктор економічних наук, професор,
професор кафедри,
Заклад освіти "Білоруський державний економічний університет",
E-mail: ludmila_sosh@mail.ru
ResearcherID: AAX-4737-2021,
ORCID: https://orcid.org/0000-0002-1402-5490

Використання логістичної регресії для аналізу результатів статистичних спостережень

У статті розглядаються питання проведення статистичного аналізу з використанням логістичних моделей упорядкованого множинного вибору, які будуються за результатами статистичних спостережень, що передбачають наявність категоріальної залежної змінної. Цю групу моделей доцільно використовувати в тому випадку, коли дискретна залежна змінна набуває кілька альтернативних значень, наприклад оцінка рівня успішності студентів (відмінно, добре, задовільно, незадовільно). Для оцінки параметрів таких моделей використовують алгоритми, засновані на елементах теорії імовірностей. Мета побудови моделі множинного вибору – визначити, які чинники та якою мірою впливають на ймовірність настання тієї чи іншої події, вибору тієї чи іншої альтернативи. У роботі досить детально викладені алгоритми розрахунку логіт-моделей бінарного і множинного вибору, а потім на прикладі конкретного завдання (статистичного аналізу результатів самооцінки здоров'я членів домашніх господарств) продемонстровано використання моделі на основі пакета SPSS. Слід зазначити, що в таких статистичних пакетах, як SPSS, STATISTICA, STATA, наявні модулі для побудови логіт- і пробіт-моделей.

Оцінка здоров'я населення охоплює об'єктивну оцінку стану його здоров'я за даними офіційної статистики про поширеність захворювань серед населення і сукупну суб'єктивну оцінку індивідуального стану здоров'я за результатами соціологічних досліджень. Важливо знати, наскільки узгоджуються об'єктивна оцінка здоров'я населення і суб'єктивне сприйняття стану здоров'я окремими індивідуумами. Оскільки первинні файли вибіркового обстеження домашніх господарств є конфіденційними, то для побудови моделі множинного вибору автором використані умовні дані, які за своїми характеристиками близькі до реальних значень. У рамках вибіркового спостереження зареєстровані такі змінні, як місце проживання, стать, вік, оцінка здоров'я, заняття спортом, куріння, дохід. При побудові моделі як залежна використовувалася змінна "здоров'я", як категоріальні змінні – "стать" і "освіта"; коваріатамі були "вік" і "дохід". Після побудови й оцінки розпізнавального сили моделі (тобто правильності передбачення залежної змінної) її специфікація зберігалась у спеціальному файлі для подальшого реконструювання.

Ключові слова: *логіт-модель, бінарний вибір, множинний вибір, логарифм шансу, метод максимальної правдоподібності, самооцінка здоров'я.*

Bibliographic description for quoting:

Soshnikova, L. A. (2021). Using the Logistic Regression in Analysis of Results from Statistical Observations. *Statystyka Ukrainy – Statistics of Ukraine, 3*, 4–11. Doi: 10.31767/su.3(94)2021.03.01

Бібліографічний опис для цитування:

Сошникова Л. А. Використання логістичної регресії для аналізу результатів статистичних спостережень (публікується англійською мовою). *Статистика України*. 2021. № 3. С. 4–11.
Doi: 10.31767/su.3(94)2021.03.01