EKONOMETRIA
ECONOMETRICS

3(45)•2014

Publishing House of Wrocław University of Economics
Wrocław 2014

Copy-editing: Marcin Orszulak

Layout: Barbara Łopusiewicz

Proof-reading: Barbara Cibis

Typesetting: Małgorzata Czupryńska

Cover design: Beata Dębska

This publication is available at www.ibuk.pl, www.ebscohost.com,
Lower Silesian Digital Library www.dbc.wroc.pl,
The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
and in The Central and Eastern European Online Library www.ceeol.com,
as well as in the annotated bibliography of economic issues of BazEkon http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

Information on submitting and reviewing papers is available on
the Publishing House's website
www.wydawnictwo.ue.wroc.pl

ISSN 1507-3866

The original version: printed
Printing: EXPOL, P. Rybiński, J. Dąbek, sp.j.
ul. Brzeska 4, 87-800 Włocławek

Contents

Streszczenia

EKONOMETRIA ECONOMETRICS 3(45) • 2014

ISSN 1507-3866

Agnieszka Sompolska-Rzechuła,
Małgorzata Machowska-Szewczyk
West Pomeranian University of Technology in Szczecin
e-mails: asompolska@zut.edu.pl; mmachowska@wi.zut.edu.pl

Anita Chudecka-Głaz, Aneta Cymbaluk-Płoska,
Janusz Menkiszak
Pomeranian Medical University in Szczecin
e-mails: anitagl@poczta.onet.pl; aneta.cymbaluk@gmail.com; nbz@list.pl

THE USE OF LOGISTIC REGRESSION IN THE OVARIAN CANCER DIAGNOSTICS

Summary: In the present elaboration an attempt has been made to build the logit model which makes it possible to specify the probability of diagnosing the ovarian carcinoma in the female patients with pathological lesion in the ovary. Based on sampling of 210 patients treated and diagnosed at the Teaching Hospital of Operative Gynaecology and Gynaecological Oncology of Women and Girls of the Pomeranian Medical University, the evaluations of the parameters of two logit models were determined and the estimation of the quality of obtained models was made. The obtained results may contribute to supporting of the ovarian cancer diagnostics.

Keywords: logit model, classification quality of female patients, ovarian cancer.

DOI: 10.15611/ekt.2014.3.11

1. Introduction

Modern medicine cannot do without statistics; the range of the applications of data analysis in this area is getting wider and wider. Statistical methods enable answering many important questions, revealing regularities, dependences and relationships occurring in medicine. One of the areas of the application of statistical methods is supporting of diagnostic decisions and it has an enormous development potential. Statistical methods may be useful e.g. for specifying the accuracy of diagnostic tests. The scope of the tasks of data analysis includes i.a. the searching for solutions of classification problems, consisting in finding such a model on the basis of which an object can be assigned to one of several classes. Building of classification models is rather

common in medical applications, e.g. diagnostics based on medical data, assigning to risk groups of falling ill with a given disease or occurrence of complications.

One of the methods serving the purpose of building the classification model is logistic regression, also referred to as the logit regression, which is very useful in a situation when the values of a dichotomous variable of *ill/healthy* or *complications/lack of complications* type should be explained by means of other variables, and the force and direction of the impact exerted by these variables on the phenomenon being analysed.

In the present elaboration an attempt has been made to build a model which can support predicting the probability of diagnosing the ovarian carcinoma in female patients with pathological lesion in the ovary.

For solving this classification problem the logistic regression has been applied. The analysis has been conducted based on data relating to patients treated and diagnosed at the Teaching Hospital of Operative Gynaecology and Gynaecological Oncology of Women and Girls of the Pomeranian Medical University in Szczecin in the period 2010–2012.

2. Methodology

The logistic regression model enables investigating the influence exerted by many independent variables $X_1, \ldots, X_k$ on the dichotomous dependent variable Y. The values of the dependent variable are coded as follows: 1 – the distinguished value – possessing the feature, 0 – not possessing the feature. In the logit regression the logistic function is used, which has values from the range $(0;1)$ and curve resembling the stretched S letter, whose analytical form is as follows [Stanisz 2007]:

$$f(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}, \quad z \in R. \tag{1}$$

The logistic regression model for the dichotomous variable Y specifies the conditional probability of taking by this variable the distinguished value and it is expressed by the following dependence [Maddala 2001; Stanisz 2007]:

$$P(Y = 1 / X_1, ..., X_k) = \frac{e^{\alpha_0 + \alpha_1 X_1 + ... + \alpha_k X_k}}{1 + e^{\alpha_0 + \alpha_1 X_1 + ... + \alpha_k X_k}}, \tag{2}$$

where $\alpha_0, \alpha_1, ..., \alpha_k$ are parameters of the model, $X_1, ..., X_k$ are independent variables that may have both the qualitative and the quantitative character.

The logistic regression model coefficients can be searched for with the maximum likelihood method [Dobosz 2004] or with the ordinary least squares method [Gruszczyński, Podgórska 1996].

Due to the model (2) nonlinearity in relation to the independent variables and parameters, by finding the logarithm the logistic model is transformed into the linear model. For this purpose the Odds Ratio concept is introduced, which is the ratio of the probability of a given event occurrence to the probability that the said event shall not occur, that is:

$$\frac{P(Y=1/X_1,...,X_k)}{1-P(Y=1/X_1,...,X_k)}=\frac{e^{\alpha_0+\alpha_1X_1+...+\alpha_kX_k}}{1+e^{\alpha_0+\alpha_1X_1+...+\alpha_kX_k}}:\frac{1}{1+e^{\alpha_0+\alpha_1X_1+...+\alpha_kX_k}}= \\ =e^{\alpha_0+\alpha_1X_1+...+\alpha_kX_k} \tag{3}$$

So the Odds Ratio expresses how many times the probability that a given event will take place increases or decreases, if there occurs a change of independent variable (at established values of independent variables).

The natural logarithm of the Odds Ratio is linear in relation to independent variables and considering the model parameters, which facilitates estimation to a high degree. It is called the logit or the logit form of the logistic model; therefore [Stanisz 2007; Cramer 2003; Kleinbaum, Klein 2002]:

$$\log itP=\ln\frac{P(Y=1/X_1,...,X_k)}{1-P(Y=1/X_1,...,X_k)}=\alpha_0+\sum_{i=1}^{k}\alpha_iX_i\,. \tag{4}$$

After estimating the parameters of the logistic regression model, it is possible to determine the theoretical values of the variable Y according to the standard principle of prediction:

$$\hat{y}_i=\begin{cases}1, & \text{for } 0.5<\hat{p}_i\le 1\\ 0, & \text{for } 0<\hat{p}_i\le 0.5\end{cases}, \tag{5}$$

where $\hat{p}_i$ – theoretical probabilities obtained from the logistic regression model estimated on the basis of a random sample.

In a situation when a sample is unbalanced, that is those in which the number of "ones" differs considerably from the number of "zeroes", for predicting theoretical values one can apply the standard principle modification and count predictions according to the optimum boundary value principle α:

$$\hat{y}_i=\begin{cases}1, & \text{for } \alpha<\hat{p}_i\le 1\\ 0, & \text{for } 0<\hat{p}_i\le \alpha\end{cases}. \tag{6}$$

The boundary value α is established as the fraction of "ones" in a sample. Then the evaluation of the correctness of the estimated model can be carried out, counting correctly and mistakenly the classified cases (see Table 1).

Table 1. Correctness of classification of cases

Expected sizes	Observed sizes		Sum
	$y_i = 1$	$y_i = 0$	
$\hat{y}_i = 1$	n_{11}	n_{12}	$n_{1\bullet}$
$\hat{y}_i = 0$	n_{21}	n_{22}	$n_{2\bullet}$
Sum	$n_{\bullet 1}$	$n_{\bullet 2}$	n

Source: the authors' own elaboration on the basis of [Dobosz 2004].

For evaluating the degree of the logistic regression model fitting to empirical data, one can use the measure called *count*-R^2, which takes values from the range $\langle 0,1 \rangle$, defined as follows [Maddala 2008]:

$$R^2_{count} = \frac{n_{11} + n_{22}}{n}. \tag{7}$$

The more the measure value is approximate to one, the better fit of the logistic model to empirical data of the investigated phenomenon is obtained; R^2_{count} stands for the percentage of correctly classified cases. The model turns out to be good in the prediction of an investigated phenomenon, when $R^2_{count} > 50\%$. This means that the classification based on the model is better than that the random one.

The quality of the built-up logistic regression model can be also evaluated by means of other measures e.g.:

- Hosmer–Lemeshow test [Hosmer et al. 2004; Hosmer et al. 2008] – a test, which for various data subgroups compares the observed sizes of an occurrence in a given subgroup of objects having the distinguished feature O_g as well as the expected sizes E_g of a distinguished value occurrence. If O_g and E_g are sufficiently similar, then one can assume that a well-fit model has been built. Usually, for the sake of calculations, observations are subdivided into G subgroups, using e.g. decyls. The hypotheses in the test take the following form:

$$H_0 : O_g = E_g \text{ for all groups,}$$

$$H_1 : O_g \neq E_g \text{ for at least one group.}$$

The value of test statistic is determined in the following manner:

$$H = \sum_{g=1}^{G} \frac{\left(O_g - E_g\right)^2}{E_g\left(1 - \frac{E_g}{N_g}\right)}, \tag{8}$$

where N_g – number of observations in group g, $g \in \{1, ..., G\}$.

This statistic has got asymptotically the distribution χ^2 with $G-2$ degrees of freedom.

- AUC (*Area Under the Curve*) [DeLong, DeLong, Clarke-Pearson 1988] – area under the ROC curve (*Receiver Operating Characteristic Curves*) – The ROC curve is formed by connecting the points in the cartesian coordinate system having the coordinates (sensitivity, 1-specificity).

Sensitivity describes the ability to detect units which have a distinguished characteristic. It is determined as follows:

$$sensitivity = \frac{n_{11}}{n_{\bullet 1}}. \tag{9}$$

Specificity describes the ability to detect units which do not have a distinguished characteristic. It is determined as the ratio of the number of observations not having the distinguished characteristic and correctly classified, to the number of all the observations not having the distinguished characteristic. It is determined according to the following formula:

$$specificity = \frac{n_{22}}{n_{\bullet 2}}. \tag{10}$$

The ROC curve is related to cut-off point. The point is a certain value of the diagnostic variable, which provides the separation of the studied population into two groups: in which the given phenomenon occurs and in which the given phenomenon does not occur.

On the basis of the logit model one can estimate the values of the probability of a phenomenon occurrence and consider it as the values of a diagnostic variable.

Let us assume that we have at our disposal a sample of *N* elements, in which each object has one of the *N* values of the diagnostic variable $\hat{p}_1, ..., \hat{p}_N$. Each of the received values of the diagnostic variable becomes the cut-off $\hat{p}_{cut}$. If the value of the diagnostic variable $\hat{p}_i \geq \hat{p}_{cut}$, then objectives are classified into a group in which a given phenomenon occurs ($\hat{y}_i = 1$); otherwise, it is considered that for objects a phenomenon does not occur ($\hat{y}_i = 0$). On the basis of those values for each cut-offs, we define the matrix of the classification of cases (see Table 1) and calculate sensitivity and specificity. Thus, we obtain *N* matrixes and *N* points. The ROC curve is created on the basis of the calculated values of sensitivity and specificity. On the abscissa axis the 1-specificity is placed, and on the ordinate axis – sensitivity. The points obtained in that manner are linked. The constructed curve, especially the area under the curve, presents the classification quality of the analyzed diagnostic variable.

Thus, the resulting ROC curve, especially the field under the curve, represents the discriminating quality of the model. When the ROC curve and the diagonal $x = y$ coincide, the decision on assigning a given case to a chosen class, made on the basis

of the model, is as good as the random assignment of the investigated cases to these groups. The classifying quality of the model is good when the area under the ROC curve is considerably bigger than the field under the straight line $x = y$, therefore bigger than 0.5.

If the growth of the diagnostic variable value is accompanied by an increase or a decrease in chances for the occurrence of an investigated phenomenon, then for the ROC curve the so-called optimum cut-off point of the predicted probability is searched for, for which the dependent variable subdivides the population into groups in the best manner: in which an investigated phenomenon occurs and does not occur. The selection of the optimum cut-off point requires professional knowledge relating to the scope of an investigation subject. Applying the advanced mathematical apparatus, one can find the point which will be the most advantageous one in terms of mathematics [Dobosz 2004].

The optimum cut-off value is calculated on the basis of sensitivity, specificity and using the quantity m − slope of the tangent to the ROC curve. Such a value of the diagnostic variable should be considered as the optimum cut-off point at which the expression: $Sensitivity - m \cdot (Specificity - 1)$ reaches the minimum [Zweig, Campbell 1993].

3. Materials

The ovarian epithelial cancers constitute a great challenge to the gynaecologist-oncologist circles. Among other reasons, due to diagnostic difficulties, especially at an early stage of the clinical progression. At present, apart from the medical imaging diagnostics methods, also laboratory investigations of tumour markers are widely used [Nolen 2010]. Most often tumour markers are macromolecular substances that may be found in blood, urine or which are associated with the surface of neoplastic cells, whose identification and measurement are useful in diagnosing patients and in planning their therapies. The most known marker which has been commonly used for many years in ovarian cancer diagnostics is the CA125 antigen. It is characterized by high sensitivity but relatively low specificity especially for women in the premenopausal period in which it may be elevated in quite numerous benign pathological states occurring in female genitals [Zurawski et al.1988]. For a long time additional markers were searched for, which could correct the specificity of differential diagnostics of ovarian cancers. It seems that such attributes characteristize the HE4 marker and the ROMA algorithm using both the CA125 and HE4 markers taking also into account the patient's hormonal status (pre- or post-menopausal) [Hellström et al. 2003; Moore et al. 2010; Moore et al. 2011; Montagnana et al. 2011]. The high effectiveness of the ROMA algorithm was described for the first time by Moore [Moore et al. 2010], who demonstrated that out of 9 investigated markers it is just the combination of CA125 and HE4 that is characterized by the best sensitivity and specificity. This method is also recommended by FDA (Food and Drug Administration).

The following set of potential diagnostic characteristics has been assumed for evaluating the probability of diagnosing the ovarian cancer in patients with pathological lesion in the ovary:

X_1 – age in years,
X_2 – hormonal status (pre- or post-menopausal),
X_3 – HE4 level (normal, above normal),
X_4 – CA125 level (normal, above normal),
X_5 – ROMA value (normal, above normal),
Y – ovarian cancer (0 – not diagnosed, 1 – diagnosed).

The CA125 level was determined by means of commercial tests – Architect i2000, Abbott Diagnostics, Abbott Park, Il, USA. The cut-off point value was equal to 35 U/ml.

The serum concentrations of HE4 by method of electrochemiluminescence "ECLIA" were determined using commercial test Elecsys HE4 produced by the Roche company, by means of cobas e 601 analyzer. As the upper norm limit and at the same time the optimum cut-off point, according to producers' recommendations, the value of 70 pmol/L has been assumed. The CA125 and HE4 tests were performed in compliance with the instruction supplied by the producers and the relevant control values were within the normal range.

The predictive index was calculated according to the following formulae:

- for premenopausal patients: $\mathrm{PI} = -12.0 + 2.38\ln(\mathrm{HE4}) + 0.0626\ln(\mathrm{CA125})$,
- for postmenopausal patients: $\mathrm{PI} = -8.09 + 1.04\ln(\mathrm{HE4}) + 0.732\ln(\mathrm{CA125})$.
- The ROMA value is determined according to the following formula:

$$\mathrm{ROMA}(\%) = \frac{\exp(\mathrm{PI})}{1+\exp(\mathrm{PI})}\cdot 100, \text{ where } \exp(\mathrm{PI}) = e^{\mathrm{PI}}. \tag{11}$$

The standard cut-off point value for ROMA was assumed 13.1% for the women in the premenopausal period and 27.7% for the women in the postmenopausal period [Moore et al. 2010].

The patients were considered as postmenopausal when the period of 12 months or shorter lapsed from the last menstruation (menopausa) but the FSH value was above 30 mIU/ml.

In order to find the best combination of characteristics significantly influencing the diagnosing of ovarian cancer, the formal selection of features was carried out by means of the backward stepwise regression and two sets of variables were obtained: X_1, X_3, X_4, X_5, as well as X_2, X_3, X_4. They create the new lists of variables which are weakly intercorrelated and – at the same time – strongly correlated with other features eliminated from the set of diagnostic characteristics.

The sampling concerns 210 patients randomly selected from all the patients that were treated and diagnosed at the Teaching Hospital of Operative Gynaecology and Gynaecological Oncology of Women and Girls of the Pomeranian Medical Universi-

ty in Szczecin in the period 2010–2012, in 84 of them (making 40%) the ovarian cancer was diagnosed. The diagnoses were established on the basis of the histopathological examination (it is the only method of definitive diagnosis, all other diagnoses are of auxiliary character). The measurements of the markers level HE4, CA125, Roma were carried out at the Central Laboratory of autonomous Public Clinical Hospital No. 2 of Pomeranian Medical University in Szczecin.

In the sampling women aged from 44 to 61 years were prevailing (more than 27%). The second most numerous group was constituted by women whose age ranged from 27 to 44 years (more than 21%). The youngest patients aged from 12 to 27 years represented *circa* 21% of the group. The group which was not so numerous, making 8% included patients, aged over 78 years. The youngest of the examined persons was aged 12 years, whereas the oldest one – 90 years. The average age in the sampled group was slightly over 46 years.

In 40% of the patients the above normal HE4 level was observed, in the case of other persons the HE4 value was within the normal range. However, in terms of CA125 level, 46% of the patients were characterized with values complying with the standard, 54% of them had CA125 values above normal. When analysing patients with their hormonal status being taken into account, for 77% of the women before menopause the normal HE4 level was observed, whereas for the women after menopause only 38% of them had a correct level of this marker. In turn, the CA125 level not exceeding normal occurred for 56% of the women before menopause and in 32% of them after menopause (see Figure 1, 2).

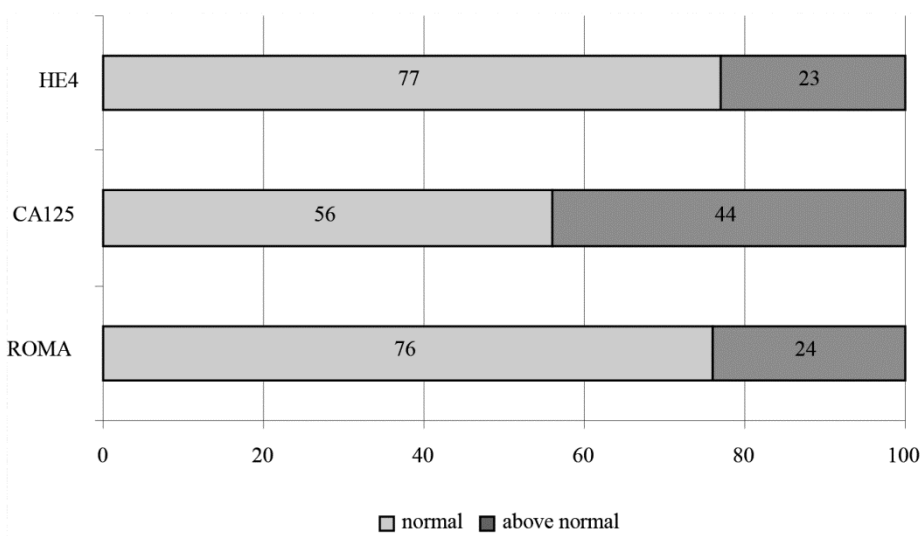

Figure 1. The structure of the patients before menopause in terms of HE4, CA125 and ROMA levels

Source: the authors' own elaboration.

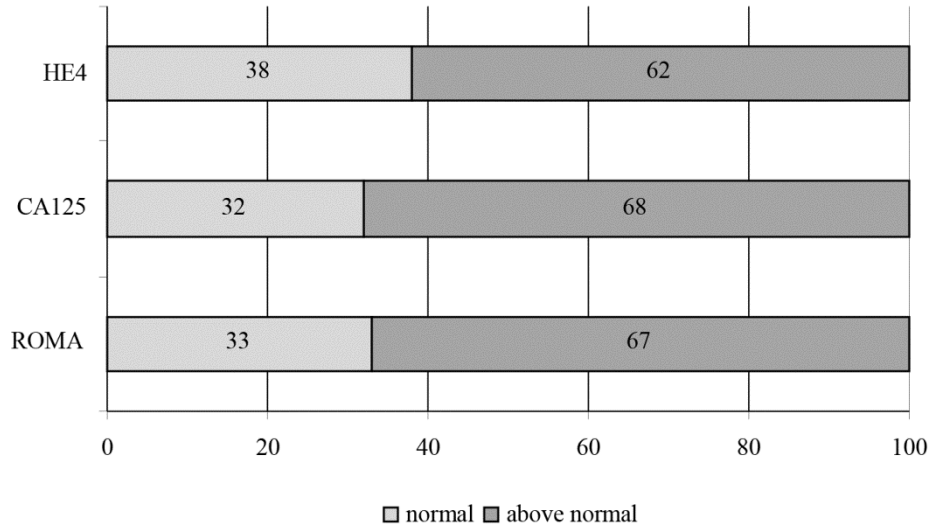

Figure 2. The structure of the patients after menopause in terms of HE4, CA125 and ROMA levels

Source: the authors' own elaboration.

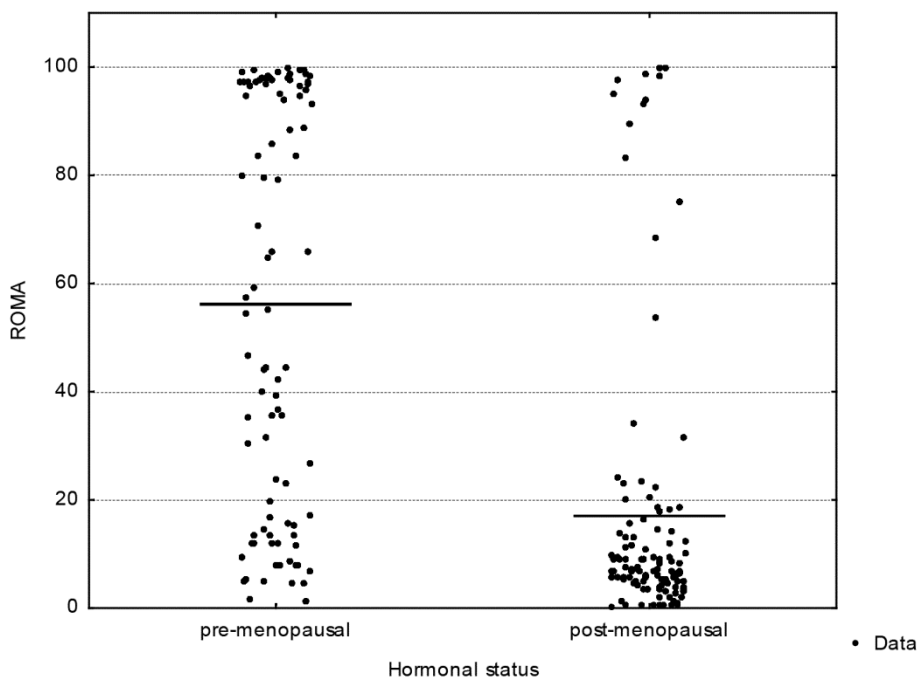

Figure 3. Scatter diagram of the ROMA values with patients' status taken into account

Source: the authors' own elaboration.

In the sample group of the patients for 57% of them the normal level of ROMA was observed, whereas 76% of the patients before menopause had the normal level of ROMA, and after menopause – 33%. The scatter of ROMA values with the status taken into account is presented in Figure3. The horizontal line indicates the average level of the ROMA value in the sample group.

The distribution of ROMA values for the women before menopause is characterized by strong right-sided asymmetry, which means that in the sample group most of the patients has a lower level of ROMA values as compared with the average value of ROMA. However, in the group of the women after menopause the distribution of ROMA values of U-shaped was observed, that is women having extreme levels of ROMA are prevailing.

4. Discussion of results

Estimates of the logit model parameters taking into account the first set of variables obtained in the case of backward regression are presented in Table 2.

Table 2. Estimates of the parameters of the logit model

Variable	Variable's name	Parameter's estimation	*p*-value	Odds Ratio
	Intercept	−10,7290	0.0000	–
X_1	Age (in years)	0.0547	0.0010	1.0562
X_3	HE4 level	1.6959	0.0028	5.4518
X_4	CA125 level	1.9887	0.0002	7.3063
X_5	ROMA value	1.2769	0.0317	3.5854

Source: the authors' own elaboration.

In the model, the following data have the positive, statistically significant influence on the dependent variable: age, HE4, CA125, ROMA levels.

Interpreting the Odds Ratios at i^{th} variable (assuming that the other variables taken into account in the model will remain unchanged), the following information is obtained:

- in each case when a woman is one year older, the odds of diagnosing the ovarian cancer increases by more than 5%;
- if the HE4 level increases above normal, the odds of diagnosing the ovarian cancer increases more than fivefold;
- exceeding the CA125 norm involves increasing the odds of diagnosing the ovarian cancer more than sevenfold;
- the odds of diagnosing the ovarian cancer is 3.6 times higher for the women with the ROMA level exceeding the norm in comparison to the patients for which the ROMA level is normal.

The evaluation of the estimated model validity was carried out by counting the correctness of the classification of women on the basis of the determined cut-off points 0.593 for model 1 and 0.50 for model 2 (see Tables 3, 4).

Table 3. Correctness of the classification of the logit model

Expected sizes	Observed sizes		Correctness of classification
	$y_i = 1$	$y_i = 0$	
$\hat{y}_i = 1$	71	10	89.05%
$\hat{y}_i = 0$	13	116	
Sensitivity, specificity	84.52%	92.06%	

Source: the authors' own elaboration.

The estimates of the logit model parameters taking into account the second set of the variables obtained in the case of the backward regression are presented in Table 4.

Table 4. Estimates of the logit model parameters

Variable	Variable's name	Parameter's estimation	*p*-value	Odds Ratio
	Intercept	−10.2177	0.000001	0.000037
X_2	Hormonal status	1.3972	0.001875	4.0438
X_3	HE4 level	2.5587	0.000001	12.9195
X_4	CA125 level	2.4376	0.000001	11.4456

Source: the authors' own elaboration.

In the model, the following data have the positive, statistically significant influence on the dependent variable: status, HE4 and CA125 levels.

Interpreting the Odds Ratios at i^{th} variable (assuming that the other variables taken into account in the model will remain unchanged), the following information is obtained:

- the entering of the postmenopausal period by a woman results in more than fourfold increase of the odds of diagnosing the ovarian cancer;
- if the HE4 level increases above normal, the odds of diagnosing the ovarian cancer increases more than twelvefold;
- exceeding the CA125 norm involves increasing of the odds of diagnosing the ovarian cancer more than elevenfold.

The evaluation of the estimated model validity was carried out by counting the correctness of the classification of women (see Table 5).

The correctness of classification was estimated by means of coefficient R^2_{count}, Hosmer-Lemeshow test and area under ROC curve. The results are presented in Table 6.

Table 5. Correctness of the classification of the logit model

Expected sizes	Observed sizes		Correctness of classification
	$y_i = 1$	$y_i = 0$	
$\hat{y}_i = 1$	76	16	88.57%
$\hat{y}_i = 0$	8	110	
Sensitivity, specificity	90.48%	87.30%	

Source: the authors' own elaboration.

Table 6. The degree of the logistic regression models fitting to the empirical data

Models	Measures of the logistic regression models fitting	Hosmer-Lemenshow test		AUC
	R^2_{count}	χ^2	p	
Model 1	89,05%	5,59	0,693	93,84%
Model 2	88,57%	3,35	0,501	92,52%

Source: the authors' own elaboration.

Based on the results included in Table 6, one can find that both models of the logistic regression are characterized by a very good fitting to the empirical data. The coefficients R^2_{count} in both models demonstrate very good the correctness of the classification of patients, obtained on the basis of the logistic regression. The Hosmer-Lemenshow test results indicate the lack of significant differences between the empirical sizes and the theoretical ones which result from the estimated models of logistic regression.

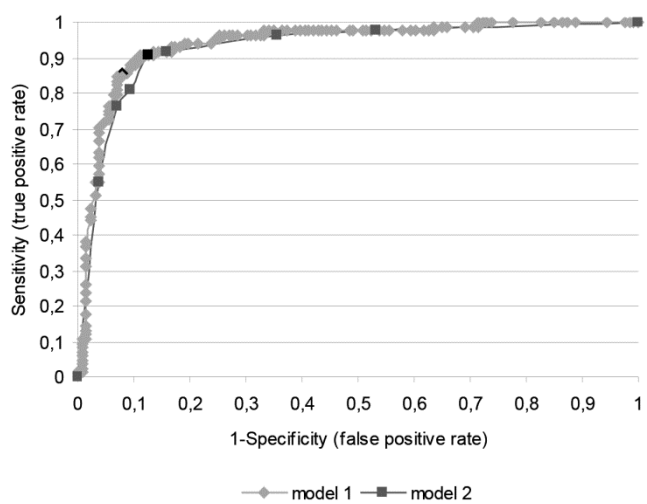

Figure 4. The ROC curve for models 1 and 2

Source: the authors' own elaboration.

When analyzing the diagrams presented in Figure 4 and the calculations in Table 6, one can find that the field under the ROC curve both in model 1 and 2 is significantly bigger than 0.5, (at the significance levels higher than 0.000001 in both models). Therefore, it is possible to classify patients on the basis of the built models. The calculated cut-off point value for the first model is equal to 0.593. The classification determined based on this cut-off point yields 89.05% of the correctly classified cases, out of which 84.52% to "yes" and 92.06% to "no". For the second model the calculated cut-off point value was equal to 0.503 and the percentage of correctly classified patients was 88.57%, out of which 90.48% of correctly estimated as having diagnosed cancer, whereas 87.30% of the cases rightly considered as not having diagnosed cancer. In Figure 4 the cut-off points in both models are marked black.

5. Conclusions

In the present elaboration the logistic regression model has been built which makes it possible to specify the probability of diagnosing the ovarian carcinoma in female patients with pathological lesion in the ovary. By means of the backward stepwise regression the two sets of factors influencing the occurrence of ovarian cancer were distinguished. The quality of the constructed models has been estimated using the correctness of classification coefficients, the Hosmer-Lemenshow test and the ROC curve. Both models are characterized by a very high quality. The model, taking into account the HE4, CA125, ROMA levels and age, has turned out to be slightly better and the following parameters in it have turned out to be the most significant ones, differentiating the probability of the occurrence of ovarian cancer: CA125 and age. However, in the model with diagnostic variables: hormonal status, HE4, CA125 both markers are significant. All the variables, both in model 1 and 2, have a positive influence on the probability of the occurrence of the ovarian carcinoma in female patients with pathological lesion in the ovary. It means that the increase of the value of any of the variables (assuming that the other variables taken into account in the model will remain unchanged) causes the increase of the probability of diagnosing the ovarian cancer.

Taking into account the combination of a number of factors influencing the diagnosing of ovarian cancer constitutes an alternative for discrimination obtained on the basis of single diagnostic parameters. The proposed method may contribute to supporting the differential diagnostics of pathological lesion in the ovary.

References

Cramer J.S., 2003, *Logit Models from Economics and Other Fields*, Cambridge University Press, Cambridge.

DeLong E.R., DeLong D.M., Clarke-Pearson D.L., 1988, Comparing the areas under two or more correlated receiver opera ng curves: A nonparametric approach, *Biometrics*, vol. 44, pp. 837–845.

Dobosz M., 2004, *Wspomagana komputerowo statystyczna analiza wyników badań*, Akademicka Oficyna Wydawnicza EXIT, Warszawa, p. 260.
Gruszczyński M., Podgórska M., 1996, *Ekonometria*, Oficyna Wydawnicza Szkoły Głównej Handlowej, Warszawa, pp. 139–141.
Hellström I., Raycraft J., Hayden-Ledbetter M. et al., 2003, The HE4 (WFDC2) protein is a biomarker for ovarian carcinoma, *Cancer Res*, vol. 63(13), pp. 3695–700.
Hosmer D.W., Lemenshow S., May S., 2008, *Applied Survival Analysis: Regression Modeling of Time to Event Data*, John Wiley & Sons, New York.
Hosmer D.W., Lemenshow S., 2004, *Applied Logistic Regression*, John Wiley & Sons, New York.
Kleinbaum D.G., Klein M., 2002, *Logistic Regression*, Springer, New York.
Maddala G.S., 2001, *Introduction to Econometrics*, 3rd ed. John Wiley & Sons.
Montagnana M., Danese E., Ruzzenente O. et al., 2011, The ROMA (Rik of Ovarian Malignancy Algorithm) for estimating the risk of epithelial ovarian cancer in women presenting with pelvic mass: is it really useful? *Clin Chem Lab Med*., vol. 49(3), pp. 521–5.
Moore R.G., Jabre-Raughley M., Brown A.K. et al., 2010, Comparison of a novel multiple marker assay vs the Risk of Malignancy Index for the prediction of epithelial ovarian cancer in patients with a pelvic mass, *Am J Obstet Gynecol*, vol. 203(3), p. 228, e1-6.
Moore G., Miller M.C., Disilvestro P. et al., 2011, Evaluation of the diagnostic accuracy of the risk of ovarian malignancy algorithm in women with a pelvic mass, *Obstet Gynecol*, vol. 118 (2 Pt 1), pp. 280–8.
Nolen B., Velikokhatnaya L., Marrangoni A. et al., 2010, Serum biomarker panels for the discrimination of benign from malignant cases in patients with the adnexal masses, *Gynecol Oncol*, vol. 117, pp. 440–445.
Stanisz A., 2007, *Przystępny kurs z zastosowaniem Statistica PL na przykładach z medycyny*. Statsoft, Kraków.
Zurawski V.R. Jr., Knapp R.C., Einhorn M. et al., 1988, An initial analysis of preoperative serum CA 125 levels in patients with early stage ovarian carcinoma, *Gynecol Oncol*, vol. 30, pp. 7–14.
Zweig M.H., Campbell G., 1993, Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine, *Clinical Chemistry*, vol. 39, pp. 561–577.

WYKORZYSTANIE REGRESJI LOGISTYCZNEJ W DIAGNOSTYCE RAKA JAJNIKA

Streszczenie: W pracy podjęto próbę budowy modelu logitowego, który umożliwia określenie prawdopodobieństwa rozpoznania raka jajnika u pacjentek z patologiczną zmianą w jajniku. Na podstawie próby 210 pacjentek leczonych i diagnozowanych w Klinice Ginekologii Operacyjnej i Onkologii Ginekologicznej Dorosłych i Dziewcząt Pomorskiego Uniwersytetu Medycznego wyznaczono oceny parametrów dwóch modeli regresji logitowej oraz dokonano oceny jakości otrzymanych modeli. Uzyskane wyniki mogą przyczynić się do wspomagania diagnostyki raka jajnika.

Słowa kluczowe: model logitowy, jakość klasyfikacji pacjentek, rak jajnika.