STATISTICS IN TRANSITION-new series, October 2010
Vol. 11, No. 2, pp. 381—402

USING ROC CURVES TO FIND THE CUT-OFF POINT IN LOGISTIC REGRESSION WITH UNBALANCED SAMPLES

Janusz Śmigielski[1][2], Anna Majdzińska[3], Witold Śmigielski[3]

1. Introduction

Logistic regression is widely used in many fields of science, e.g. medicine, psychology and anthropology. It was introduced in the 19th c. and its graphic form was developed by P. F. Verhulst and R. F. Pearl, who were the first to use the logistic model in practice to model population increase. The full model of logistic regression was used by Finney in 1972 [Stanisz, 2000, p. 205].

Logistic regression is used in models with a dichotomous endogenous variable, i.e. one taking only values 0 and 1 (e.g. healthy persons Y_i=0, sick persons Y_i=1). The probability of variable Y_i taking value 0 or 1 can be estimated by means of maximum likelihood method (see section 2). As far as the dichotomous variable is concerned, a frequently occurring problem is unbalanced sample, i.e. having the number of the values Y_i=0 considerably different from the number of values Y_i=1, for example, the number of healthy persons is usually much larger than of the sick ones. The classical method to find the cut-off point for the estimated probability P(Y_i=1) (in order to transform this probability into values 0 or 1 of the endogenous variable), may turn out quite ineffective. Therefore, the optimal cut-off value should be sought with methods other than the classical ones.

In the paper we propose using the concept of the receiver operator characteristic (ROC) curves in order to find an optimal cut-off point in logit models based on unbalanced samples. The proposed method will be compared with some other popular methods discussed in the literature.

[1] The authors wish to express their gratitude to Prof. dr. hab. Agnieszka Rossa for her valuable comments and guidelines, as well as all other assistance she extended during the writing of the article.
[2] Department of Medical Informatics and Statistics, Medical University in Łódź.
[3] Department of Demography, University of Łódź.

2. Logistic regression – the model's theoretical foundations

2.1. The form of a logistic regression model

It has already been mentioned in the introduction that the logit models are used to describe the influence of a number of exogenous variables $X_{1i}, X_{2i},..., X_{ki}$ on a dichotomous variable Y_i, which is a variable presenting two states of a described phenomenon observed for an i-th unit, i.e. 1 (a success) and 0 (a failure).

One reason for using a logit model in such a case flows among others from the undesirable properties of the linear regression model having a dichotomous variable as a dependent variable. One of the main drawbacks of the linear regression model with a binary dependent variable is that the model does not guarantee probability estimates $P(Y_i=1)$ within the interval [0,1]. Moreover, because the random error is heteroscedastic in this case and does not have the normal distribution, applying the Least Squares Method (LSM) to estimating the model parameters is not effective and the estimates of the error's variance can even be negative.

The traditional linear regression model is as follows:

$$Y_i = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + ... + \alpha_k x_{ki} + \varepsilon_i, \tag{1}$$

where $x_{1i}, x_{2i},..., x_{ki}$ are known values of the variables $X_{1i}, X_{2i},..., X_{ki}$, while ε_i stands for a random term whose expected value equals 0.

When variable Y_i is dichotomous, the conditional expected value of variable Y_i (given assumed values of the explanatory variables) is expressed by the formula:

$$E(Y_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i) = 1 \cdot P(Y_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}_i)$$

$$+ 0 \cdot P(Y_i = 0 \mid \boldsymbol{X}_i = \boldsymbol{x}_i) = P(Y_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}_i) \tag{2}$$

where $\boldsymbol{X_i} = [X_{1i}, X_{2i},..., X_{ki}]^T$, $\boldsymbol{x_i} = [x_{1i}, x_{2i},..., x_{ki}]^T$.

Denoting $p_i = P(Y_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}_i)$, we obtain from (1) and (2):

$$p_i = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + ... + \alpha_k x_{ki} \tag{3}$$

The structural parameters in (3) show the amount of change in probability p_i (i.e. its increase or decrease) for a unit growth in the value of the exogenous variable at the given parameter, when the values of all other variables stay unchanged. The conclusion can be formulated that the structural parameters in the model (3) should not be outside the interval [-1,1], otherwise a unit growth in the exogenous variable will be accompanied by a change in probability p_i larger than a unit, which is unacceptable for probability. Therefore, if the obtained parameter

estimates do not belong to the interval, then their interpretation, as presented above, becomes pointless [Gruszczyński, 2002, pp. 54–56; Pruska, 2001, p. 91].

To eliminate such interpretation problems, the right-hand side of equation (3) can be transformed to obtain estimates p_i always within the interval [0,1]. This can be done by replacing (3) with the model:

$$p_i = F(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + ... + \alpha_k x_{ki}) \tag{4}$$

or, equivalently,

$$Y_i = F(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + ... + \alpha_k x_{ki}) + \varepsilon_i . \tag{5}$$

where F is a continuous and increasing function having following properties $\lim_{z\to\infty} F(z) = 1,\ \lim_{z\to-\infty} F(z) = 0.$ Such a model is called a binomial model.

In practice, the cumulative distribution function (CDF) of the logistic distribution[1] is frequently applied, which is defined by the formula:

$$F(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}} . \tag{6}$$

Assuming therefore that F is the CDF of the logistic distribution, equation (4) can be written as follows [see Stanisz, 2000, pp. 207–208]:

$$p_i = P(Y_i = 1 \mid \mathbf{X}_i = \mathbf{x}_i) = \frac{e^{\left(\alpha_0 + \sum_{j=1}^{k} \alpha_j x_{ji}\right)}}{1+e^{\left(\alpha_0 + \sum_{j=1}^{k} \alpha_{ji} x_{ji}\right)}} \tag{7}$$

and equivalently

$$q_i = 1 - p_i = P(Y_i = 0 \mid \boldsymbol{X}_i = \boldsymbol{x}_i) = 1 - \frac{e^{\left(\alpha_0 + \sum_{j=1}^{k} \alpha_j x_{ji}\right)}}{1+e^{\left(\alpha_0 + \sum_{j=1}^{k} \alpha_{ji} x_{ji}\right)}} \tag{8}$$

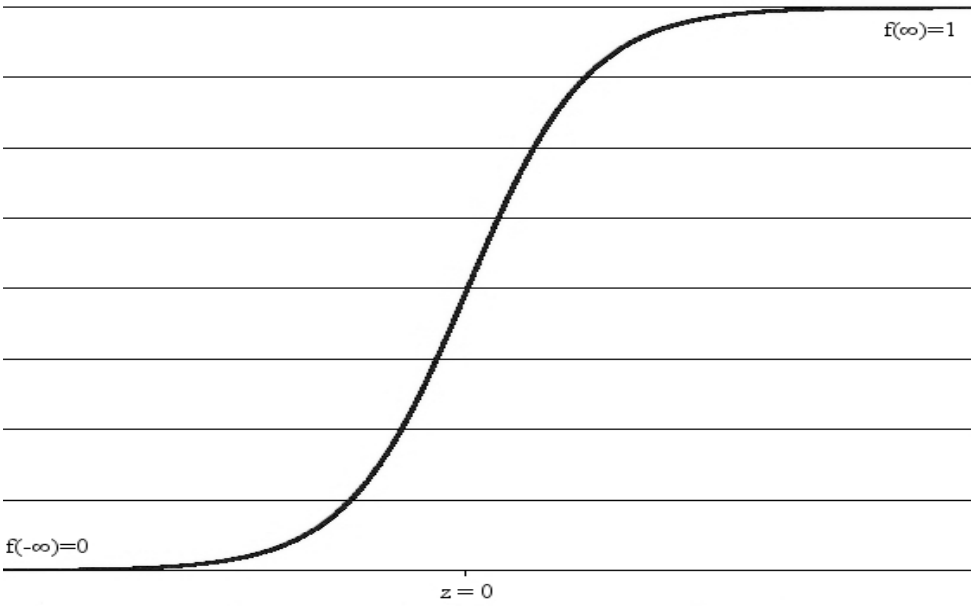
where α_j, $j = 0,...,k$ are the regression coefficients. The special case of binomial model defined in (7) is called a logistic regression model or a logit model.

Further, probabilities p_i and q_i in the models (7) and (8) will be generally denoted as $p(y_i \mid x_1, x_2, ..., x_k)$, where $y_i = 1$ or $y_i = 0.$

The example of the logistic function is graphically presented in Figure 1.

[1] To this end, the CDF of any continuous distribution can be applied, but usually the CDFs of the logistic or normal distribution are used.

Figure 1. Graphic representation of the logistic function



Source: developed by the authors.

The logistic curve is *S*-shaped. The probability p_i varies within the interval [0,1], so it can be treated as the probability of success occurring for an object characterised by the linear combination $z_i = \alpha^T \boldsymbol{x}_i$ [Gruszczyński et al., 2009, p. 142]. Initially, the function takes values close to 0, but after reaching some threshold value it grows rapidly and takes values approaching 1.

As mentioned, in the logistic model, the dichotomous dependent variable Y_i for the *i*-th unit takes two values denoted as 1 (a success) and 0 (a failure). In other words, the logistic model relates the probability of occurrence of either of the two possible results of variable Y_i to variables $\boldsymbol{X}_i$.

2.2. The likelihood function and parameter estimation for the logit model

Coefficients α_j of the logit model are usually estimated using the Maximum Likelihood Method (ML). A sufficiently large sample is needed for the purpose, i.e. $n > 10(k+1)$, where n is the sample size and k stands for the number of model parameters.

Let $[x_{1i}, x_{2i}, ..., x_{ki}]^T$ be the observation vector for the *i*-th sample unit and $\alpha_1,...,\alpha_k$ the unknown parameters of the logistic regression function to be estimated based on the sample. The likelihood function is given by the formula:

$$L = \prod_{i=1}^{n} p(y_i \mid x_{1i},...,x_{ki}) \tag{9}$$

where $p(y_i \mid x_{1i},...,x_{ki})$ stands for the probability of the dependent variable Y_i in the given regression model taking value y_i for known values x_{1i}, x_{2i}, ..., x_{ki}.

When Y_i equals 1 or 0, then the probability (9) can be written as:

$$L = \prod_{i=1}^{n} p_i^{y_i} (1-p_i)^{1-y_i} \tag{10}$$

where p_i denotes the probability of $Y_i = 1$. In the case of a logistic regression model this probability is given by (7).

It is assumed in this paper that the estimators of the structural parameters $\alpha_1,...,\alpha_k$ are the ML-estimators [see Stanisz, 2000, pp. 208-209]. „An ML-estimator is a vector of parameters ensuring the largest probability of obtaining the observed values of the variables" [see Welfe A., 2003, p. 55].

When probabilities p_i are given by the formula (7), then finding the values of the parameters $\alpha_1,...,\alpha_k$ that maximise the function L (or, equivalently, the function ln L) involves the application of relevant numerical methods, such as the Newton-Raphson algorithm.

2.3. The odds ratio and the interpretation of logit model's parameters

After transforming (7) , we obtain

$$\ln \frac{p_i}{1-p_i} = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + ... + \alpha_k x_{ki} \tag{11}$$

The expression under logarithm on the left-hand side in (11)

$$\frac{p_i}{1-p_i} \tag{12}$$

is known as an odds ratio, as it shows the probability of event (a success) occurrence in relation to the probability of its non-occurrence (a failure) [see Gruszczyński, 2001, p. 58; Gruszczyński et al., 2009, p. 169]:

Model (11) is equivalent to model (8). It is linear with respect to its parameters $\alpha_1,...,\alpha_k$ and with respect to the explanatory variables. The left-hand side of equation (11) is called a logit[1], i.e. it is the logarithm of the odds ratio.

[1] Analogous transformation of (4) where F stands for the CDF of the normal distribution is called a probit model.

The odds ratio is a tool applied to interpreting the parameters of the logit model. In general terms, the interpretation is as follows: if an *j*-th explanatory variable increases by a unit, then the odds ratio will change e^{α_j} times; in other words, the value of e^{α_j} shows how many times the odds ratio will increase or decrease for the *j*-th variable increasing by a unit (for *j*=1,…,*k*), *ceteris paribus*. A value of $e^{\alpha_j} > 1$ means that the odds of $Y_i = 1$ are growing, while for $e^{\alpha_j} < 1$ the odds are falling. The parameter α_j alone indicates the increase in the logarithm of the odds ratio [Gruszczyński et al., 2009, p. 169].

2.4. Forecasting the explained variable with a logit model

The binomial model, i.e. one describing a dichotomous variable Y_i (the logit model in this case), is mainly used for forecasting probabilities p_i, that is the probabilities of obtaining „a success", as well as forecasting the value of the dependent variable for new units with unknown values of Y_i.

The estimated probability p_i is usually transformed into a dichotomous variable Y_i using the so-called *standard forecasting method* that fixes the threshold value *c* at the level 0,5, i.e. with accepting the rule that $\hat{Y}_i = 1$ when $\hat{p}_i > 0{,}5$ and $\hat{Y}_i = 0$ when $\hat{p}_i \leq 0{,}5$ [Jeziorska-Pąpka, 2007, p. 277], where $\hat{p}_i$ denotes the estimated probabilitity that Y_i=1, and $\hat{Y}_i$ denotes the predicted outcome of the dependent variable Y_i. The above rule flows from the simple assumption that $\hat{Y}_i = 1$ when $\hat{p}_i > \hat{q}_i$, which reduces to the inequality $\hat{p}_i > 0{,}5$.

The standard rule performs well with a balanced sample, i.e. with a sample for which the number of observed successes is equal or "close" to the number of observed failures [Gruszczyński, 2001, p. 80]. However, when we deal with an unbalanced sample, then this approach may lead to poor predictive properties. „Upon fitting a logit model it is then invariably found that the estimated prediction probabilities $\hat{p}_i$ are quite high for Y_i=1, the outcome with the greater share, and very low for the outcome with lesser share" [Cramer, 1999, p. 3].

In the literature the following solutions are proposed to find a threshold value *c* for an unbalanced sample [Gruszczyński, 2001, p. 81; Jeziorska-Pąpka, 2007, p. 277; Dudek and Dybciak, 2006, p. 85]:

a) *Cramer's optimal threshold value*: $\hat{y} = 1$ when $\hat{p}_i > c$ and $\hat{y} = 0$ when $\hat{p}_i \leq c$ where c is the threshold value representing the share of 1s in the sample.
b) *Sample balancing*, by drawing two subsamples of the same size separately for both classes of the dependent variable. This sample composition is called a matched sample [Gruszczyński, 2001, p. 68]. In such a case the optimal cut-off point *c* is the value 0,5.

In our paper, we shall present the concept of using another method of finding the cut-off point c for the logit model's outcomes when the model is estimated from an unbalanced sample. The proposed method uses the properties of the so-called Receiver Operating Characteristic Curves (ROC) . The relevant theory can be found in section 3.

2.5. Goodness-of-fit of logit models

The literature of the subject usually proposes the following measures of model's goodness-of-fit [see Gruszczyński, 2001, pp. 64-66; Maddala, 1992, pp. 332-334]:

a) The Efron's determination coefficient R^2:

$$\text{b) } R^2_{Efrona} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{p}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{n}{n_1 n_0}\sum_{i=1}^{n}(y_i - \hat{p}_i)^2 \tag{13}$$

since

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = n_1 - n\left(\frac{n_1}{n}\right)^2 = \frac{n_1 n_0}{n},$$

where n_1 and n_0 are the numbers of actual outcomes in a sample for which $Y_i=y_i$, where $y_i = 1$ and $y_i = 0$, respectively.

c) The McFadden's determination coefficient R^2 based on the maximum likelihood function:

$$R^2_{McFadden} = 1 - \frac{\ln \hat{L}_{UR}}{\ln \hat{L}_R} \tag{14}$$

where $\ln \hat{L}_R$ is the maximal value of the logarithm of the likelihood function in a model containing only a constant, while $\ln \hat{L}_{UR}$ represents a maximal value of the logarithm of the likelihood function for a unrestricted model.

d) *Count* R^2 – is a measure of prediction accuracy, indicating the share of the correctly predicted outcomes of the model in their total number:

$$R^2 = \frac{n_{00} + n_{11}}{n} \tag{15}$$

where n_{00} and n_{11} are the numbers of observations for which, respectively, $\hat{Y}_i = Y_i = 0$ and $\hat{Y}_i = Y_i = 1$ (see table 1).

Table 1. The crossclassification of predicted and observed values of Y in a sample

Empirical	Forecasted		Total
	Y=1	Y=0	
Y=1	n_{11}	n_{10}	$n_{1.}$
Y=0	n_{01}	n_{00}	$n_{0.}$
Total	$n_{.1}$	$n_{.0}$	n

Source: Gruszczyński et al., 2009, p. 171.

The above measures take values within the interval [0,1].

e) The *ex post* crossclassification table (see table 1) allows constructing another measure of prediction accuracy (used in this paper). The measure is founded on the Yule's coefficient Q [see Domański, 2001, p. 184]. Using the notation from table 1, the measure Q can be defined as follows:

$$Q = \frac{n_{11} \cdot n_{00} - n_{10} \cdot n_{01}}{n_{11} \cdot n_{00} + n_{10} \cdot n_{01}} \tag{16}$$

The coefficient Q takes values in the interval [-1, 1]. The more it approaches 1, the better prediction accuracy of a given model.

3. The theoretical foundation of *ROC* curves

The *ROC* curve is a statistical tool that is used in many important fields of life and science. It was used for the first time during World War II for receiver's signal analysis. In the 1950s, the *ROC* curve was employed in signal detection theory and psychophysics, whereas in the 1979s and 1980s it turned out to be of use in medicine, particularly radiology, cardiology and epidemiology, and then in other fields of science, e.g. psychology, and the earth sciences, as well as financing and insurance [Krzanowski and Hand, 2009, p. 13].

ROC curves are used to assign the objects to two subpopulations according to the value of a continuous random variable, as they allow selecting a threshold value ensuring a possibly high probabilities of correct classifications.

Let us assume that a population of objects G consists of two subpopulations G_0 and G_1, such that $G = G_0 \cup G_1$ and $G_0 \cap G_1 = \emptyset$, where G_1 is the subpopulation of objects for which Y=1, and G_0 is the subpopulation of objects with Y=0. In other words, the variable Y indicates unit's membership to either of the two subpopulations – G_0 or G_1.

Let us assume that a unit has been drawn from a population G with an unknown value of the indicator Y. We can assign the unit to either G_0 or G_1 using the estimate $\hat{p}$ of the probability p=P(Y=1) that has been obtained from a fitted

logit model. In doing this, we are assuming that the unit will be assigned to G_1, when the inequality $\hat{p} > c$ is met, but for $\hat{p} \le c$ it will go to G_0. The point c will be called the decision threshold (or the cut-off point).

Let us consider the estimator $\hat{p}$ of the probability p derived from the adjusted logit model. Assume that H stand for the CDF of the estimator $\hat{p}$, i.e.

$$H(y) = P(\hat{p} \le y) \tag{17}$$

We will denote by H_0, H_1 the conditional CDFs of $\hat{p}$ in the subpopulation G_0 and G_1, respectively:

$$H_0(y) = P(\hat{p} \le y \mid G_0) \tag{18}$$

$$H_1(y) = P(\hat{p} \le y \mid G_1) \tag{19}$$

It is worth noting that for a given decision threshold $y=c$ the values $\overline{H}_1(c) = 1 - H_1(c)$ and $H_0(c)$ represent probabilities that a unit from G_1 or G_0, respectively, is correctly classified according to the classification rule just described.

Using the assumed notation, a *ROC* curve can be defined as follows:

$$ROC = \left\{ \left(\overline{H}_0(y), \overline{H}_1(y) \right), \quad y \in R \right\},$$

where

$$\overline{H}_1(y) = 1 - H_1(y), \ \overline{H}_0(y) = 1 - H_0(y). \tag{20}$$

In other words, the *ROC* curve is a set of points on a plane with the coordinates $(\overline{H}_0(y), \overline{H}_1(y))$ for $y \in R$.

In general, the theoretical ROC curve is usually not known, because the distributions H_0 and H_1 are unknown, so it is replaced by its empirical counterpart, i.e. by ROC_{emp} of the form:

$$ROC_{emp} = \left\{ \left(\hat{\overline{H}}_0(y), \hat{\overline{H}}_1(y) \right), \quad y \in R \right\}, \tag{21}$$

where $\hat{\overline{H}}_0(y)$ and $\hat{\overline{H}}_1(y)$ are the estimators of, respectively, $\overline{H}_0(y)$ and $\overline{H}_1(y)$. The estimators can be, for instance, derived as the complements of the empirical cumulative distribution functions from a training sample, i.e. a sample with known population membership indicators Y. Such a sample can be written as the following sequence:

$$(\hat{p}_1, Y_1), (\hat{p}_2, Y_2), \ldots, (\hat{p}_n, Y_n), \tag{22}$$

where Y_i, (i = 1, 2, …, n) are the observed outcomes of the indicator variable Y, while $\hat{p}_i$ are the estimated probabilities p_i that Y_i =1, i = 1, 2, …, n (i.e. probabilities derived from the logit model).

By plotting the ROC_{emp} curve we can find the empirical decision threshold $y=c$, i.e. a point producing as high empirical probabilities $\widehat{H}_0(y)$ and $\widehat{\overline{H}}_1(y)$ as possible, where $\widehat{H}_0(y) = 1 - \widehat{\overline{H}}_0(y)$. This procedure reduces to finding $y=c$ with coordinates $(\widehat{\overline{H}}_0(y), \widehat{\overline{H}}_1(y))$ on the plotted empirical ROC curve located the closest to the left upper corner of the unit square, i.e. the closest to the point with coordinates (0,1). Then $y=c$ is treated as the optimal decision threshold.

3. Examples of application of the presented cut-off rules

3.1. Characteristics of the databases and logit models used in the experiments

For the purposes of this article, four different datasets were used to fit the logit models. Each of the models contained two explanatory variables (statistically significant at the significance level 0.05). All the analysed datasets were unbalanced and were assembled based on observations of actual events. The first three models were fitted by using "moderately-sized" datasets (with n below 200). The statistical data were obtained mainly from own questionnaire surveys. The fourth dataset, containing 594 observations, was gathered from databases published in the Internet.

In the first model, a binary dependent variable „Emigration" represented the attitudes of students in their final years of study to economic emigration. The questionnaire survey[1] asked the students the following question: „If you received identical job offers, one in Poland and one from abroad, which one would you choose?". If the student declared to work in Poland, the answer was arbitrarily assigned 0. Answers indicating respondents' inclination to work abroad were given 1[2]. The explanatory variables used in the logit model were respondent's sex and the expected level of earnings (4 different levels of earnings were included). Table 2 presents results of the statistical inference received for the model. Since the article does not aspire to make a detailed analysis of the conclusions offered by the model, let us only state that males expecting substantial salaries were more frequent to choose emigration.

[1] The dataset that the authors used in their research was compiled on the basis of a questionnaire survey that was conducted among the students of the Faculty of Sociology and Economics, University of Łódź, and of the Medical University in Łódź in 2006 for the purposes of Joanna Śmigielska's master thesis: „Dobór i rekrutacja pracowników". The authors were authorised by the thesis author to use the database in this article.

[2] Observations where the marked answer was „I don't know" were omitted from the investigation.

Table 2. Results of the statistical inference for the logit model with the dependent variable „Emigration"

N=95	Model: Logistic regression (logit) (Emigration) Dependent variable: Emigration Loss: Maximum likelihood estimator, MSE max. 1 Final loss: 46,913010240 Chi2(2)=17,683 p=,00014 Constant	Earnings	Sex
Estimate	-5,78256	1,1367	1,33638
Standard error	1,46061	0,4538	0,51454
t(92)	-3,95902	2,5046	2,59725
Significance level p	0,00015	0,0140	0,01094
-95%CL	-8,68345	0,2353	0,31447
+95%CL	-2,88167	2,0380	2,35830
Wald's chi-square	15,67381	6,2729	6,74570
Significance level p	0,00008	0,0123	0,00940
Odds ratio	0,00308	3,1164	3,80525
-95%CL	0,00017	1,2653	1,36953
+95%CL	0,05604	7,6754	10,57295

Source: calculated by the authors with the Statistica software.

In order to explain some selected results contained in the table 2, we refer here to the odds ratio corresponding to the binary, explanatory variable "Sex". It follows from the last column and 9th row of the table 2 that the odds of having emigration inclination is 3,80525 times greater for males (Sex=1) than for females (Sex=0).

In the second model, the binary dependent variable „Kids" was the students' attitude to having multi-children families in the future. As in the first survey, the respondents were students in the final years, studying at the tertiary education institutions in Łódź. For a respondent wishing to have at least three children the observation was assigned 1, otherwise it was 0. The explanatory variables in this model were respondent's sex and the place of permanent residence. The statistical inference results for the model are presented in table 3. In this case, let us only state that males coming from small towns or countryside declared their wish to have a multi-children family the most frequently[1].

[1] More information on the same survey and the conclusions it provided can be found in the article: A. Majdzińska, W. Śmigielski, Wpływ religijności na decyzje dotyczące planowania życia rodzinnego studentów Uniwersytetu Łódzkiego, [in:] (ed.) J. T. Kowaleski, A. Rossa, Przyszłość demograficzna Polski, Folia Oeconomica 231, Wydawnictwo Uniwersytetu Łódzkiego, Łódź 2009 or on the website:
http://www.demografia.uni.lodz.pl/pubonline/Folia_Majdzinska_Smigielski.pdf.

Table 3. Results of the statistical inference for the logit model with the dependent variable „Kids"

	Model: Logistic regression (logit) (Kids) Dependent variable: Kids. Loss: Maximum likelihood estimator. MSE max. 1 Final loss: 61,550783827 Chi2(2)=9,2604 p=,00976		
N=129	Constant	Sex	Origin
Estimate	-4,10963	1,090751	0,64740
Standard error	1,02160	0,472151	0,27923
t(126)	-4,02275	2,310176	2,31850
Significance level p	0,00010	0,022505	0,02203
-95%CL	-6,13134	0,156379	0,09481
+95%CL	-2,08792	2,025124	1,19999
Walds chi-square	16,18249	5,336914	5,37544
Significance level p	0,00006	0,020885	0,02043
Odds ratio	0,01641	2,976510	1,91056
-95%CL	0,00217	1,169269	1,09945
+95%CL	0,12395	7,577050	3,32007

Source: calculated by the authors with the Statistica software.

In the third model, the dependent variable „Taxes" illustrated students' declarations as to their compliance with tax regulations after they become economically active. The data were collected in the course of a questionnaire survey[1]. If respondents declared that they would not pay taxes in the future according to the laws in force, the value assigned to the variable "Taxes" was 1, otherwise it was given 0. The explanatory variables that were statistically significant in this case were binary variables called „Ethics" and „Law". The former took the value 1, when the respondent believed that tax evasion was immoral, otherwise the value was 0. The variable „Law" was assigned 1 when the respondent believed that the possibility of being penalised by the tax administration if irregularities were found was the reason for paying taxes. For other answers, the variable took the value 0. The inference results for the model are presented in table 4.

[1] The database that we used in the investigation was compiled during the questionnaire survey conducted in 2006 among the students of the Faculty of Economics and and Sociology, University of Łódź, and of the Medical University in Łódź for the purposes of W. Śmigielski's master's thesis: „Unikanie opodatkowania jako zagadnienie etyczno-moralne w nauczaniu społecznym Kościoła Katolickiego".

Table 4. Results of the statistical inference for the logit model with the dependent variable „Taxes"

	Model: Logistic regression (logit) (taxes) Dependent variable: Taxes. Loss: Maximum likelihood estimator. MSE max. 1 Final loss: 33,443165686 Chi2(2)=25,873 p=,00000		
N=176	Constant	Ethics	Law
Estimate	-2,46129	-3,36931	1,69001
Standard error	0,60362	1,06492	0,71031
t(173)	-4,07757	-3,16390	2,37925
Significance level p	0,00007	0,00184	0,01844
-95%CL	-3,65270	-5,47122	0,28802
+95%CL	-1,26989	-1,26740	3,09200
Walds chi-square	16,62655	10,01028	5,66083
Significance level p	0,00005	0,00156	0,01735
Odds ratio	0,08532	0,03441	5,41951
-95%CL	0,02592	0,00421	1,33378
+95%CL	0,28086	0,28156	22,02098

Source: calculated by the authors with the Statistica software.

As far as the parameter estimates corresponding to variables „Ethics" and „Law" are analysed (table 4), it is worth noting that they have different signs. Almost all persons believing that tax evasion was immoral declared themselves as honest tax payers. On the other hand, the belief that the fear of the tax administration is what makes people pay taxes was frequently coupled with respondents' unwillingness to declare that they will pay their taxes in future as required by the law. This may suggest that the respondents' opinion about the effectiveness of the Polish tax inspectors is rather moderate.

In the fourth model, the binary dependent variable „Juventus" was the score of a football game played by Juventus Torino in the Italian *Serie A*. If the team did not lose the game, then the dependent variable was given 1, otherwise the assigned value was 0. The explanatory variables in the model were „Goals" and „Home/away". The first of them represented the number of goals the team scored during a game and the second variable was a binary one indicating whether the team played at home (Home/Away=1) or away (Home/Away=0)[1]. Table 5 presents detailed results of the statistical inference concerning this model.

[1] Match scores starting with the season 1991/92 until the last games in 2009 were obtained from the website www.wikipedia.pl for „Serie A".

Table 5. Results of the statistical inference for the logit model with the dependent variable „Juventus"

	Model: Logistic regression (logit) (Juventus) Dependent variable: Juventus. Loss: Maximum likelihood estimator. MSE max. 1 Final loss 184,75434375 Chi2(2)=128,84 p=0,0000		
N=594	Constant	Home/Away	Goals
Estimate	-0,92362	0,566605	-1,56049
Standard error	0,49799	0,274142	0,19004
t(591)	-1,85471	2,066829	-8,21158
Significance level p	0,06414	0,039185	0,00000
-95%CL	-1,90165	0,028194	-1,93372
+95%CL	0,05442	1,105016	-1,18726
Walds chi-square	3,43993	4,271783	67,43002
Significance level p	0,06365	0,038758	0,00000
Odds ratio	0,39708	1,762274	0,21003
-95%CL	0,14932	1,028595	0,14461
+95%CL	1,05593	3,019274	0,30505

Source: calculated by the authors with the Statistica software.

Analysing the parameter estimates (table 5) we intuitively arrive at rather obvious conclusions that Juventus odds of having a favourable result (i.e. not losing a match) increase for a match played at home and with the number of goals the team scores.

3.2. Determination of the optimal cut-off points for logit models and assessment of the models' prediction accuracy

The problem of low prediction accuracy for a dependent variable in an unbalanced sample in the case of the logit model was already brought up by J. Cramer in his article [Cramer J.S, 1999]. It should be noted that the prediction accuracy involving a binary endogenous variable depends not only on the appropriate specification of model (7) and on the correct estimation of its parameters, but also on the methodology one uses to find the cut-off point c, in which the estimates $\hat{p}_i$ of probabilities p_i are transformed into values 0 or 1 of the variable Y_i. The problem was already discussed in section 2.4.

Regarding the question of finding an optimal cut-off point c, Cramer proposed an alternative method to the standard forecasting method. Regarding the eight logit models that the author analysed, the comparison of *count* R^2 for forecasts with a standard cut-off point c at 0,5 and a cut-off point found using the Cramer's approach indicates that the standard method provides a higher *count* R^2 in each of the models. However, the Cramer's method produces more accurate predictions for one of two values of dependent variable Y which is characterized by a smaller share in the training sample. In one model only the result was actually identical. The Cramer's approach will be called hereafter the Cramer's rule.

The conclusion is that the standard method of finding the cut-off point generally provides good results for the case of unbalanced samples as far as the *count* R^2 is concerned; however, the frequency of correct forecasts for one of the dependent variable's values (i.e. 1 or 0) is quite frequently below 0,5, and sometimes it is exactly 0. This especially applies to this value of the dependent variable that has a smaller share in the unbalanced sample. The Cramer's rule for finding the cut-off point almost always improves the frequency of the correct *ex post* forecasts for such a value, however, the frequency of the correct forecasts for the second, predominating value of Y is worse. As a result, the *count* R^2 is usually smaller then in the standard method. Considering the context, however, the *count* R^2 does not seem to offer the complete picture of the model quality. It is therefore recommended to use in the comparisons also other measures of model's goodness-of-fit, such as Efron's R^2, which is equivalent to the ordinary coefficient of determination as far as the binomial model is concerned.

In the next section, we shall examine a proposed method for seeking an optimal cut-off point for unbalanced samples that builds on the properties of the ROC curves. This method will be called the *ROC rule*. The *ex post* prediction results obtained by means of the *ROC* rule and those provided by the *standard method* and the *Cramer's rule* are contrasted in tables 6 and 7.

Table 6 shows the frequencies with which the predicted values of the dependent variable agree with its actual values, i.e. the accuracy of the *ex post* forecasts for the four logit models discussed in section 3.1. Then, table 7 presents the basic prediction accuracy measures for the particular models. Since both the *count* R^2 values and the values of the *Q coefficient* depend on the method used to find the cut-off point, the point was determined separately for each of the three different rules. The best results are printed in bold to facilitate the analysis of the tables 6–7.

Table 6. Correct *ex post* forecasts for the dependent variable in the four logit models as produced by the three methods for determining the cut-off point

Dependent variable (logit model)	Sample characteristics		Percentages of correct predictions of outcomes (for two values *WD* and *WN* of the dependent variable)					
			WD			*WN*		
	n	WN%	standard method	Cramer's rule	ROC rule	standard method	Cramer's rule	ROC rule
Emigration	95	**27%**	**81,16%**	79,71%	79,71%	38,46%	**57,69%**	**57,69%**
Kids	129	**21%**	**96,08%**	50,00%	50,00%	18,52%	**70,37%**	**70,37%**
Taxes	145	**9%**	**100,00%**	67,38%	67,38%	0,00%	**92,31%**	**92,31%**
Juventus	594	**15%**	**94,47%**	75,49%	91,70%	39,77%	**80,68%**	52,10%

n – sample size. D – value of the dependent variable accounting for the larger share in the sample. WN – value of the dependent variable accounting for the smaller share in the sample. WN %– the percentage share of WN's values in a dataset.

Source: calculated by the authors with the Statistica software.

Table 7. Selected measures of prediction accuracy for the four logit models as produced by the three methods for determining the cut-off point

	Sample characteristics			Count R^2			Q coefficient		
Dependent variable (logit model)	n	WN%	Efron's R^2	Standard rule	Cramer's rule	ROC rule	Standard rule	Cramer's rule	ROC rule
Emigration	95	**27%**	0,14	0,69	**0,74**	**0,74**	0,46	**0,68**	**0,68**
Kids	129	**21%**	0,08	**0,80**	0,54	0,54	**0,70**	0,41	0,41
Taxes	145	**9%**	0,21	**0,93**	0,69	0,69	×	**0,92**	**0,92**
Juventus	594	**15%**	**0,25**	0,86	0,76	**0,87**	0,84	0,86	**0,87**

n – sample size. WN – value of the dependent variable accounting for the smaller share in the sample. WN % – the percentage share of WN's values in a sample. × – denotes that the percentage of correct predictions took value 0 and the Q coefficient could not be computed.

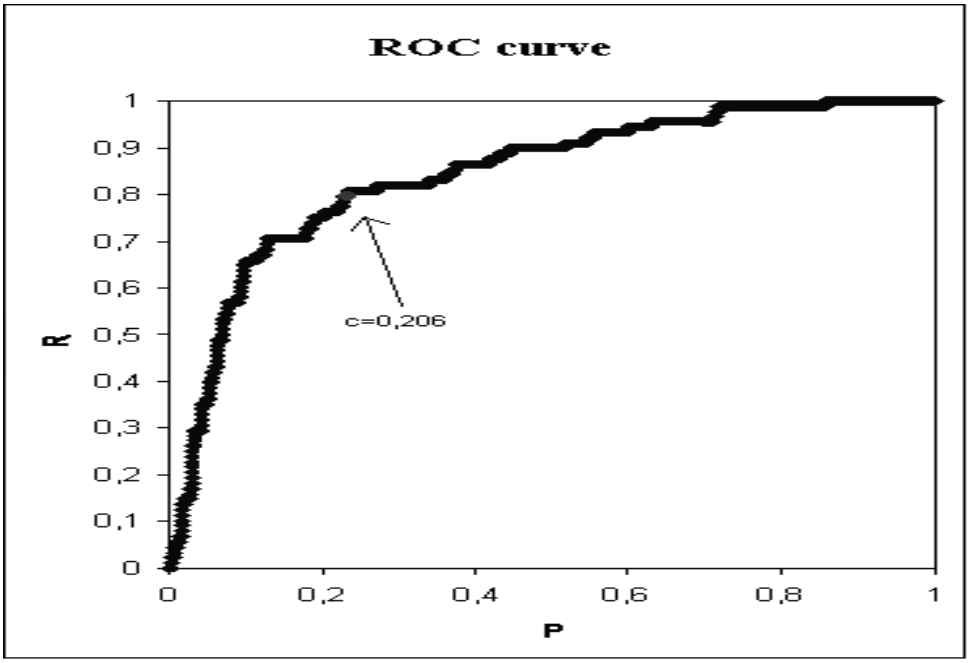
Source: calculated by the authors with the Statistica software.

It is worth noting that the *count* R^2 computed using the standard rule for three of the four models exceeds 0.8 (thus the results are similar as discussed in the Cramer's article), which is a relatively high value. On the other hand, the percentage of consistency between the predicted values of dependent variables and their actual values which represent the smaller share in the respective datasets

is lower than 50% in the case of the standard rule for each of the four models, but for the variable „Taxes" it is 0% (see table 6).

Another notable fact is that the prediction accuracy measures (i.e. *count* R^2 and *Q coefficient*) obtained by employing the *Cramer's rule* and the *ROC rule* is identical for the first three models. In the authors' opinion, this situation arises from the relatively limited sample sizes. As far as the forecasts of the variable „Juventus" are concerned, the *ROC rule* performs the best, as it provides the highest values of *count* R^2 and *Q coefficient*. Figure 2 illustrates the empirical ROC curve for this case. The empirical ROC curves for the other models are included in Appendix 1.

Figure 2. The empirical ROC curve with the optimal cut-off point for the logit model with the dependent variable „Juventus"



Legend[1]*:* $P=\hat{\overline{H}}_0(y)$, $R=\hat{\overline{H}}_1(y)$.

Source: calculated by the authors with the Microsoft Excel worksheet.

Table 8 compares the cut-off point values calculated using the *Cramer's rule* and the *ROC rule* for the four models.

[1] The same symbols apply to the other ROC curves.

Table 8. The cut-off points as produced by the *Cramer's rule* and the *ROC rule* for the four logit models

Dependent variable (model)	Cut-off point		Absolute diference $\|c_1-c_2\|$
	Cramer's rule (point c_1)	ROC rule (point c_2)	
Emigration	0,274	0,262	0,012
Kids	0,209	0,217	0,008
Taxes	0,074	0,079	0,005
Juventus	0,148	0,206	0,058

Source: calculated by the authors with the Microsoft Excel worksheet.

Analysing the data in table 8 we find that for training samples containing relatively small numbers of observations the cut-off point provided by the *ROC rule* is almost fully consistent with that obtained using the *Cramer's rule* (hence, the forecasts provided by both rules in the first three models are the same). However, in the model with the dependent variable „Juventus" the points found using both rules are noticeably different, which affects the *ex post* forecast results.

As shown by the Efron's R^2 values[1] the best fit of the model is in the cases of the variable „Juventus". Hence, we decided to assess the quality of the model again, this using the data from the so-called test sample that was not involved in model estimation. The test sample was compiled using the scores of the matches played by *Juventus Torino in Serie A* in the four successive seasons: from 1987/88 to 1990/1991 (n=132 matches in total). Table 9 presents results for the logit model with parameter estimates given in table 5, but forecast results derived with respect to the test sample using the three discussed rules.

Table 9. Forecast results for the dependent variable „Juventus" received for the test sample

Rule	Percentage of correct forecasts in the test sample (for two values WD_t and WN_t of the dependent variable)		Values of prediction accuracy measures in the test sample	
	WD_t	WN_t	Count R^2	**Q**
Standard	**84,16%**	70,97%	**0,81**	**0,86**
Cramer's	67,74%	**75,25%**	0,73	0,73
ROC	**84,16%**	67,74%	0,80	0,84

WD_t – value of the dependent variable accounting for the larger share in the test sample.
WN_t – value of the dependent variable accounting for the smaller share in the test sample.

Source: calculated by the authors with the Microsoft Excel.

[1] As we can read in statistical articels Effron's R^2 value is Fast always not very high. See more: Morrison D. G., *Upper Bounds for Correlation Between Binary Outcomes and Probabilistic Predictions*, JASA, vol. 67, no. 337/.

It is worth noting that the Q values in table 9 are close to 1, however they are higher than for the training sample data (table 7). Compared with the first model, though, the percentage of the correct forecasts for those values of the dependent variable, which are represented by the smaller portion of the test observations, is higher (this concerns values of the dependent variable denoted by WN_t in table 9). The best, i.e. the greatest, value of Q is obtained with the *standard rule*, although it is only slightly better than the *ROC rule*. The poorest results were obtained for the cut-off point determined using the *Cramer's rule.*

4. Summation and final conclusions

When the logit models are based on unbalanced samples then it becomes necessary to use methods allowing the determination of a cut-off point that are alternative to the *standard rule*. The *Cramer's rule* usually increases the pecentage of the correct forecasts of the dependent variable, but it frequently involves lower values of *count* R^2 and/or of the *Q coefficient*. The *ROC rule* application to computing the optimal cut-off points that the authors proposed has not yielded better results for the models with relatively small sample sizes, but for the model based on a large number of observations (almost 600 in this study) the procedure improved the frequency of the correct forecasts concerning those values of the dependent variable which are represented by a smaller part of the observations. It allowed also obtaining higher values for both *count* R^2 and *Q coefficient* (it must be borne in mind, though, that *count* R^2 provides rather general information about the prediction accuracy).

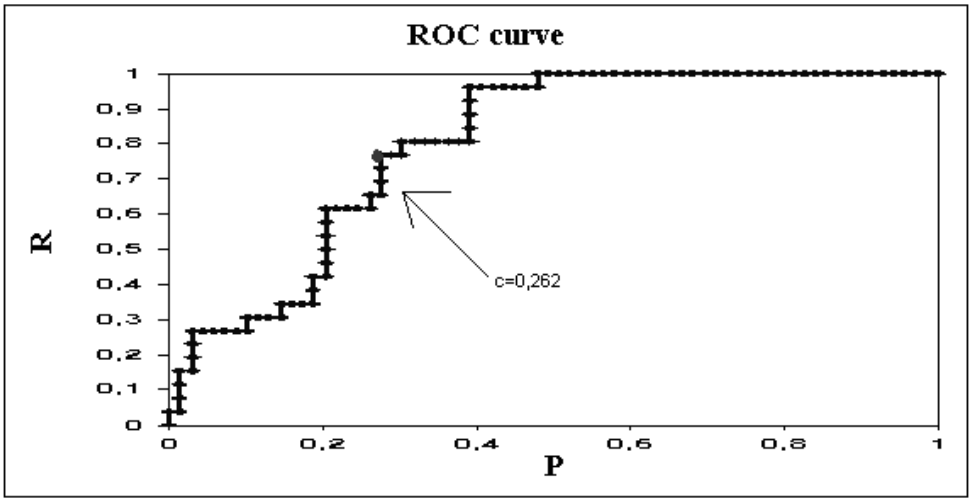
The authors are aware that further research is needed to confirm the conclusions offered by the study. This article aims at demonstrating that using the ROC curve is justified when the binary dependent variables are forecasted using, for instance, the logit models. The presented examples show that the method can produce results that are better or comparable with those offered by methods that are more common in the literature and applied more often.

Acknowledgements

We acknowledge an anonymous reviewer for valuable comments and suggestions which improved the manuscript.

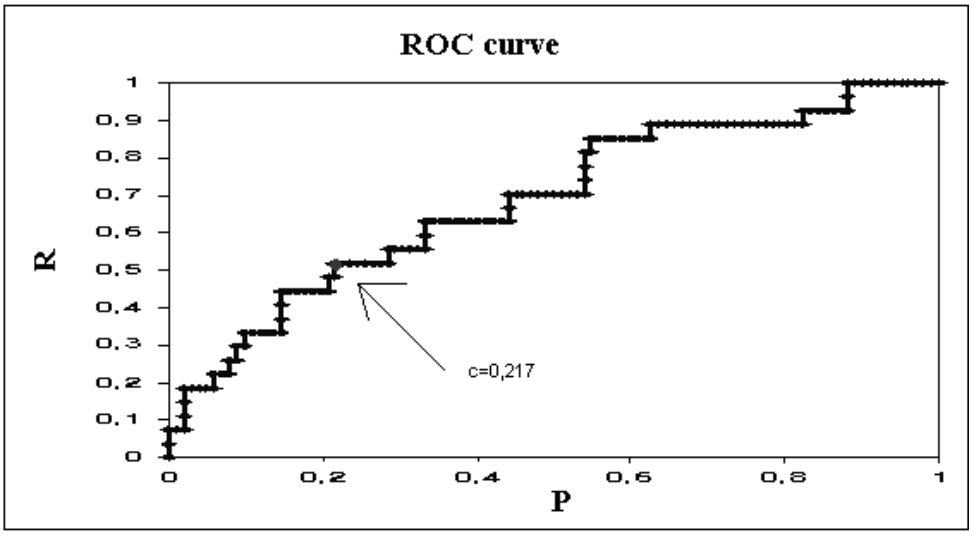
Appendix

Figure A.1. The empirical ROC curve for the model with the dependent variable „Emigration"



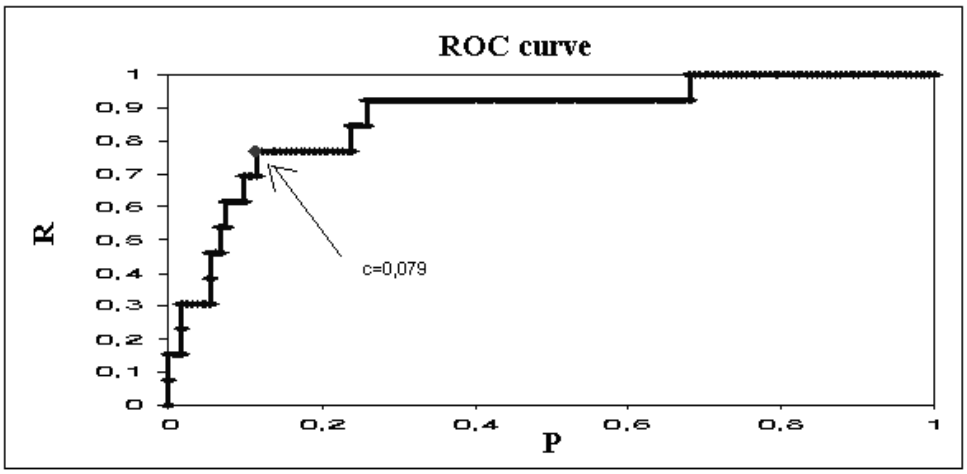
Source: calculated by the authors with the Microsoft Excel worksheet.

Chart 3. The empirical ROC curve for the model with the dependent variable „Kids"



Source: developed by the authors using the Microsoft Excel worksheet.

Chart 4. The empirical ROC curve for the model with the dependent variable „Taxes"



Source: calculated by the authors with the Microsoft Excel worksheet.

REFERENCES

CRAMER J. S. (1999). Predictive performance of the binary logit model in unbalanced samples, [in]: Journal of the Royal Statistical Society. Series D (The Statistician), Vol. 48, No. 1, pp. 85-94, Blackwell Publishing, Oxford.

DOMAŃSKI CZ. (ed) (2001) Metody statystyczne. Teoria i zadania, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

DUDEK H., DYBCIAK M. (2006) Zastosowanie modelu logitowego do analizy wyników egzaminu [in]: Zeszyty Naukowe Szkoły Głównej Gospodarstwa Wiejskiego w Warszawie, Ekonomika i Organizacja Gospodarki Żywnościowej, nr 60, Wydawnictwo SGGW, Warsaw.

GRUSZCZYŃSKI M. (2001) Modele i prognozy zmiennych jakościowych w finansach i bankowości, Monografie i opracowania, no. 490, Szkoła Główka Handlowa, Warsaw.

GRUSZCZYŃSKI M., KUSZEWSKI T. AND PODGÓRKSA M. (2009) Ekonometria i badania operacyjne. Podręcznik dla studiów licencjackich, PWN, Warsaw.

JEZIORSKA-PĄPKA M. (2007) Zastosowanie modeli dwumianowych do opisu asymetrii informacji na rynku ubezpieczeń na przykładzie polis komunikacyjnych OC [in]: Dynamiczne Modele Ekonometryczne, Wydawnictwo Uniwersytetu Mikołaja Kopernika, Toruń.

KRZANOWSKI W., HAND D. (2009) ROC curves for continuous data, Monographs on statistics and applied probability, 111, CRC Press, New York.

MADDALA G. S. (1992) Introduction to Econometrics – 2nd ed., Macmillan Publishing Company, New York.

MAJDZIŃSKA A., ŚMIGIELSKI W. (2009) Wpływ religijności na decyzje dotyczące planowania życia rodzinnego studentów Uniwersytetu Łódzkiego, [in:] (ed.) J. T. Kowaleski, A. Rossa, Przyszłość demograficzna Polski, Folia Oeconomica 231, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

MORRISON D. G. (1972) *Upper Bounds for Correlation Between Binary Outcomes and Probabilistic Predictions*, JASA, vol. 67, no. 337.

PRUSKA K. (2001) Modele probitowe i logitowe w programach nauczania studiów ekonomicznych [in:] Metody analizy cech jakościowych w procesie podejmowania decyzji (conference proceedings), Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

ROSSA A. (2007) Asymptotic tests for receiver operating characteristic curves [in]: Statistics in Transition – new series, vol. 8, no. 3, GUS, Warsaw.

STANISZ A. (2000) Przystępny kurs statystyki z wykorzystaniem programu STATISTICA PL na przykładach z medycyny, vol. II, Kraków.

WELFE A. (2003) Ekonometria. Metody i zastosowanie, PWE, Warsaw.

WELFE A., BRZESZCZYŃSKI J. AND MAJSTEREK M. (2002) Angielsko-polski, polsko-angielski słownik terminów metod ilościowych, Polskie Wydawnictwo Ekonomoiczne, Warsaw.