PRZEGLĄD STATYSTYCZNY
R. LXIV – ZESZYT 3 – 2017

JAN PURCZYŃSKI[1], KAMILA BEDNARZ-OKRZYŃSKA[2]

THE RAYBIT MODEL AND THE ASSESSMENT OF ITS QUALITY IN COMPARISON WITH THE LOGIT AND PROBIT MODELS

1. INTRODUCTION

A prevailing amount of methods of econometric model analysis refers to the situation when variables (both dependent and explanatory) are continuous variables. This is the case of a quantitative model – a quantitative dependent variable. If the variable can take a finite number of values, it is referred to as a discrete or a qualitative variable. Gruszczyński (2012) draws attention to an increasing importance of qualitative models, as they constitute a basic tool for describing microeconometric models used in empirical corporate finance. In the case when a variable takes only two values it is called a dichotomous variable (also binomial or binary variable).

The simplest method of solving an equation with a binary dependent variable is a linear probability model, the solution of which has one vital drawback, namely, a possibility of obtaining the probability which falls outside the interval [0;1] (Maddala, 1992). In order to get rid of this drawback, it is assumed that the probability corresponds to the cumulative distribution function of a random variable. In the case of a logistic distribution, a logit model is obtained, and in the case of a normal distribution – a probit model (Maddala, 1992).

In the literature other types of transformations can be found. Nerlove (1973) provides the following formulas:

$$F(x) = \frac{1}{2}\left(1 + \sin x\right), \quad -\frac{\pi}{2} \leq x \leq \frac{\pi}{2}, \tag{1}$$

$$F(x) = \frac{1}{2} + \frac{1}{\pi}\arctan x, \quad -\infty < x < \infty, \tag{2}$$

$$F(x) = \tanh x, \quad -\infty < x < \infty. \tag{3}$$

[1] University of Szczecin, Faculty of Management and Economics of Services, Department of Quantitative Method, 8 Cukrowa St., 71-004 Szczecin, Poland, corresponding author – e-mail: jan.purczynski@wzieu.pl.

[2] University of Szczecin, Faculty of Management and Economics of Services, Department of Quantitative Method, 8 Cukrowa St, 71-004 Szczecin, Poland.

McFadden (1984) lists the following transformations: the cumulative distribution function of the Student's t-distribution, the cumulative distribution function of the Cauchy distribution and the arctan model (equation (2)). Finney (1973) lists four transformations: the arctan model, the rational function, the $\sin^2 x$ and parabolic function.

In this paper the author proposes his own model, in which the probability is expressed by a Rayleigh cumulative distribution function, hence the name of the model – raybit. The Rayleigh distribution is a special case of the Weibull distribution, the cumulative distribution function of which is given by (Rine, 2009):

$$F(x) = 1 - \exp\left[-\left(\frac{x-a}{b} \right)^c \right]. \tag{4}$$

By assuming in equation (4) $a = 0$, $b = 1$ and $c = 2$, the Rayleigh cumulative distribution function is obtained:

$$F(x) = 1 - \exp\left(- x^2 \right). \tag{5}$$

While conducting computer simulations described in section 5 of the paper, it was observed that the values of parameters a nad b (equation (4)) do not affect the results of the proposed method. In order to obtain the simplest form of the Rayleigh cumulative distribution function (equation (5)), $a = 0$ and $b = 1$ were assumed.

The continuous random variable with a Weibull distribution (Rayleigh) has been widely applied in modeling physical and economic phenomena (Polakow, Dunne, 1999; Celik, 2003).

The random variable with a Weibull distribution is also applied in binary variable analysis, yet the cumulative distribution function is given by a relation other than equation (4) (Chou, 1983):

$$F(x) = \exp\left[- \exp\left(- x \right) \right]. \tag{6}$$

This misunderstanding is explained by Train (2009), namely, the distribution (6) is also called Gumbel and type I extreme value, and quite often is mistakenly referred to as the Weibull distribution. The Gumbel distribution is often used in modeling extreme values (Koutsoyiannis, 2003).

Hence it can be concluded that the Rayleigh distribution (equation (5)) has not been used in modeling a discrete variable yet.

2. PROBABILITY MODELS FOR A BINARY VARIABLE

It is assumed that a variable Y can take two values: one or zero, corresponding to the fact of making or not making a decision – an occurrence of an event A.

The subject of the analysis are the models of a binary variable for grouped data.

If among n_i of decision-makers, y_i of them made a sensible decision, then a quotient

$$p_i = \frac{y_i}{n_i}, \ (i = 1,2,...,I) \tag{7}$$

represents an empirical frequency of making a decision in an i-th group of decision-makers.

The easiest model is a linear model of probability (Judge et al., 1980):

$$\boldsymbol{p} = \boldsymbol{X\alpha} + \boldsymbol{\varepsilon}, \tag{8}$$

where:
$\boldsymbol{p}$ – I-dimensional vector of empirical probabilities,
$\boldsymbol{X}$ – $[I \times (k+1)]$ dimensional matrix including k number of explanatory variables,
$\boldsymbol{\alpha}$ – $(k+1)$ vector of parameters,
$\boldsymbol{\varepsilon}$ – I-dimensional vector of random elements.
Based on equation (8), the following can be observed

$$p_i = P_i + \varepsilon_i, \tag{9}$$

where:
p_i – empirical probability of an occurrence of an event A for an i-th value of a vector of explanatory variables,
P_i – probability of an occurrence of an event A for an i-th value of a vector of explanatory variables,
ε_i – a disturbance: $E(\varepsilon_i) = 0$ and $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$.

Since a variable y_i (equation (7)) has a binomial distribution, the variance of a disturbance is given by relation (Judge et al., 1980)

$$V(\varepsilon_i) = \frac{P_i(1 - P_i)}{n_i}, \tag{10}$$

which means that the disturbances appearing in equation (9) are heteroskedastic.

Due to the drawback mentioned in the Introduction of this paper (p. 1), the linear probability model will not be further discussed.

It is assumed that the probability P_i, with which the decision in question is made in an i-th group of decision-makers, is a function F of a variable $\boldsymbol{x_i}^T\boldsymbol{\alpha}$

$$P_i = F(\boldsymbol{x_i}^T\boldsymbol{\alpha}), \tag{11}$$

where F is a cumulative distribution function, $\boldsymbol{x_i}^T$ is an i-th row of an explanatory variable matrix.

The most commonly applied cumulative distribution functions are as follows:

– a logit model, hereafter referred to as LOG

$$P_i = L(\boldsymbol{x}_i^T \boldsymbol{\alpha}) = \left[1 + e^{-\boldsymbol{x}_i^T \boldsymbol{\alpha}} \right]^{-1}, \tag{12}$$

where L denotes the cumulative distribution function of a logistic distribution

– a probit model, hereafter referred to as PRO

$$P_i = \Phi(\boldsymbol{x}_i^T \boldsymbol{\alpha}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\boldsymbol{x}_i^T \boldsymbol{\alpha}} e^{-\frac{t^2}{2}} dt, \tag{13}$$

where Φ denotes the cumulative distribution function of a standardized normal distribution.

Depending on the model, a vector v is called:

– an observed logits $$v_i = \ln\left(\frac{p_i}{1 - p_i} \right), \quad \text{LOG} \tag{14}$$

– an observed probits $$v_i = \Phi^{-1}(p_i), \quad \text{PRO} \tag{15}$$

where $\Phi^{-1}(\ )$ – the inverse function to the cumulative distribution function of a standardized normal distribution.

The following relations can be observed (Amemiya, 1981; Judge et al., 1980):

$$v_i = \boldsymbol{x}_i^T \boldsymbol{\alpha} + u_i \ , \quad E(u_i) = 0,$$

– for the logit model $$V(u_i) = \frac{1}{n_i P_i (1 - P_i)}, \tag{16}$$

– for the probit model $$V(u_i) = \frac{P_i (1 - P_i)}{n_i \left\{ \varphi \left[\Phi^{-1}(P_i) \right] \right\}^2}, \tag{17}$$

where φ – a standard normal density.

In this paper, the Rayleigh cumulative distribution function, given by equation (5), is considered as a function F in equation (11),

$$P_i = R(\boldsymbol{x}_i^T \boldsymbol{\alpha}) = 1 - \exp\left[-\left(\boldsymbol{x}_i^T \boldsymbol{\alpha} \right)^2 \right]. \tag{18}$$

A vector $\boldsymbol{v}$ of the observed raybit is given by formula:

$$v_i = \sqrt{-ln(1-p_i)}. \tag{19}$$

From equations (18) and (19) it follows that

$$R^{-1}(P_i) = \boldsymbol{x}_i^T \boldsymbol{\alpha} = \sqrt{-\ln(1-P_i)}. \tag{20}$$

Starting from equation

$$\sqrt{-ln(1-p_i)} = \sqrt{-\ln(1-P_i-\varepsilon_i)},$$

and adopting approximate formulas, applicable for small values δ ($\delta \approx 0$):

$$\ln(1 \pm \delta) \approx \pm \delta,$$

$$\sqrt{1+\delta} \approx 1 + \frac{1}{2}\delta,$$

the following is derived

$$R^{-1}(p_i) = \boldsymbol{x}_i^T \boldsymbol{\alpha} + \frac{\varepsilon_i}{2 \cdot (1-P_i) \cdot \sqrt{-\ln(1-P_i)}} = \boldsymbol{x}_i^T \boldsymbol{\alpha} + \eta_i. \tag{21}$$

From equations (10) and (21), the following is obtained:

$$V(\eta_i) = \frac{P_i}{4n_i(1-P_i)\ln\dfrac{1}{1-P_i}}. \tag{22}$$

Which means that the random variable η_i is heteroskedastic.

In the analysis of each model the following three steps can be singled out (Judge et al., 1980; Jajuga, 1989):

A. The first step

Estimation of a vector $\boldsymbol{\alpha_0}$ of parameters $\boldsymbol{\alpha}$

$$\boldsymbol{\alpha_0} = \left(\boldsymbol{X}^T \boldsymbol{W}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W}^{-1} \boldsymbol{v}, \tag{23}$$

where: $\boldsymbol{W}$ is a diagonal covariance matrix (of a size $I \times I$), where the elements on the main diagonal equal:

$$w_i = \left[n_i p_i (1-p_i)\right]^{-1}, \qquad \text{LOG} \tag{24}$$

$$w_i = \frac{p_i(1-p_i)}{n_i\left\{\varphi\left[\Phi^{-1}(p_i)\right]\right\}^2}, \qquad \text{PRO} \quad (25)$$

$$w_i = \frac{p_i}{4n_i(1-p_i)\ln\dfrac{1}{1-p_i}}. \qquad \text{RAY} \quad (26)$$

The estimation of theoretical probability:

$$p_{0i} = L\left(\boldsymbol{x}_i^T\boldsymbol{\alpha_0}\right), \qquad \text{LOG} \quad (27)$$

$$p_{0i} = \Phi(\boldsymbol{x}_i^T\boldsymbol{\alpha_0}), \qquad \text{PRO} \quad (28)$$

$$p_{0i} = R(\boldsymbol{x}_i^T\boldsymbol{\alpha_0}). \qquad \text{RAY} \quad (29)$$

B. The second step

By applying the ordinary least squares (OLS), the following is obtained:

$$\boldsymbol{\alpha_1} = \left(\boldsymbol{X^T X}\right)^{-1}\boldsymbol{X^T v}, \qquad (30)$$

where: $\boldsymbol{v}$ is defined by formulas (14), (15), (19).

The estimation of theoretical probability

$$p_{1i} = L\left(\boldsymbol{x}_i^T\boldsymbol{\alpha_1}\right), \qquad \text{LOG} \quad (31)$$

$$p_{1i} = \Phi(\boldsymbol{x}_i^T\boldsymbol{\alpha_1}), \qquad \text{PRO} \quad (32)$$

$$p_{1i} = R(\boldsymbol{x}_i^T\boldsymbol{\alpha_1}). \qquad \text{RAY} \quad (33)$$

C. The third step

Estimation of a vector $\boldsymbol{\alpha_2}$ of parameters $\boldsymbol{\alpha}$

$$\boldsymbol{\alpha_2} = \left(\boldsymbol{X^T W_1}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X^T W_1}^{-1}\boldsymbol{v}, \qquad (34)$$

where: $\boldsymbol{v}$ is defined by formulas (14), (15), (19).

$\boldsymbol{W}_1$ is a diagonal covariance matrix, where the elements on the main diagonal equal:

$$w_{1i} = \left[n_i p_{1i}(1-p_{1i})\right]^{-1}, \qquad \text{LOG} \quad (35)$$

$$w_{1i} = \frac{p_{1i}(1-p_{1i})}{n_i\left\{\varphi\left[\Phi^{-1}(p_{1i})\right]\right\}^2}, \qquad \text{PRO} \quad (36)$$

$$w_{1i} = \frac{p_{1i}}{4n_i(1-p_{1i})\ln\frac{1}{1-p_{1i}}}, \qquad \text{RAY} \quad (37)$$

where p_{1i} is given by equations (31), (32) and (33).

The estimation of theoretical probability:

$$p_{2i} = L\left(\boldsymbol{x}_i^T \boldsymbol{\alpha}_2\right), \qquad \text{LOG} \quad (38)$$

$$p_{2i} = \Phi(\boldsymbol{x}_i^T \boldsymbol{\alpha}_2), \qquad \text{PRO} \quad (39)$$

$$p_{2i} = R(\boldsymbol{x}_i^T \boldsymbol{\alpha}_2). \qquad \text{RAY} \quad (40)$$

In the literature (Judge et al., 1980; Jajuga, 1989) there are two alternative methods described: the probability p_0 and the probability p_2 (in which case the probability p_1 is used to determine p_2). In this paper the probability p_1 is taken into account in the same way as p_0 and p_2.

The forms of the likelihood function for the logit and probit models can be found in the paper by Chow (1983). In the case of the raybit model, the likelihood function is given by:

$$L(\boldsymbol{\alpha}) = \prod_{i=1}^{I}\left[1 - e^{-\left(\boldsymbol{x}_i^T \boldsymbol{\alpha}\right)^2}\right]^{n_i'}\left[e^{-\left(\boldsymbol{x}_i^T \boldsymbol{\alpha}\right)^2}\right]^{n_i - n_i'},$$

where n_i' is the number of decision-makers for whom a variable $y_i = 1$ (equation (7)).

The log-likelihood function of the model is given by:

$$lnL(\alpha_0, \alpha_1) = \sum_{i=1}^{I} n_i' \cdot ln\left[1 - e^{-(\alpha_0 + \alpha_1 x_{1i} + \ldots, \alpha_k x_{ki})^2}\right] - \sum_{i=1}^{I}(n_i - n_i')(\alpha_0 + \alpha_1 x_{1i} + \ldots \alpha_k x_{ki})^2. \quad (41)$$

The necessary condition for extremum leads to the set of equations:

$$\sum_{i=1}^{I}\left\{\frac{n_i'}{1 - \exp\left[-(\alpha_0 + \alpha_1 x_{1i} + \ldots \alpha_k x_{ki})^2\right]} - n_i\right\}(\alpha_0 + \alpha_1 x_{1i} + \ldots \alpha_k x_{ki}) = 0, \quad (42a)$$

$$\sum_{i=1}^{I}\left\{\frac{n_i'}{1 - \exp\left[-(\alpha_0 + \alpha_1 x_{1i} + \ldots \alpha_k x_{ki})^2\right]} - n_i\right\}(\alpha_0 + \alpha_1 x_{1i} + \ldots \alpha_k x_{ki}) \cdot x_{1i} = 0, \quad (42b)$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots ,$$

$$\sum_{i=1}^{I}\left\{\frac{n_i'}{1 - \exp\left[-(\alpha_0 + \alpha_1 x_{1i} + \ldots \alpha_k x_{ki})^2\right]} - n_i\right\}(\alpha_0 + \alpha_1 x_{1i} + \ldots \alpha_k x_{ki}) \cdot x_{ki} = 0. \quad (42c)$$

3. ESTIMATING THE ERROR OF THE MODEL

The most popular measure of goodness of fit of a model is the mean square error (MSE):

$$MSE = \frac{1}{I}\sum_{i=1}^{I}(p_i - \hat{p}_i)^2, \tag{43}$$

where:
p_i – empirical probability (equation (7)),
$\hat{p}_i$ – the estimation of theoretical probability.

As $\hat{p}_i$ the results of the following four methods (p_{0i}, p_{1i}, p_{2i}, p_{MLi}) are taken.

Guzik et al. (2005) recommends equation (43) as a criterion of goodness of fit of a theoretical probability model.

Another measure is the mean absolute error (MAE):

$$MAE = \frac{1}{I}\sum_{i=1}^{I}|p_i - \hat{p}_i|. \tag{44}$$

Due to the heteroskedasticity of the disturbance, many authors (cf. Amemiya, 1981; Jajuga, 1989; Maddala, 2006) propose a criterion called the Weighted Mean Squared Error (WMSE):

$$WMSE = \sum_{i=1}^{I}\frac{n_i(p_i - \hat{p}_i)^2}{p_i(1-p_i)}. \tag{45}$$

The main problem lies in the fact that the variance of MSE (equation (43)) and MAE (equation (44)) depends heavily on the value of the empirical probability. Therefore, a recommended measure of goodness of fit is the weighted mean squared error (equation (45)). This issue was discussed in the paper by Purczyński et al. (2015), where computer simulations were carried out using a random number generator with a binominal distribution. As a result of these studies, yet another measure of goodness of fit was proposed, namely the Weighted Mean Absolute Error (WMAE):

$$WMAE = \sum_{i=1}^{I}\frac{n_i|p_i - \hat{p}_i|}{\sqrt{p_i(1-p_i)}}. \tag{46}$$

Adopting as a criterion a constant value of a variance for a changing empirical probability, it was shown, in the aforementioned paper, that the least useful measure of goodness of fit of the model is MSE (equation (43)), a slightly better measure is MAE (equation (44)), still better is WMSE (equation (45)), however the best and the mostly recommended one is the weighted mean absolute error (equation (46)).

4. COMPUTATIONAL EXAMPLES

A computational example was conducted based on the data taken from Household Budget Survey in 2012, CSO Warsaw 2013, which refers to the likelihood of possessing the PC by a household. The data presented in table 1 refer to the year 2012. Column 5 includes the number of households n_i' equipped with the PC.

Table 1.

Households equipped with PCs

Lp.	Number of residents in thousand	Surveyed residents in thousand x_{1i}	Available income per person x_{2i}	Households surveyed n_i	Empirical probability p_i	Number of households possessing PC n_i'
	1	2	3	3	4	5
1	less than 20	10	1199.58	4296	0.652	2801
2	20–99	60	1272.82	6447	0.676	4358
3	100–199	150	1320.44	2719	0.707	1922
4	200–499	350	1497.20	3455	0.722	2495
5	500 and more	870	2011.66	4768	0.769	3667
6	rural	0.4	1027.63	15742	0.642	10106

Source: Household Budget Survey in 2012, GUS, Warsaw.

Column 1 of table 1 contains the number of residents of Polish towns in which the people, included in the survey and given in column 3, live. Column 2 (rows 1–4) contains the values which correspond to the center of the interval. In the case of towns of the population 500,000 and more, the mean was calculated for five Polish towns fulfilling this condition. Row 6 represents rural residents. Starting from the number of rural residents and the number of villages, the average number of rural residents was estimated at 360 persons, which was rounded off to 0.4 thousand. The household possessing the PC was chosen as the first model, where an explanatory variable x_{1i} was the number of residents (column 2). Table 2 contains the results of calculations in the form of errors of the following models: logit, probit, raybit.

As far as labeling is concerned, the model errors MAE0, MAE1, MAE2 demonstrate the results of calculations obtained using equation (44) for estimating the probability *p0* – equations (27), (28) and (29). However MAEML represents the results of the Maximum Likelihood Method for the error given by equation (44).

Taking into account the data included in table 2, the values of errors for particular methods obtained for a given equation were compared. For instance, for the data included in column 1 and rows 1, 2, 3 it can be noticed that in the case of equation

(44) (MAE) and the results obtained for the estimation of the probability p_0, the raybit model yields the smallest error. By conducting further comparisons, it was observed that the raybit method yielded the smallest errors in 13 cases. In the remaining three cases, the logit method yielded the smallest errors.

Table 2.

Errors of the models: logit, probit and raybit for explanatory variable x_{1i}

		MAE0	MAE1	MAE2	MAEML	MSE0	MSE1	MSE2	MSEML
		1	2	3	4	5	6	7	8
1	Logit	0.01347	0.012963	0.01346	0.01347	0.0002569	0.0002120	0.0002532	0.0002550
2	Probit	0.01371	0.01320	0.01370	0.01371	0.0002640	0.0002178	0.0002619	0.0002634
3	Raybit	0.01346	0.012960	0.01345	0.01346	0.0002571	0.0002119	0.0002528	0.0002548
4		WMAE0	WMAE1	WMAE2	WMAEML	WMSE0	WMSE1	WMSE2	WMSEML
5	Logit	955.388	1143.383	962.856	959.318	31.592	42.104	31.559	31.570
6	Probit	973.572	1157.956	978.093	975.187	32.589	42.896	32.571	32.581
7	Raybit	953.510	1143.878	961.957	958.031	31.543	42.159	31.507	31.518

Source: own elaboration.

Another model related to the household possessing the PC assumed an explanatory variable x_{2i} as an available income per person (column 3 in table 1). The results of calculations are shown in table 3 with the labeling identical as in table 2.

Table 3.

Errors of the models: logit, probit and raybit for explanatory variable x_{2i}

	MAE0	MAE1	MAE2	MAEML	MSE0	MSE1	MSE2	MSEML
Logit	0.009538	0.010287	0.009561	0.009557	0.0001530	0.0001492	0.0001523	0.0001526
Probit	0.009732	0.010510	0.009720	0.009723	0.0001580	0.0001535	0.0001586	0.0001578
Raybit	0.009545	0.010289	0.009558	0.009563	0.0001530	0.0001492	0.0001523	0.0001525
	WMAE0	WMAE1	WMAE2	WMAEML	WMSE0	WMSE1	WMSE2	WMSEML
Logit	454.606	583.21	456.855	456.444	14.4814	15.332	14.473	14.4754
Probit	473.849	604.317	472.204	467.372	14.956	15.890	14.960	14.939
Raybit	456.119	582.982	456.191	457.701	14.4815	15.338	14.472	14.4749

Source: own elaboration.

On the basis of the data included in table 3 it can be concluded that the smallest errors are obtained through the raybit method – in 9 cases, and the logit method – in 7 cases. Considering the total values of errors of MAE, MSE, WMAE and WMSE presented in tables 2 and 3, it can be noticed that the smallest values of the aforementioned errors were obtained for the following probabilities: p_0 – 6 cases, p_1 – 10 cases, p_2 – 6 cases, p_{ML} – 2 cases. It only validates the application of the probability p_1 in the same way as p_0 and p_2. There was no point in examining the model with two explanatory variables x_{1i} and x_{2i}, since they are strongly correlated – the Pearson's correlation coefficient equaling 0.9985.

5. RESULTS OF COMPUTER SIMULATIONS

In order to verify the applicability of particular models (logit, probit, raybit), computer simulations were conducted.

In accordance with equation (47) a random variable S with a Bernoulli distribution was determined (Devroye, 1986) and takes the value:

$$s_k = \begin{cases} 1 & for \quad r_k \le P \\ 0 & for \quad r_k > P \end{cases}, \tag{47}$$

where $r_k \in [0;1]$ are uniform random variables,
P – theoretical probability,
$k = 1,2,\ldots, M$.

The observed value of a binomially distributed random variable Z is given by (Devroye, 1986):

$$z = \sum_{k=1}^{M} s_k.$$

The generated empirical value of probability was derived from:

$$p = \frac{z}{M}, \tag{48}$$

where M is the number of random variable in Bernoulli process.

The calculations were conducted for M = 50. The interval [0 ; 1] was divided into ten sub-intervals of the length 0.1 each. For each sub-interval of the form $[A_n; A_{n+1}]$, where $A_n = 0.1 \cdot n$; $n = 0,1,2,...,9$ the values of the theoretical probability were determined:

$$P_{in} = A_n + \frac{i}{100}, \text{ where } i = 0,1,...,10. \tag{49}$$

From equation (48) the empirical probability p_i was determined. For the values of the theoretical probability obtained from equation (49), a random number

generator with a binomial distribution was used, which provided the values of the empirical probability p_i. For these values, the logit, probit and raybit methods were applied. For the obtained estimations $\hat{p}_i$ the error measures were calculated (equtions (43),(44),(45),(46)).

During the computer simulations, for each value i and n (equation (49)), $K = 16000$ repetitions were made. The repetitions consisted in restarting the random number generator. The error measures were calculated as a mean from K repetitons.

The results of the computer simulations are presented in table 4. Rows 1 and 7 contain the values of the theoretical probability P (equation (49)).

In rows 2 and 8 next to the names of the models, in brackets, the numbers of cases for which a given model yielded the smallest errors are provided. The total number of cases for particular probability sub-intervals $[A; A+0.1]$ is 16 – four methods (p_0, p_1, p_2, p_{ML}) multiplied by four criteria of an error. The total number of resaults, for 10 probability sub-intervals equals 160. By adding up the figures in brackets the number of cases with the smallest error is obtained: LOG 30 (17.5%), PRO 74 (46.25%), RAY 58 (36.3%).

Table 4.

Errors of models: logit, probit and raybit obtained through computer simulations

1	Probability	$P \in [0; 0.1]$	$P \in [0.1; 0.2]$	$P \in [0.2; 0.3]$	$P \in [0.3; 0.4]$	$P \in [0.4; 0.5]$
	1	2	3	4	5	6
2	Model	LOG (1) PRO (2) RAY (13)	LOG (4) PRO (2) RAY (10)	LOG (4) PRO (3) RAY (9)	LOG (3) PRO (5) RAY (8)	LOG (5) PRO (3) RAY (8)
3	MAE	p_1	p_{ML}	p_{ML}	p_0	p_0
4	MSE	p_{ML}	p_{ML}	p_{ML}	p_{ML}	p_{ML}
5	WMAE	p_1	p_2	p_0	p_2	p_0
6	WMSE	p_0	p_0	p_2	p_0	p_2
7	Probability	$P \in [0.5; 0.6]$	$P \in [0.6; 0.7]$	$P \in [0.7; 0.8]$	$P \in [0.8; 0.9]$	$P \in [0.9; 1.0]$
8	Model	LOG (4) PRO (11) RAY (1)	LOG (1) PRO (11) RAY (4)	LOG (3) PRO (10) RAY (3)	LOG (3) PRO (12) RAY (1)	LOG (0) PRO (15) RAY (1)
9	MAE	p_0	p_0	p_{ML}	p_{ML}	p_{ML}
10	MSE	p_{ML}	p_{ML}	p_{ML}	p_{ML}	p_{ML}
11	WMAE	p_0	p_1	p_2	p_2	p_1
12	WMSE	p_2	p_0	p_0	p_0	p_0

Source: own elaboration.

It shows that in terms of the goodness of fit, the probit model is the best one, the raybit model is worse and the logit model is the worst. Furthermore it should be noticed that the raybit model is substantially better (in fact twice as good) compared with the logit model.

The following rows (from 3 to 6) contain the information about which equation that defines the probability leads to the smallest value of a selected error measure. In the case of MAE, it is as follows: p_{ML} (five times), p_0 (four times) and p_1 (once). In the case of MSE, there is a clear advantage of the probability determined by applying MLE (p_{ML}) – all ten cases.

In the case of WMAE the following was observed: $p_0(3)$, $p_1(3)$ and $p_2(4)$. WMSE takes the smallest value for: $p_0(7)$ and $p_2(3)$.

Table 5 was compiled on the basis of the results included in table 4. The only difference are the intervals, which now take the form $[0 ; A]$, where $A = 0.1, 0.2, 0.3 \ldots 1$. Rows 2 and 4 in table 5 contain the sum of subsequent columns in rows 2 and 8 in table 4.

Table 5.

Errors of the models: logit, probit and raybit obtained through computer simulations (cont.)

1	Probability	$P \in [0; 0.1]$	$P \in [0; 0.2]$	$P \in [0; 0.3]$	$P \in [0; 0.4]$	$P \in [0; 0.5]$
	1	2	3	4	5	6
2	Model	LOG (1) PRO (2) RAY (13)	LOG (5) PRO (4) RAY (23)	LOG (9) PRO (7) RAY (32)	LOG (12) PRO (12) RAY (40)	LOG (17) PRO (15) RAY (48)
3	Probability	$P \in [0; 0.6]$	$P \in [0; 0.7]$	$P \in [0; 0.8]$	$P \in [0; 0.9]$	$P \in [0; 1.0]$
4	Model	LOG (21) PRO (26) RAY (49)	LOG (22) PRO (37) RAY (53)	LOG (25) PRO (47) RAY (56)	LOG (28) PRO (59) RAY (57)	LOG (28) PRO (74) RAY (58)

Source: own elaboration.

The data shown in table 5 shows the advantage of the raybit model for $P \in [0; A]$ where $A = 0.1, 0.2, \ldots 0.8$. It is only for $P \in [0; 0.9]$ and $P \in [0; 1.0]$ that the probit model gains the advantage.

The logit model performs worst of all analyzed models for any value from the interval $P \in [0; A]$.

The data shown in table 4 demonstrates a variability in the number of cases when a given method yields the smallest error in relation to the value of the probability. In order to explain this phenomenon, the following numerical experiment was conducted. A random number generator was replaced by the values of the theoretical probability $P_i = 0.01 \cdot (1 + i)$, where $i = 0,1,\ldots,98$, which were used in place of the values of the empirical probability. Applying equations (27)–(29) and (38)–(40) the values of the

probability p_0 and p_2 were determined. The results of the calculations are shown in figures 1–4, where a dashed line represents the raybit model and a solid line – the linear model. Figure 1 proves that the results for the raybit model for $P_i \in [0.01; 0.5]$ are very similar to the results for the linear model, which results in very small values of the error. This is the reason why the raybit model has a clear advantage over other models for this probability interval.

The probability $pPRO_{0,i}$ obtained for the probit model for the same interval shows much larger nonlinearity. However for the interval $P_i \in [0.5; 0.99]$ the probit model fits well with the linear model.

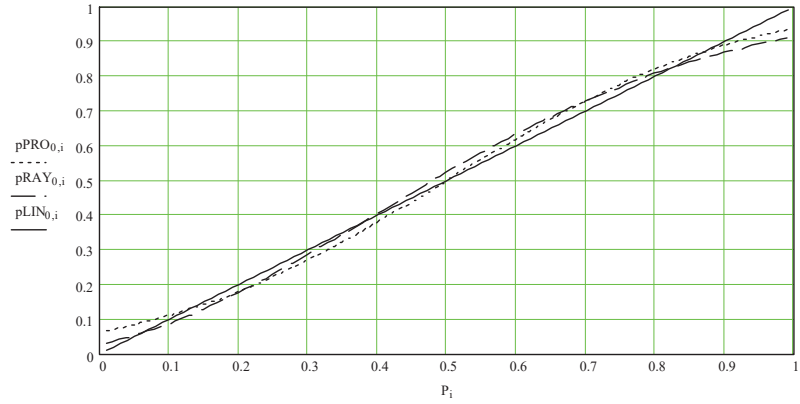

Figure 1. The results of the probability p_0 calculations for $P_i \in [0.01; 0.99]$
Applied labeling: dotted line $pPRO_{0,i}$ (probit model), dashed line $pRAY_{0,i}$ (raybit model), solid line $pLIN_{0,i}$ (linear model).
Source: own elaboration.

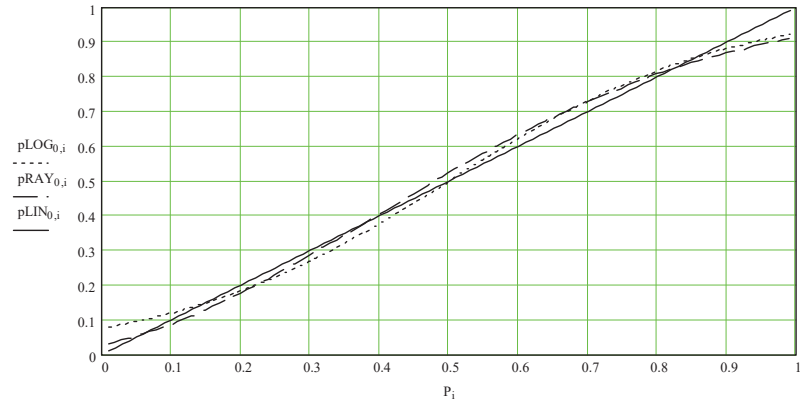

Figure 2. The results of the probability p_0 calculations for $P_i \in [0.01; 0.99]$
Applied labeling: dotted line $pLOG_{0,i}$ (logit model), dashed line $pRAY_{0,i}$ (raybit model), solid line $pLIN_{0,i}$ (linear model).
Source: own elaboration.

On the basis of figure 2 it can be noticed that the probability $pLOG_{0,i}$ obtained for the logit model shows much larger nonlinearity (than the raybit model), especially for $P_i \in [0.01; 0.2]$ and $P_i \in [0.6; 0.99]$.

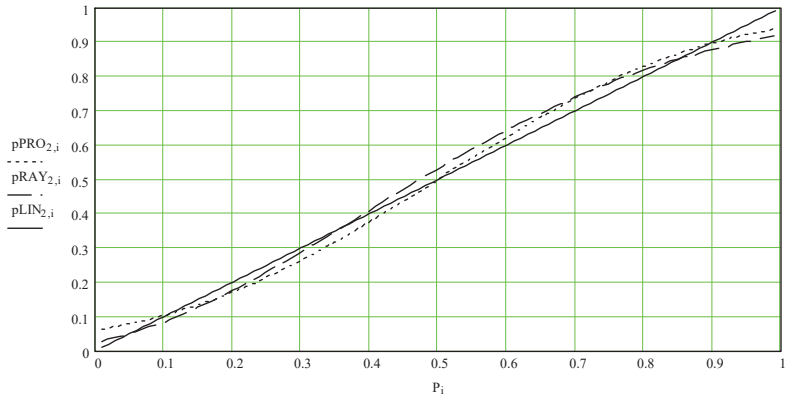

Figure 3. The results of the probability p_2 calculations for $P_i \in [0.01; 0.99]$
Applied labeling: the same as in figure 1.
Source: own elaboration.

The situation described in relation to figure 1 can be also observed in figure 3.

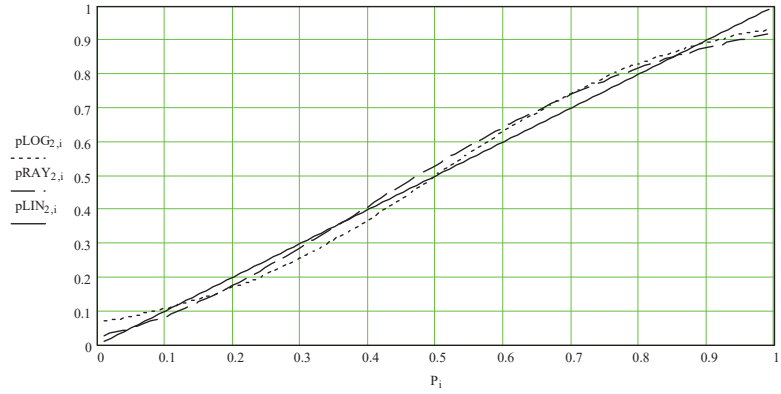

Figure 4. The results of the probability p_2 calculations for $P_i \in [0.01; 0.99]$
Applied labeling: the same as in figure 2.
Source: own elaboration.

The results observed in figure 2 can be also observed in figure 4.

6. CONCLUSION

In the paper the estimation of the parameters of qualitative econometric models was discussed including: the logit model, the probit model and the raybit model. The following methods of estimation were considered. The generalized least squares (equation (23)), where the elements of a diagonal covariance matrix are determined on the basis of the empirical probability. The method leads to the estimation of a theoretical probability labelled as p_0 (equations (27)−(29)). The next method is two-step. As a first step, using OLS, the estimation of the probability p_1 was determined (equations (31)−(33)). As a second step, GLS was used, where the elements of a diagonal covariance matrix were determined on the basis of the probability p_1. Consequently, the estimation of the probability labelled as p_2 was obtained (equations (38)−(40)). Although the probability p_1 was used to calculate the probability p_2, it was also treated as yet another value of the theoretical probability estimation. The last method of estimation of a qualitative econometric model was the maximum likelihood method, where the probability estimation was labelled as p_{ML}.

With reference to the computational examples (tables 2 and 4), the raybit model, proposed in this paper, proved to be the best out of the three models under study. In computer simulations this model showed clear advantage for probability $P \in [0; 0.8]$ (table 5). Only for $P \in [0; 0.9]$ and $P \in [0; 1.0]$ the probit model performs best. Despite the fact that for the above mentioned probability intervals the raybit model is worse than the probit model, it still has its advantages, namely, the analytical forms of the cumulative distribution function as well as the inverse function to the cumulative distribution function.

It should be noticed that in the whole probability interval the raybit model yields a smaller error than the logit model.

It means that while analyzing a binomial qualitative variable, along with the classic logit and probit models, it is worth taking into account the results of the raybit model.

REFERENCES

Amemiya T., (1981), Qualitative Response Models: A Survey, *Journal of Economic Literature*, December, 1483–1536.

Budżety gospodarstw domowych w 2012 r., (2013), Household Budget Survey in 2012, GUS, Warsaw.

Celik A. N., (2003), A Statistical Analysis of Wind Power Density Based on the Weibull and Rayleigh Models at the Southern Region of Turkey, *Elsevier, Renewable Energy,* 29, 593–604.

Chow G. C., (1983), *Econometrics*, Mc-Graw-Hill Inc, New York.

Devroye L., (1986), *Non-Uniform Random Variate Generation*, Springer-Verlag, New York.

Finney D. J., (1983), *Probit Analysis*, Cambridge University Press.

Gruszczyński M., (2012), *Empiryczne finanse przedsiębiorstw. Mikroekonometria finansowa,* Difin, Warszawa.

Guzik B., Appenzeller D., Jurek W., (2005), *Prognozowanie i symulacje. Wybrane zagadnienia*, Wydawnictwo Akademii Ekonomicznej w Poznaniu, nr 168, Poznań.

Jajuga K., (1989), Modele z dyskretną zmienną objaśnianą, in: Bartosiewicz S. (ed.), *Estymacja modeli ekonometrycznych*, PWE, Warszawa.

Judge G. G., Griffiths W. E., Hill R. C., Lee T. C., (1980), *Theory and Practice of Econometrics*, Wiley, New York.

Koutsoyiannis D., (2003), On the Appropriateness of the Gumbel Distribution in Modelling Extreme Rainfall, *Proceedings of the ESF LESC Exploratory Workshop held at Bologna*, Italy, October 24–25, 303–319.

Maddala G. S., (1983), *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge University Press.

Maddala G. S., (1992), *Introduction to Econometrics*, Macmillan Publishing Company, New York.

McFadden D., (1984), Econometric Analysis of Qualitative Response Models, in: Gribliches Z., Intriligator M. D., (eds.), *Handbook of Econometrics*, Vol. II, Elsevier Science Publishers BV.

Nerlove M., Press S. J., (1973), *Univariante and Multivariante Log-Linear and Logistic Models*, R-1406-EDA/NIH, Dec., RAND Santa Monica, CA, 90406.

Polakow D. A., Dunne T. T., (1999), Modelling Fire-Return Interval *T*: Stochasticity and Censoringin the Two-Parameter Weibull Model, *Elsevier, Ecological Modelling*, 121, 79–102.

Purczyński J., Porada-Rochoń M., (2015), Ocena jakości modeli ze zmienną dychotomiczną, *Logistyka*, 3, 4064-4073.

Rinne H., (2009), *The Weibull Distribution. A Handbook*, CRC PressTaylor & Francis Group, LLC.

Train K., (2009), *Discrete Choice Methods with Simulation*, Cambridge University Press.

MODEL RAYBITOWY I OCENA JEGO JAKOŚCI W PORÓWNANIU Z MODELEM LOGITOWYM I PROBITOWYM

Streszczenie

W pracy zaproponowano nowy model dla zmiennej objaśnianej zero-jedynkowej (binarnej, dychotomicznej). Nazwa modelu raybit wynika stąd, że prawdopodobieństwo odpowiada dystrybuancie rozkładu Rayleigha. Ocenę jakości modeli przeprowadzono z wykorzystaniem 4 definicji błędu: MSE, MAE, WMSE, WMAE. Rozpatrzono dwa przykłady obliczeniowe, które wykazały, że model raybitowy prowadzi do mniejszych wartości błędu, niż model logitowy i probitowy. Wykonano symulacje komputerowe z wykorzystaniem generatora liczb losowych o rozkładzie dwumianowym. Przeprowadzone symulacje wykazały, że dla wartości prawdopodobieństwa teoretycznego z przedziału $P_i \in [0; 0{,}8]$ model raybitowy przewyższa pozostałe dwa modela prowadząc do mniejszej wartości błędu.

Słowa kluczowe: jakościowe modele ekonometryczne, model logitowy, model probitowy

THE RAYBIT MODEL AND THE ASSESSMENT OF ITS QUALITY IN COMPARISON WITH THE LOGIT AND PROBIT MODELS

Abstract

A new model for a dependent variable taking the value 0 or 1 (binary, dichotomous) was proposed. The name of the proposed model – the raybit model – stems from the fact that the probability corresponds to the Rayleigh cumulative distribution function. The assessment of the quality of selected models was conducted with the use of four definitions of error: MSE, MAE, WMSE, WMAE. Two computational examples were considered, which proved that the raybit model yields smaller values of error than the logit and probit models. Computer simulations were conducted using a random number generator with a binomial distribution. They proved that for the values of the theoretical probabilityfor the interval $P_i \in [0; 0.8]$ the raybit model outperforms the other two models yielding a smaller value of error.

Keywords: qualitative econometric models, logit model, probit model