Folia Oeconomica Stetinensia
Volume 18 (2018) Issue 1
DOI: 10.2478/foli-2018-0014

GOODNESS OF FIT MEASURES OF MODELS WITH BINARY DEPENDENT VARIABLE WHICH TAKE INTO ACCOUNT HETEROSKEDASTICITY OF A RANDOM ELEMENT

Prof. Jan Purczyński

West Pomeranian University of Technology
Department of Signal Processing and Multimedia Engineering
26. Kwietnia 10, 71-126 Szczecin, Poland
e-mail: JanPurczynski@zut.edu.pl

Kamila Bednarz-Okrzyńska, Ph.D.

University of Szczecin
Faculty of Management and Economics of Services
Department of Quantitative Methods
Cukrowa 8, 71-004 Szczecin, Poland
e-mail: kamila.bednarz@wzieu.pl

Received 30 October 2017, Accepted 30 March 2018

Abstract

The paper tackles a problem which arises during the analysis of binary models, and which is the heteroskedasticity of a random element manifested by the variable value of variance. In the paper, the following probability models, used in the analysis of a dichotomic variable, were considered: a logit model, probit model, and raybit model, which is a model proposed by the authors. The following measures of goodness of fit, present in the field literature, were considered: MSE, MAE, WMSE, and WMAE. A new measure of goodness of fit of a model was proposed, which limits the amplitude of varying values of variance.

Keywords: qualitative econometric models, logit model, probit model

JEL classification: C15, C25, C63

Introduction

One of the problems arising during the analysis of binary models is the heteroskedasticity of a random element manifested by the variable value of variance. In the paper, the following probability models, used in the analysis of a dichotomic variable, will be considered: a logit model, probit model, and raybit model, which is a model proposed by the authors. The aim of this paper is to modify the currently applied measures of goodness of fit of theoretical probability to empirical data in a way that would maximally reduce the impact of heteroskedasticity. Three models of probability and three methods of estimation, including the Maximum Likelihood Method (MLM), will be analyzed.

As a main tool, computer simulations will be used with the application of a Bernoulli distribution random number generator.

As a research result, a modified form of the Weighted Mean Squared Error (WMSE) and Weighted Mean Absolute Error (WMAE) will be expected.

In the estimation of a binary dependent variable, the ordinary least squares (OLS) can be used. However, the solution obtained through this method has one vital drawback, namely, the theoretical probability may fall outside the interval [0, 1]. In order to avoid this drawback, it is assumed that the probability corresponds to a cumulative distribution function of the random variable distribution. In the case of the logistic distribution, the logit model is obtained, and in the case of the normal distribution – the probit model.

In the paper by Purczyński and Bednarz-Okrzyńska (2017), another model was proposed, in which the probability is expressed by Rayleigh cumulative distribution function, hence the name of the model – raybit. The paper promises to analyze binary decision problems. However, in equation (1), the authors move from a single decision problem to the results of the individual decisions of group members, which are summarized as group ratios.

1. Probability models for binary variable

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

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

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

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

The easiest model is a linear model of probability:

$$p = X\alpha + \varepsilon \tag{2}$$

where:

p – I – dimensional vector of empirical probabilities,

X – $[I \times (k + 1)]$ dimensional matrix including k number of explanatory variables,

α – $(k + 1)$ vector of parameters,

ε – I – dimensional vector of random elements.

Based on equation (2), the following can be observed

$$p_i = P_i + \varepsilon_i \tag{3}$$

where:

p_i – empirical probability of occurrence of event A for i-th value of the vector of explanatory variables,

P_i – probability of occurrence of event A for i-th value of the vector of explanatory variables,

ε_i – disturbance: $E(\varepsilon_i) = 0$ and $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$.

Since variable y_i (equation (1)) has binomial distribution, the variance of a disturbance is given by the relation (Judge et al., 1980):

$$V(\varepsilon_i) = \frac{P_i(1 - P_i)}{n_i} \tag{4}$$

which means that the random variable appearing in equation (4) is heteroskedastic.

For the linear probability model, the following relation is observed:

$$P_i = x_i^T\alpha \tag{5}$$

where x_i^T is i-th row of explanatory variable matrix.

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

$$P_i = F(x_i^T\alpha) \tag{6}$$

where F is the cumulative distribution function.

Two models are most commonly applied:

– a logit model, hereafter referred to as LOG

$$P_i = L(x_i^T\alpha) = [1 + e^{-x_i^T\alpha}]^{-1} \tag{7}$$

where L denotes the cumulative distribution function of a logistic distribution, and

– a probit model, hereafter referred to as PRO

$$P_i = \Phi(x_i^T\alpha) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_i^T\alpha} e^{-\frac{t^2}{2}} dt \tag{8}$$

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

In the paper by Purczyński and Bednarz-Okrzyńska (2017), the probability model applying Rayleigh cumulative distribution function was proposed:

– a raybit model, hereafter referred to as RAY

$$P_i = R(x_i^T\alpha) = 1 - \exp[-(x_i^T\alpha)^2] \tag{9}$$

where R denotes Rayleigh cumulative distribution function.

In the analysis of each model, the following three steps can be singled out (Jajuga, 1989):

A. The first step: estimation of vector α_0 of parameters α

$$\alpha_0 = (X^T W^{-1} X)^{-1} X^T W^{-1} v \tag{10}$$

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

$$w_i = [n_i p_i (1 - p_i)]^{-1} \text{ LOG} \tag{11}$$

$$w_i = \frac{p_i(1 - p_i)}{n_i \{\varphi[\Phi^{-1}(p_i)]\}^2} \text{ PRO} \tag{12}$$

where:

φ(t) denotes the density function of the standardized normal distribution,

$\Phi^{-1}(p_i)$ is the inverse function to the cumulative distribution function of the standardized normal distribution

$$w_i = \frac{p_i}{4n_i\left(1-p_i\right)\ln\frac{1}{1-p_i}} \qquad \text{RAY} \tag{13}$$

Depending on the model, vector *v* is given by the following formula:

$$v_i = \ln\left(\frac{p_i}{1-p_i}\right) \qquad \text{LOG} \tag{14}$$

$$v_i = \Phi^{-1}\left(p_i\right) \qquad \text{PRO} \tag{15}$$

$$v_i = \sqrt{-\ln\left(1-p_i\right)} \qquad \text{RAY} \tag{16}$$

Estimation of the theoretical probability:

$$p0_i = L(x_i^T\alpha0) \qquad \text{LOG} \tag{17}$$

$$p0_i = \Phi(x_i^T\alpha0) \qquad \text{PRO} \tag{18}$$

$$p0_i = R(x_i^T\alpha0) \qquad \text{RAY} \tag{19}$$

B. The second step

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

$$\alpha1 = (X^TX)^{-1}X^Tv,$$

where *v* is defined by formulas (14)–(16).

The estimation of theoretical probability $p1_i$ is derived from the formulas analogous to (17), (18), and (19).

C. The third step

Estimation of vector α2 of parameters α

$$\alpha2 = (X^TW1^{-1}X)^{-1}X^TW1^{-1}v \tag{20}$$

where *v* is defined by (14)–(16).

*W*1 is a diagonal covariance matrix, where the elements on the main diagonal equal:

$$w1_i = [n_i p1_i(1-p1_i)]^{-1} \qquad \text{LOG} \tag{21}$$

$$w1_i = \frac{p1_i(1-p1_i)}{n_i\left\{\varphi\left[\Phi^{-1}\left(p1_i\right)\right]\right\}^2} \quad \text{PRO} \tag{22}$$

$$w1_i = \frac{p1_i}{4n_i\left(1-p1_i\right)\ln\frac{1}{1-p1_i}} \quad \text{RAY} \tag{23}$$

Estimation of theoretical probability:

$$p2_i = L(x_i^T\alpha 2) \quad \text{LOG} \tag{24}$$

$$p2_i = \Phi(x_i^T\alpha 2) \quad \text{PRO} \tag{25}$$

$$p2_i = R(x_i^T\alpha 2) \quad \text{RAY} \tag{26}$$

The last method of theoretical probability estimation is the Maximum Likelihood Method (MLM). The description of MLM in relation to logit and probit models can be found in the paper by Chow (1995). The application of MLM for the raybit model was described in the paper by Purczyński and Bednarz-Okrzyńska (2017). The theoretical probability obtained by means of MLM will be labeled as pM_i.

2. Estimating the error of a model

The most popular measure of goodness of fit of a model is the Mean Squared Error (MSE):

$$MSE = \frac{1}{I}\sum_{i=1}^{I}(p_i - pt_i)^2 \tag{27}$$

where:

p_i – empirical probability (equation (1)), and

pt_i – theoretical probability.

As pt_i, the results of the following three methods ($p0_i$, $p2_i$, pM_i) are taken. Guzik, Appenzeller, Jurek (2005) recommend equation (27) as a criterion of goodness of fit of a theoretical probability model.

Another measure is the mean absolute error (MAE):

$$MAE = \frac{1}{I}\sum_{i=1}^{I}|p_i - pt_i| \tag{28}$$

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

$$WMSE = \sum_{i=1}^{I} \frac{n_i (p_i - pt_i)^2}{p_i (1 - p_i)} \tag{29}$$

The main problem lies in the fact that the variance of MSE (equation (27)) and MAE (equation (28)) depend heavily on the value of the empirical probability. Therefore, a recommended measure of goodness of fit is WMSE (equation (29)). This issue was discussed in the paper by Purczyński and Porada-Rochoń (2015), where computer simulations were carried out using a random number generator with binominal distribution. As a result of those studies, yet another measure of goodness of fit was proposed, namely Weighted Mean Absolute Error (WMAE):

$$WMAE = \sum_{i=1}^{1} \frac{n_i \left| p_i - pt_i \right|}{\sqrt{p_i (1 - p_i)}} \tag{30}$$

In order to examine the phenomenon of heteroskedasticity, the variances of the particular measures of goodness of fit of a model were calculated. Figure 1 shows the values of variance of MSE measure (equation (27)), determined just to estimate probability *p*0 (equations (17)–(19)). In the case of estimation *p*2 and *pM*, distributions analogous as in Figure 1 are obtained.

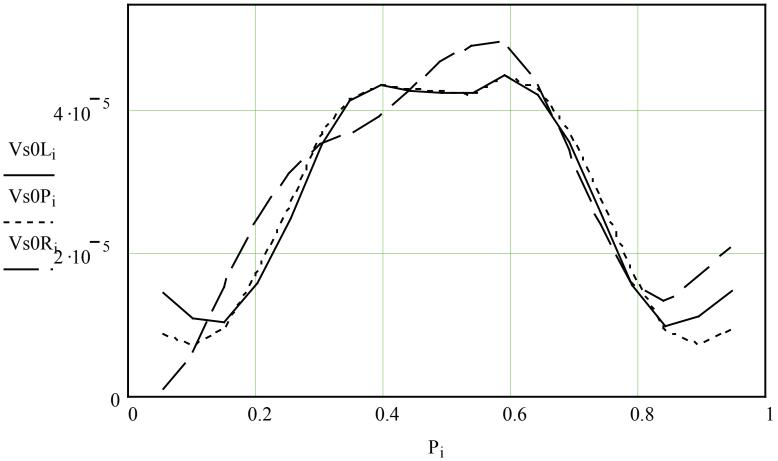

Figure 1. Values of variance *Vs*0 of equation (27) for estimation of *p*0 as a function of empirical probability. The solid line *Vs*0*L* corresponds to the logit model, the dotted line *Vs*0*P* corresponds to the probit model, and the dashed line *Vs*0*R* – to the raybit model.

Source: author's own study.

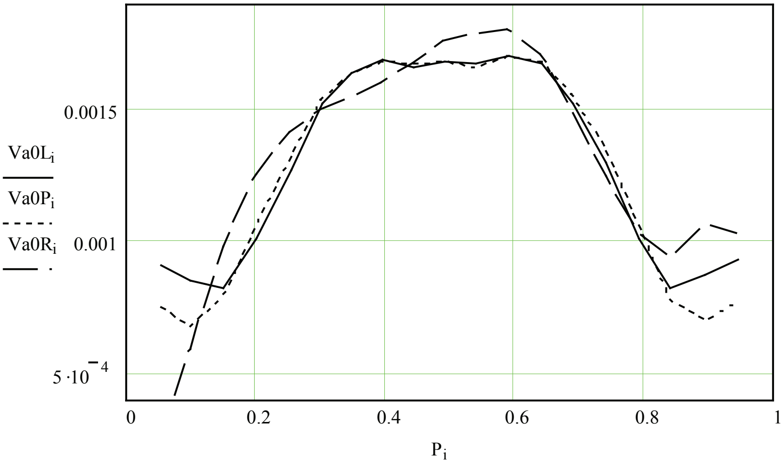

Figure 2. Values of variance *Va*0 of equation (28) for estimation of *p*0 as a function of empirical probability. The solid line *Va*0*L* corresponds to the logit model, the dotted line *Va*0*P* corresponds to the probit model, and the dashed line *Va*0*R* – to the raybit model.

Source: author's own study.

3. Measures of goodness of fit of a model taking into account the phenomenon of heteroskedasticity

Figure 3 presents the values of variance of WMSE measure (equation (29)) determined to estimate probability *p*2 (equations (24)–(26)).

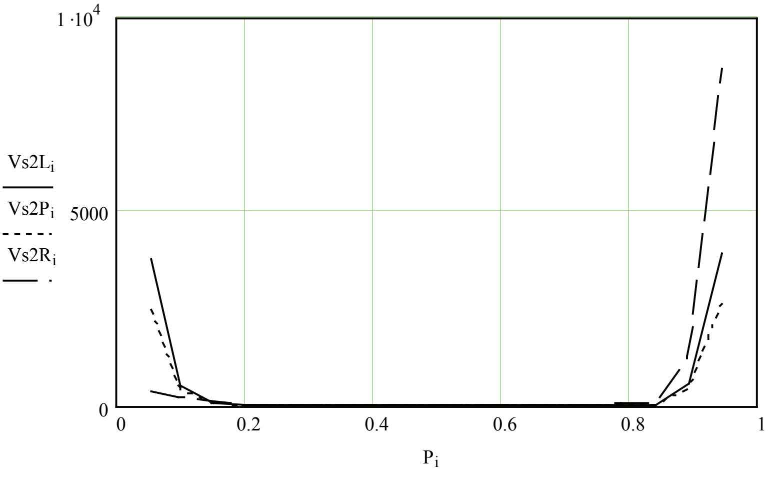

Figure 3. Values of variance *Vs*2 of equation (29) for the estimation of *p*2 as a function of empirical probability. Applied labeling: the same as in Figure 1.

Source: author's own study.

Figure 4 present the values of variance of WMAE measure (equation (30)) determined to estimate probability *p*2.

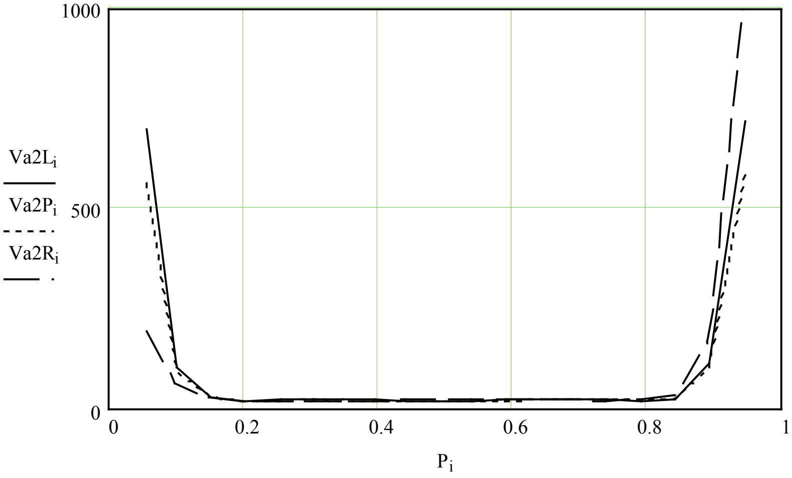

Figure 4. Values of variance *Va*2 of equation (30) for the estimation of *p*2 as a function of empirical probability. Applied labeling: the same as in Figure 2.

Source: author's own study.

As already mentioned, the above measures (WMSE and WMAE) were introduced in order to limit the variability of variance. However, as shown in Figures 3 and 4, the situation worsens substantially, namely, the range of the variance is larger than in the case of the two other measures of goodness of fit – MSE and MAE (Figures 1 and 2). It is further supported by the results presented in Table 1. The reason for that is the fact that equations (29) and (30) lead to very large values in the case when probability is close to zero or one. Therefore, a modified form of both equations is considered:

$$WMSE` = \sum_{i=1}^{I} \frac{n_i (p_i - pt_i)^2}{(P0 + p_i)(P1 - p_i)} \qquad 31$$

$$WMAE` = \sum_{i=1}^{I} \frac{n_i |p_i - pt_i|}{\sqrt{(P0 + p_i)(P1 - p_i)}} \qquad 32$$

Figure 5 presents the values of variance of expression (31) obtained for parameters *P*0 and *P*1, which lead to the smallest variations of variance. In the case of the logit and probit models, *P*0 = 0.1 and *P*1 = 1.1, and for the raybit model *P*0 = 0.03 and *P*1 = 1.13.

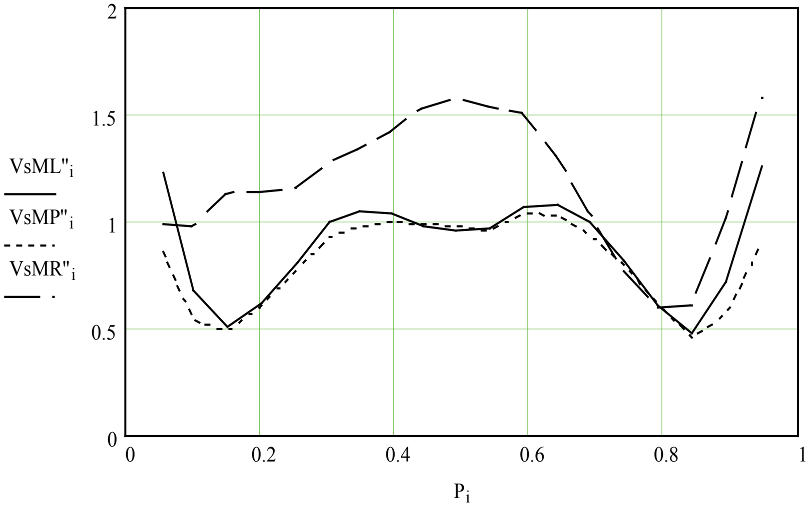

Figure 5. Values of variance *VsM* of expression (31) for estimation *pM* as a function of empirical probability. Applied labeling: the same as in Figure 1.

Source: author’s own study.

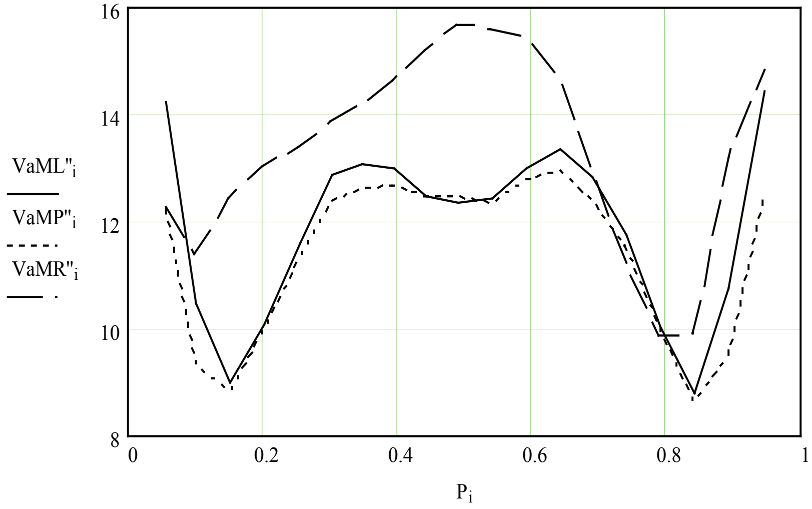

Figure 6. Values of variance *VaM* of expression (32) for estimation *pM* as a function of empirical probability. Applied labeling: the same as in Figure 2.

Source: author’s own study.

With reference to Figures 1 and 2, Table 1 was created, which includes the values of the ratio of the maximum value to the minimum value of MSE variance (equation (27)) and MAE variance (equation (28)).

Table 1. Ratio of the maximum value to the minimum value of MSE variance (equation (27)) and MAE variance (equation (28))

	vS0	vS2	vSM	vA0	vA2	vAM
Logit	4.535	19.844	8.828	2.073	4.404	2.875
Probit	6.429	24.130	9.673	2.500	4.775	3.088
Raybit	45.182	63.417	51.500	7.200	8.577	7.143

Source: author's own study.

Based on Table 1, it can be noticed that the ratio of the maximum value to the minimum value of MAE (vA) is roughly equal to the root of the ratio of the maximum value to the minimum value of MSE variance (vS). It stems from the fact that error MAE represents the mean value of the absolute error, and error MSE equals the mean value of the square of the error. Based on Table 1, it can be also concluded that, the smallest variations in the values of variance can be observed for the logit model – both for MSE and MAE. A slightly larger amplitude of variations of variance is characteristic for the probit model, and the largest variations are observed for the raybit model.

Table 2 includes the results of the calculations for the modified forms of errors (equations (31) and (32)).

Table 2. Ratio of the maximum value to the minimum value of WMSE variance (equation (31)) and WMAE variance (equation (32))

	vS0	vS2	vSM	vA0	vA2	vAM
P0 = 0; P1 = 1						
Logit	12,801.64	1,832.74	5,638.21	107.58	39.46	68.32
Probit	8,227.54	1,215.22	3,916.98	84.72	31.73	56.85
Raybit	18,637.88	3,938.28	9,721.50	125.40	57.56	91.94
P0 = 0.1; P1 = 1.1 for LOG and PRO as well as P0 = 0.03; P1 = 1.13 – RAY						
Logit	4.443	3.392	2.643	2.105	1.804	1.639
Probit	4.443	3.491	2.277	1.826	1.816	1.500
Raybit	5.089	4.078	2.677	2.029	1.986	1.600

Source: author's own study.

The results presented in the upper part of Table 2 (the case of $P0 = 0$ and $P1 = 1$) correspond to the classical form of WMSE (equation(29)) and WMAE (equation(30)). The comparison of the results in Table 1 and 2 shows that equations (29) and (30) yield slightly worse results than equations (27) and (28).

The results presented in the lower part of Table 2 were obtained for the weighted sums (equations (31) and (32)) for the optimal values of parameters $P0$ and $P1$. A substantially smaller amplitude of variations of variance can be observed in relation to both the upper part of Table 2 (WMSE, WMAE) and the results included in Table 1 (MSE, MAE)). When narrowing down to the results obtained with MLM, for vSM, the smallest variations can be observed for the probit model, and similar values for both the logit and raybit models. However, when arranging the models in the order of increasing value vAM, the following sequence is obtained: probit, raybit, and logit.

The results presented in Tables 1 and 2 were obtained for the interval of empirical probability $p_i \in [0.05; 0.95]$. Stretching of the interval would result in larger values of the ratio of the maximum and minimum values of variance presented in Tables 1 and 2.

Conclusions

The paper tackles the problem arising during the analysis of binary models, which is the heteroskedasticity of a random element manifested by the variable value of variance. The following probability models, used in the analysis of a dichotomic variable, were analyzed: logit model, probit model, and raybit model, which is a model proposed by the authors. The following measures of goodness of fit, present in the field literature, were applied: MSE, MAE, WMSE, and WMAE (eq. (27)–(30)). As a result of computer simulations, performed with the use of a Bernoulli distribution random number generator, the relation between the variance and the value of the empirical probability was determined for the aforementioned measures of goodness of fit of a model (Figures 1–4). Another result of the computational experiment was the proposed form of a measure of goodness of fit of a model (equations (31) and (32)) which limits the amplitude of varying values of variance.

References

Amemiya, T. (1981). Qualitative Response Models: A Survey. *Journal of Economic Literature*, December, 483–536.

Chow, G.C. (1995), *Ekonometria*. Warszawa: Wydawnictwo Naukowe PWN.

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

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

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

Maddala, G.S. (2006). *Ekonometria*. Warszawa: Wydawnictwo Naukowe PWN.

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

Purczyński, J., Bednarz-Okrzyńska, K. (2017). The raybit model and the assessment of its quality in comparison with the logit and probit models. *Przegląd Statystyczny*, *3*, 305–322.