DE GRUYTER
OPEN

ECONOMETRICS. ADVANCES IN APPLIED DATA ANALYSIS
Year 2018, Vol. 22, No. 1
ISSN 1507-3866
e-ISSN 2449-9994

STATISTICAL MODELS IN ENTERPRISES DEFAULT RISK ASSESSMENT – AN EXAMPLE OF APPLICATION

Aneta Ptak-Chmielewska
SGH Warsaw School of Economics, Warsaw, Poland
e-mail: aptak@sgh.waw.pl

Piotr Kuleta
e-mail: piotr.kuleta@gmail.com

DOI: 10.15611/eada.2018.1.07
JEL Classification: C52, C58, G33

Abstract: Default risk assessment is crucial in the banking activity. Different models were developed in the literature using the discriminant analysis, logistic regression and data mining techniques. In this paper the logistic regression was applied to verify models proposed by R. Jagiełło for different sectors. As an alternative, the logistic regression model with the nominal variable SECTOR was applied on the pooled sample of enterprises. The dynamic approach using the Cox regression survival model was estimated. Including the nominal variable SECTOR only slightly increases the predictive power of the model (in the case of "defaults"). The predictive power of the Cox regression model is lower, the only advantage is the higher accuracy classification in the case of "defaulted" enterprises.

Keywords: default risk, logistic regression, Cox model.

1. Introduction

The first models predicting the insolvency of enterprises date from the 1930s and the subsequent years, among them are works by P.J. Fritz Patrick, and more advanced and spread multivariate models by W.H. Beaver and E.I. Altman date from the 1960s [Jagiełło 2013, pp. 5-6].

The very first papers by Z. Hellwig and U. Siedlecka on the early warning system in Poland were published in the 1980s and the 1990s [Pociecha (ed.) 2014, p. 14]. The multivariate models were launched in Poland by E. Mączyńska in the 1990s and continued by M. Pogodzińska and S. Sojak, J. Gajdka and T. Stos [Jagiełło 2013, pp. 32 and further]. The works by D. Hadasik (D. Appenzeller) in the late 90s are still based on the discriminant analysis with a different set of financial ratios, number of instances and homogeneity of the sample. In the works (the start of the 21st century) by A. Hołda, D. Wierzba, S. Sojak, J. Stawicki, B. Prusak, the tests of the accuracy of such models were performed using different ratios and samples achieving the prediction accuracy at a very high level, i.e. above 90% [Jagiełło 2013, pp. 41 and further]. In two research centers, works on discriminant analysis were conducted increasing the quality and accuracy of the prediction by increasing the homogeneity of samples (E. Mączyńska and M. Zawadzki, M. Hamrol and B. Czajka, cited by [Jagiełło 2013, pp. 48-52]).

For the first time the logit model in the prediction of bankruptcy was used by J. Ohlson in 1980 [Pociecha (ed.) 2014, p. 24]. Models utilizing the logistic regression in predicting bankruptcy risk in Poland are included in works by A. Hołda, M. Gruszczyński, P. Stępień and T. Strąk, D.Wędzki, B. Prusak, published in 2000-2005 [Jagiełło 2013, pp. 53 and further; Pociecha (ed.) 2014, pp. 25-26]. In this article we do not aim to review the current literature on bankruptcy. A more detailed literature review was published in Pociecha [Pociecha (ed.) 2014].

The methods using the multivariate statistics and logit models are called the probabilistic approach. Machine learning methods, decision trees and neural networks (data mining) are called the nonparametric approach [Pociecha (ed.) 2014, pp. 16-17].

Due to more effective computers and more advanced methods, the machine learning methods became more and more popular and widely used. The review of data mining models used in the research of prediction of the bankruptcy of enterprises was published in work edited by J. Pociecha [Pociecha (ed.) 2014].

In R. Jagiełło [2013], the sample was split according to the sector of activity, and separate models were estimated for each sector using a sample of 80 enterprises in each model. Different ratios were significant in models for different sectors. The high prediction accuracy of such models was the inspiration for this research and led to the following research hypothesis: including information about the sector of activity significantly increases the effectiveness of the model predicting the default of enterprises.

The main goal of the paper was the comparison of R. Jagiełło's model with the conclusion about the possibility of the usage of such a model in practice. The authors' own model was proposed. In this paper the pooled model using information about the sector of activity was applied. Two types of models were used: the logistic regression and the survival model.

2. The research sample – a description

The sample used in this paper consisted of 3,112 financial statements from small and medium enterprises (SMEs). The financial statements came from the homogenous period of years 2004-2008. The data sample included 494 financial statements of enterprises recognized as defaulted. In this group of enterprises (defaulted) bankrupted and non-performing enterprises were included (see Table 1).

Table 1. Research sample by sectors

Sector	*Default*	*N*	*N*-sector
1 − Manufacturing	0 – no	1199	1410
	1 – yes	211	
2 − Construction	0 – no	372	444
	1 – yes	72	
3 − Transportation	0 – no	140	168
	1 – yes	28	
4 − Trade	0 – no	606	739
	1 – yes	133	
5 − Services	0 – no	301	351
	1 – yes	50	

Source: own elaboration.

Table 2. Sectors according to classification PKD 2007 (Polish Classification of Activities)

Sector	Sections PKD
Manufacturing	C, D, E
Construction	F
Transportation	H
Trade	G
Services	I, J, K, L, M, N, O, P, Q, S

Source: own elaboration.

3. The research methods

The logistic regression model was proposed as the research method. The financial ratios proposed by Jagiełło [2013] were included in the model. To compare, the semi-parametric Cox survival model was applied.

3.1. The logistic regression

The general form of the logistic regression model is as follows:

$$Y \sim B(1, p),$$

$$p = E(Y \mid X) = \frac{\exp(\beta X)}{1 + \exp(\beta X)},$$

where: $B(1, p)$ is the binomial distribution with the probability of success p; Y – the dependent variable; $X = (X_1, \ldots, X_k)$ – the independent explanatory variables; β – structural parameters in model.

The $P(Y = 1)$ takes the values from interval [0;1]. A cut-off point is an important element in the logistic regression model. The estimation based on the balanced sample usually takes 0.5 as a cut-off value. The structure of the sample (percentage of bankrupted enterprises) determines the cut-off value.

The interpretation of results is usually based on odds ratios (the ratio of odds in two groups or in the change of one unit in the explanatory variable). Logistic regression requires a lot of different assumptions to be fulfilled. The most important assumptions are: randomness of the sample, big sample, no collinearities in explanatory variables, and independence of observations.

3.2. The survival models

In the semiparametric model (the proportional hazards Cox model) only the regression part is parametrically specified (the interaction between processes), the *time* distribution is not parametrically specified (the nonparametric approach). It is assumed that the interval variable T means *time* to event occurrence. In the Cox regression model the dependent variable *time* is estimated as a hazard function:

$$h(t) = h_0(\mathrm{t})\exp(\alpha_1 x_1 + \cdots + \alpha_k x_k),$$

where: $h_0(t)$ – the baseline hazard; X_1, X_2,…X_k – the explanatory variables.

Cox [1972] proposed the partial maximum likelihood method for the estimation of such models. The likelihood function is divided into two parts: the first, including only parameters and the second, including parameters and the hazard function. The basic assumption in the Cox model is the hazard proportionality assumption. When this assumption is violated, the model becomes the non-proportional hazards model by including the interaction between variable X and the *time* of the process *t* [Blossfeld, Rohwer 2002].

4. Financial ratios used in model estimation

The approach to the estimation of models predicting bankruptcy is evolving. The types of utilized models and number of ratios are changing. In the 1970s and the 1980s, the most frequently used models consisted of the discriminant models and the logistic regression models. Later on the number of papers utilizing the neural network increased and the number of papers using discriminant functions decreased. Neural networks nowadays hold the dominant position in the world literature. Still, the logit models are frequently used and a new type of predictive models appeared. The most important issue in the estimation of predictive models is the selection of financial ratios. The financial ratios in the classic approach of the economic analysis of an enterprise's condition are organized in five areas [Sierpińska, Jachna 2004, pp. 144-145]:

- financial liquidity,
- profitability,
- market value of shares and equity,
- efficiency,
- debt and debt service.

In the models predicting bankruptcy in different publications, as many as 752 financial ratios were used in total. Only 18 of them are used in at least 10 publications. Those ratios are presented in Table 3.

Table 3. The most frequently used financial ratios in the bankruptcy prediction models

Financial ratios groups	Number of models
Net income/assets	54
Current liquidity ratio	51
Working capital/assets (assets coverage ratio)	45
Profit retained/assets (profitability ratio)	42
Profit before taxes/assets (profitability ratio)	35
Sales/assets (rotation of assets)	32
Quick liquidity ratio	30
Total liabilities/assets (debt ratio)	27
Current assets/assets (assets structure ratio)	26
Net income/net assets (net rotation ratio)	23
Total liabilities/assets (debt ratio)	19
Cash/assets (assets structure)	18
Market value of own equity/accountant value of debt (debt ratio)	16
Operational cash flow/assets (profitability ratio)	15
Operational cash flow/total liabilities (debt ratio)	14
Current liabilities/assets (debt ratio)	13
Operational cash flow/total liabilities (debt ratio)	12
Liquid assets/assets (assets structure ratio)	11

Source: [Pociecha (ed.) 2014] (cited following: [Bellovary, Giacomini, Akers 2007, p. 42]).

5. The extension of the Robert Jagiełło model

In his logistic regression model for different sectors of SMEs, R. Jagiełło used the sample of 400 enterprises [Jagiełło 2013, p. 8]. This means that he used 80 enterprises for each sector, and from among them, 40 were classified as "good" (according to the Polish Accounting Standards as of 31st December 2008) and 40 as "bad" enterprises (according to the Polish Accounting Standards as of 31st December 2009). The models were based on 16 financial ratios (see Table 4) with a high discriminatory power and covering all aspects of the financial situation of an enterprise (calculated as of 31st December 2008). Based on those financial ratios the models were estimated separately for each sector.

Table 4. The financial ratios used in the discriminant model

Ratio	Name	definition
X1	current liquidity	$\frac{\text{current assets}}{\text{short – term liabilities}}$
X2	quick ratio	$\frac{\text{current assets – inventory – prepayments}}{\text{short – term liabilities}}$
X3	cash liquidity	$\frac{\text{cash}}{\text{short – term liabilities}}$
X4	capital share in assets	$\frac{\text{current assets – short\_term liabilities}}{\text{total assets}}$
X5	gross margin	$\frac{\text{gross profit/loss on sale}}{\text{operating expenses}}$
X6	operating profitability of sales	$\frac{\text{profit/loss on operating activities}}{\text{total revenues}}$
X7	operating profitability of assets	$\frac{\text{profit/loss on operating activities}}{\text{total assets}}$
X8	net profitability of equity	$\frac{\text{net profit/loss}}{\text{equity}}$
X9	rotation of assets	$\frac{\text{total revenues}}{\text{total assets}}$
X10	rotation of current assets	$\frac{\text{total revenues}}{\text{current assets}}$
X11	rotation of receivables	$\frac{\text{total revenues}}{\text{receivables}}$
X12	rotation of inventory	$\frac{\text{total revenues}}{\text{inventory}}$
X13	capital ratio	$\frac{\text{equity}}{\text{total liabilities}}$
X14	coverage of short-term liabilities by equity	$\frac{\text{equity}}{\text{short\_term liabilities}}$
X15	coverage of fixed assets by equity	$\frac{\text{equity}}{\text{fixed assets}}$
X16	share of net financial surplus in total liabilities	$\frac{\text{net profit/loss + amortisation + Interest}}{\text{total liabilities}}$

Source: [Jagiełło 2013, pp. 21-22].

The author proposed models based on the logistic regression and the discriminant analysis. The selection of significant ratios was performed based on the discriminant analysis. In the next step, the logistic regression models were performed on the ratios selected in the discriminant analysis.

5.1. The R. Jagiełło logistic regression model for different sectors

For each group (by sector) the author proposed the model based on 3-4 financial ratios. The coefficients from the logistic regression models are presented in Table 5. Different ratios were significant in different models. The most common set of ratios was significant in the model for Construction and Transportation.

Table 5. The coefficients in logistic regression models – the Jagiełło model

Coefficient	Manufacturing	Construction	Trade	Transportation	Services
Intercept	–4.98	–5.94	–8.05	–5.46	–7.59
X2					5.39
X3			8.05		
X4		10.43		3.17	
X5	27.79		14,36		22.26
X8		4.86		10.20	
X9	0,28		1.03		
X10		0.12		0.44	
X11					0.24
X13	5.22	4.72			
X14			1.62		1.12
X16				2.43	
Effectiveness	93.8%	87.5%	86.3%	87.5	90.0%

Source: [Jagiełło 2013, pp. 95 and further].

6. The proposal of the model with the nominal variable SECTOR

As an alternative proposal to the R. Jagiełło model, there is one pooled model for all sectors with and without the nominal variable SECTOR, which should increase the predictive power of the model. The discriminatory power of such models will be compared using the area under curve ROC.

6.1. The sample used for the model estimation

The sample used for the verification of the R. Jagiełło model and the proposal of a new model consisted of 388 observations from the original sample. For each of the five sectors (industry, construction, transportation, trade, services) only 40 “good” and 40 “bad-defaulted” enterprises were randomly selected. An exception was made

Table 6. The sample description

Sector	Default	*N*	*N*-sector
1 − Manufacturing	0 − no	40	80
	1 − yes	40	
2 − Construction	0 − no	40	80
	1 − yes	40	
3 − Transportation	0 − no	40	68
	1 − yes	28	
4 − Trade	0 − no	40	80
	1− yes	40	
5 − Services	0 − no	40	80
	1 − yes	40	

Source: own elaboration.

Table 7. The effectiveness of the R. Jagiełło model, the original and on the sample used in this research

Effectiveness	Manufacturing	Construction	Trade	Transportation	Services
R. Jagiełło Model	93.8%	87.5%	86.3%	87.5	90.0%
Sample	45.0%	41.3%	50.0%	54.4%	45.0%

Source: [Jagiełło 2013, pp. 95 and further]; own elaboration.

for transportation because the sample for this sector was too small and only 28 "defaults" were available.

The effectiveness of the R. Jagiełło model on our sample was very weak, much lower comparing to the effectiveness of the original model proposed by R. Jagiełło (see Table 7). In such a situation, the model based on the ratios used by R. Jagiełło, but on the pooled sample with and without the variable SECTOR was proposed. The logistic regression model was applied.

6.2. Variables selection

The selection of variables was based on their accuracy power (AR). 0.1 was set up as a cut-off value for the discriminatory power. The selection based on the accuracy ratio (AR) includes more ratios in the model comparing to the reference model based on the t-test. The financial ratios below the AR cut-off value (0.1) were eliminated.

The Pearson correlation coefficient was used in the detection of correlation. The correlation coefficients for pairs: (X1 – X2), (X2 – X14), (X4 – X13) are too high (see Table 9). To eliminate the correlation between financial ratios, the variables X2 and X4 were eliminated.

Table 8. The significance of variables – discriminatory power

Ratio	GINI	*p*-value (t^2)
X1	0.215	0.38
X2	0.142	0.425
X3	0.082	0.092
X4	0.214	0.0001
X5	0.22	0.008
X6	0.317	0.299
X7	0.33	0.001
X8	0.071	0.317
X9	0.006	0.889
X10	0.04	0.596
X11	–0.134	0.503
X12	–0.024	0.4716
X13	0.366	<0.001
X14	0.372	0.271
X15	–0.234	0.228
X16	0.361	0.021

Source: own elaboration in SAS 9.4.

Table 9. The correlation matrix

	X1	X2	X4	X5	X6	X13	X14	X16
X1	1.00	0.96	0.16	0.01	0.01	0.15	0.69	–0.26
X2	0.96	1.00	0.12	0.01	0.01	0.13	0.77	–0.12
X4	0.16	0.12	1.00	0.10	0.03	0.92	0.02	–0.00
X5	0.01	0.01	0.10	1.00	–0.00	0.08	0.00	0.03
X6	0.01	0.01	0.03	–0.00	1.00	–0.02	0.00	0.00
X13	0.15	0.13	0.92	0.08	–0.02	1.00	0.08	0.03
X14	0.69	0.77	0.02	0.00	0.00	0.08	1.00	0.31
X16	–0.26	–0.12	–0.00	0.03	0.00	0.03	0.31	1.00

Source: own elaboration in SAS 9.4.

6.3. The logistic regression model without the variable SECTOR

For the variable selection in the final model without the variable SECTOR, the stepwise selection method was used ($\alpha = 0.05$ significance level SLE and SLS). Only two ratios, i.e. X7 and X13, were significant in the model. The results of the model are presented in Table 10.

Table 10. The results for the logistic regression model without the variable SECTOR

Variable	Coefficient	*p*-value	Odds ratio (95% CI)
Intercept	0.3302	0.0207	
x7	–2.7509	0.0001	0.064 (0.016-0.263)
x13	–1.6786	<0.0001	0.187 (0.087-0.399)

Source: own elaboration in SAS 9.4.

The predictive power of the model without SECTOR was at the level of 64.6% of correctly classified enterprises. This model is characterized by a high specificity, the model correctly classifies 74.5% of "good" enterprises. Correctly classified "defaults" amounted to 54.0%.

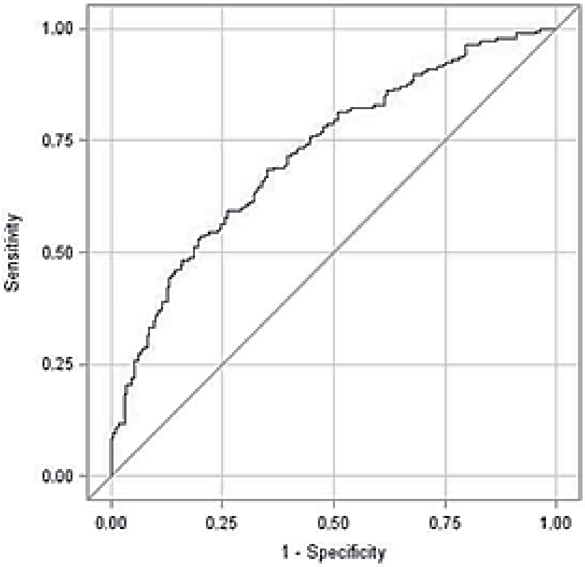

Figure 1. The ROC curve in the logistic regression model without the variable SECTOR

Source: own elaboration in SAS 9.4.

The area under the ROC curve (AUC) should be at least at the level 0.75 – 0.8 to assume the predictive power of the model as satisfactory. For this model, the AUC was a little lower (AUC = 0.7227), which means that the predictive power of the model without the variable SECTOR is rather low.

6.4. The logistic regression model with the variable SECTOR

For the variable selection in the final model with the variable SECTOR, the stepwise selection method was used ($\alpha = 0.05$ significance level SLE and SLS). Only two ratios, i.e. X7 and X13, were significant in the model. The results are presented in Table 11. The variable SECTOR was not significant (*p*-value at the level 0.22). Only

Table 11. The results for the logistic regression model with the variable SECTOR

Variable		Coefficient	*p*-value	Odds ratio (95% CI)
Intercept		0.5026	0.0619	
Sector	1 – Manufacturing	–0.0446	0.8961	0.956 (0.490-1.867)
Sector	2 – Construction	–0.00539	0.9875	0.995 (0.506-1.955)
Sector	3 – Transportation	–0.7583	0.0441	0.468 (0.224-0.980)
Sector	4 – Trade	–0.0233	0.9459	0.977 (0.498-1.916)
x7		–2.9328	<0.0001	0.053(0.013-0.226)
x13		–1.7922	<0.0001	0.167(0.077-0.361)

Source: own elaboration in SAS 9.4.

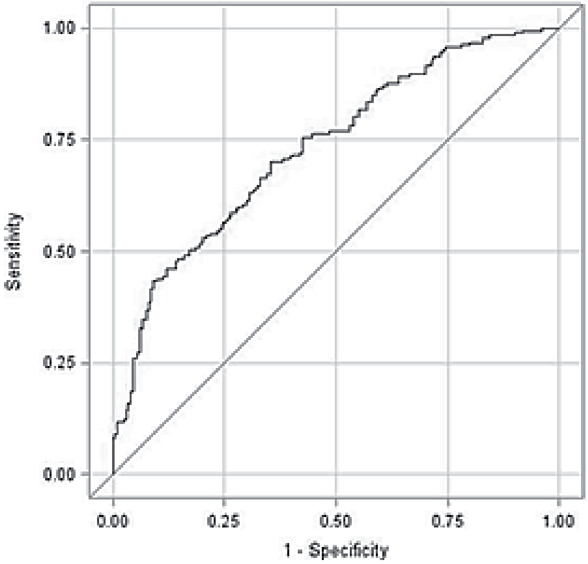

Figure 2. The ROC curve in the logistic regression model with the variable SECTOR

Source: own elaboration in SAS 9.4.

the difference between Transportation and Services (the reference level) is significant at the 0.05 level.

The predictive power of the model with the variable SECTOR amounted for 63.8% of correctly classified enterprises, which is lower comparing to the previous model without the variable SECTOR. There were 72.5% correctly classified "good" enterprises and 54.5% correctly classified "defaults". The predictive power of the model with variable SECTOR is still low AUC = 0.7333 (see Figure 2) but a little higher comparing to the model without this variable (AUC = 0.7227).

6.5. The results for the survival model estimation – the Cox model

The survival model was proposed as an alternative approach. The semiparametric Cox regression model represents the dynamic approach. Contrary to the logistic regression where the dependent variable is binary (0 − good, 1 − *default*), in the survival model the dependent variable is an interval variable measuring the time of the process. The ratios selected in the logistic regression were used as explanatory variables. On the 0.05 significance level only ratios X1, X6 and X16 are significant. On 0.1 significance level additionally the ratios X5 and X7 are significant. The variable SECTOR is not significant (*p*-value 0.26).

Table 12. The results for the Cox survival model

Variable		Coefficient	*p*-value	HR (95% CI)
X1		–0.30150	0.0015	0.740 (0.614-0.891)
X5		–0.000133	0.0724	1.000 (1.000-1.000)
X6		–0.01398	0.0100	0.986 (0.976-0.997)
X7		–0.58178	0.0617	0.559 (0.304-1.029)
X13		–0.14776	0.1594	0.863 (0.702-1.060)
X16		–0.59684	0.0011	0.551 (0.385-0.788)
Sector	1 Manufacturing	0.08032	0.7215	1.084 (0.697-1.685)
Sector	2 Construction	–0.01716	0.9389	0.983 (0.634-1.525)
Sector	3 Transportation	–0.48515	0.0609	0.616 (0.371-1.022)
Sector	4 Trade	–0.09215	0.6840	0.912 (0.585-1.421)

Source: own elaboration in SAS 9.4.

For the assessment of the predictive power and the accuracy of this model the predictor from the Cox model was used.

The predictive power of the model amounted to 61.0% of correctly classified enterprises, which is lower comparing to the logistic regression. There were 63.5% correctly classified “good” enterprises which was much less comparing to logistic regression, but there were 58.3% of correct classifications among ”defaults” which is much more, comparing to logistic regression.

The discriminatory power of the Cox regression model is rather low AUC = 0.705. This value is lower comparing to the logistic regression model with variable SECTOR (AUC = 0.733).

7. The summary and conclusions

The proposed verification of the R. Jagiełło model and an alternative model proposal delivered the following conclusions:

- In the evaluation of the predictive power of models for bankruptcy/defaults prediction the definition of the “event” is crucial. The differences in definition may cause a significant drop of the predictive power of the model.

- Using small samples in models' estimation (in this case 80 entities) causes a low level of accuracy classification of such models. The application of such a model on the independent sample does not confirm the ability of the model to predict the bankruptcy of enterprises.
- Including the nominal variable SECTOR only slightly increases the classification accuracy (in the range of "defaulted" enterprises) comparing to the model estimated on the pooled sample of all enterprises (different sectors in one model).
- An alternative approach was proposed using the dynamic survival model (the semi-parametric Cox regression model). There is an advantage of such an approach in the case of "defaults" classification accuracy. The general predictive power of the survival model was lower comparing to the logistic regression model.

Concluding – the goal of this research was achieved. The effectiveness of models built on the small samples is rather low and the R. Jagiełło model cannot be used in practice. Authors' model proposal was not effective despite some advantages.

Bibliography

Bellovary J., Giacomino D., Akers M., 2007, *A review of bankruptcy prediction studies: 1930 to present*, Journal of Financial Education, vol. 33, Winter.

Blossfeld H.P., Rohwer G., 2002, *Techniques of Event History Modeling. New Approaches to Causal Analysis*, Lawrence Elbaum Associates Publishers, London.

Cox D.R., 1972, *Regression models and life tables*, Journal of the Royal Statistical Society (Series B), no. 34, pp. 187-202.

Jagiełło R., 2013, *Analiza dyskryminacyjna i regresja logistyczna w procesie oceny zdolności kredytowej przedsiębiorstw*, Materiały i Studia NBP, no. 286, Warszawa.

Pociecha J. (ed.), 2014, *Statystyczne metody prognozowania bankructwa w zmieniającej się koniunkturze gospodarczej*, Uniwersytet Ekonomiczny w Krakowie, Kraków.

Sierpińska M., Jachna T., 2004, *Ocena przedsiębiorstwa według standardów światowych*, PWN, Warszawa.

MODELE STATYSTYCZNE W OCENIE RYZYKA NIEWYPŁACALNOŚCI PRZEDSIĘBIORSTW – PRZYKŁAD ZASTOSOWAŃ

Streszczenie: Ryzyko niewypłacalności podmiotów (*default*) jest krytyczne w działalności bankowej. W literaturze przedmiotu opracowano różne modele bazujące na analizie dyskryminacyjnej, regresji logistycznej i technikach *data mining*. W artykule zastosowano regresję logistyczną do weryfikacji skuteczności modelu zaproponowanego przez R. Jagiełłę dla różnych sektorów. Jako alternatywę zaproponowano model regresji logistycznej ze zmienną nominalną SEKTOR na łącznej próbie danych. Oszacowano dynamiczny model przeżycia – model Coxa. Włączenie do modelu zmiennej nominalnej SEKTOR tylko nieznacznie zwiększa moc dyskryminacyjną modelu (w obszarze *default*). Moc dyskryminacyjna modelu Coxa jest niższa, z wyjątkiem klasyfikacji podmiotów w sytuacji *default*, w której wyższa trafność klasyfikacji stanowi przewagę modelu Coxa.

Słowa kluczowe: ryzyko *default*, regresja logistyczna, model Coxa.