The problem of outliers in the research on the financial standing of construction enterprises in Poland

Barbara Pawełek[1], Jadwiga Kostrzewska[2], Artur Lipieta[3]

Abstract

The analysis of an enterprise's financial standing is an important element of company management, while deterioration of the financial standing may result in the enterprise's bankruptcy. Financial indicators are used for the evaluation of enterprises' financial standing. Thus, the data from financial statements is the basis for the examination of the financial position. The evaluation of the quality of the data includes (inter alia) the identification of outliers. The purpose of the article is to present the results of a pilot empirical study regarding the influence of the selection of the method of detecting and eliminating outliers on the effectiveness of the logit model constructed on the basis of samples including or omitting the detected outliers in the scope of the classification of enterprises into the "healthy" ones and the ones at a bankruptcy risk. The pilot research has used the one-dimensional (quantiles) and multi-dimensional (depth function) methods of detecting outliers. For the analysis of changes in the distribution of financial indicators, the sign test and Wilcoxon signed-rank test have been applied. The evaluation of classification effectiveness of the logit model was based on sensitivity and specificity measures. The research covered construction enterprises operating in Poland in 2009.

Keywords: *outliers, financial standing, financial indicator, logit model, classification*

JEL Classification: C250, C530, G330

1. Introduction

The results of the analysis of the financial standing are used (inter alia) in the research regarding the threat of going bankrupt. Financial indicators taken from financial statements are used to evaluate enterprises' financial standing. The assessment of the quality of the financial data includes, for example, the detection of outliers. Studies focusing on the prediction of enterprise bankruptcy contain considerations regarding the issue of outliers. The proposed problem solutions range from ignoring (Spicka, 2013), to replacement or removal of values considered as measurement errors or errors of data introduction (Pociecha et al., 2014), to change or elimination of outliers (De Andrés et al., 2011; Shumway, 2001; Wu et al., 2010). Thus, empirical studies reveal doubts regarding the

[1] Corresponding author: Cracow University of Economics, Department of Statistics, Rakowicka 27, 31-510 Kraków, Poland, e-mail: barbara.pawelek@uek.krakow.pl.

[2] Cracow University of Economics, Department of Statistics, Rakowicka 27, 31-510 Kraków, Poland, e-mail: jadwiga.kostrzewska@uek.krakow.pl.

[3] Cracow University of Economics, Department of Statistics, Rakowicka 27, 31-510 Kraków, Poland, e-mail: artur.lipieta@uek.krakow.pl.

selection of the correct approach to the problem of outliers. Should outliers be detected or not? If so, how should they be detected, and what to do with the knowledge of outliers?

This paper refers to the research on the essence and significance of the problem of outliers in the prediction of enterprise bankruptcy (Tsai and Cheng, 2012). The empirical study is aimed to check the usability of the selected methods for the identification of outliers (one-dimensional and multi-dimensional), procedures for constructing the logit model used for prediction of the risk of bankruptcy and ways of verifying the classification effectiveness of the estimated logit models in the context of the knowledge of outliers.

The article is aimed at presenting the results of a pilot empirical study regarding the influence of the selection of the method of detection and elimination of outliers on the effectiveness of the logit model constructed on the basis of samples including or omitting the detected outliers in the scope of the classification of enterprises as “healthy” ones and the ones at a risk of bankruptcy.

2. Research methodology

The financial data of construction enterprises in Poland for the years 2005-2009 was downloaded from the EMIS Intelligence – Polska website. The data base contains information on 371 objects, including seven bankrupt enterprises. In the paper “healthy” enterprises are also referred to as ‘non-bankrupts’ (NB) while bankrupt enterprises are referred to as ‘bankrupts’ (B). What is a weak point of the analysed database is a small number of bankrupt enterprises, which hindres the creation of a test sample for the logit model. The pilot research covered construction enterprises in Poland in 2009.

The research used 14 financial indicators (table 1), divided into four groups in line with the classification referring to important characteristics of the financial standing of enterprises, such as liquidity (R_{01}–R_{03}), liability (R_{04}–R_{06}), profitability (R_{07}–R_{10}) and productivity (R_{11}–R_{14}).

Pairs of distribution values of financial indicators for successive years of the period of 2005-2009 were compared with the use of non-parametric tests for dependent samples, such as the sign test and the Wilcoxon signed-rank test (Aczel, 2006; Domański and Pruska, 2000). Wilcoxon signed-rank test is stronger than the sign test. When applied for variables measured on an interval scale or a higher scale, its strength is close to that of parametric *t* test for dependent samples. What is significant is that each of the applied tests requires weaker assumptions than their parametric equivalents.

Symbol	Description	Symbol	Description
R_{01}	Current liquidity ratio	R_{08}	Net profitability
R_{02}	Quick liquidity ratio	R_{09}	ROE
R_{03}	Cash Ratio	R_{10}	ROA
R_{04}	Total Debts to Assets	R_{11}	Accounts Receivable Turnover
R_{05}	Debt to Equity	R_{12}	Fixed Asset Turnover
R_{06}	Long-term debt to Equity	R_{13}	Total Asset Turnover
R_{07}	Gross profitability	R_{14}	Operation cost to sales revenues

Table 1. Financial indicators.

The data depth concept is an issue connected with non-parametric robust multi-dimensional statistical analysis, developed as part of the exploratory data analysis. It allows for defining the linear order of multi-dimensional observations with the use of multi-dimensional median, defined as the multi-dimensional centre of the observation set. There are many proposals of functions, called the depth functions, which subordinate to each observation from a given distribution a positive number constituting a measure of its deviation from the centre with regard to the distribution (Kosiorowski, 2012). To identify outliers in the multi-dimensional space, the projection depth based on the normal standardised distribution was used (Kosiorowski, 2008).

The following logit model was used for the prediction of the threat of bankruptcy of construction enterprises in Poland:

$$P\left(y_i = \text{bankrupt } \left|\mathbf{x_i}\right.\right) = \frac{\exp(\mathbf{x_i}\boldsymbol{\beta})}{1+\exp(\mathbf{x_i}\boldsymbol{\beta})} \tag{1}$$

where $\mathbf{x_i}$ – vector of independent variables for *i-th* object, $\boldsymbol{\beta}$ – vector of parameters. To evaluate the effectiveness of the classification of the logit model, the model sensitivity measure (i.e. percentage of bankrupt enterprises classified correctly by the model to the collection of enterprises at risk) and the model specificity measure (i.e. percentage of healthy enterprises correctly classified by the model to the group of enterprises under no threat of bankruptcy) (Pociecha et al., 2014) were used.

3. Analysis of the distribution of financial indicators

The statistical analysis of changes in the distributions of values of particular financial indicators in successive years of the period from 2005 to 2009 was carried out in the group of all enterprises and separately in groups of bankrupts (B) and non-bankrupts (NB).

The distributions of the indicator values of all analysed enterprises were compared on the basis of box plots (with the median and with the mean) and selective descriptive statistics. The character of distributions maintained for particular financial indicators in the successive years is similar but not necessarily the same. The greatest differences occur in outliers, concentration of values around the mean or the median, differentiation degree. Throughout the analysed period a strong positive asymmetry of distributions was observed for the values of the following indicators: current liquidity ratio (R_{01}), quick liquidity ratio (R_{02}), cash ratio (R_{03}), long-term debt to equity (R_{06}), accounts receivable turnover ratio (R_{11}) and fixed asset turnover ratio (R_{12}). The asymmetry of the distributions of these financial indicators is maintained in all years of the analysed period, but its intensity differs. As regards total debts to assets (R_{04}) and operation cost to sales revenues (R_{14}), distributions of values are close to symmetrical throughout the analysed period.

Non-parametric tests (the sign test and the Wilcoxon signed-rank test) were used to evaluate the statistical significance of differences between pairs of distributions of the values of financial indicators of construction enterprises in two successive years. Differences were observed for 31 pairs of distributions, while for 25 pairs no statistically significant differences were observed.

The analysed database for the years 2005-2009 included seven construction enterprises which were declared bankrupt. Because of a small share of bankrupt companies among all analysed enterprises (ca. 1.89%), their impact on the distribution of values of individual financial indicators may not be considerable. An exception is when the values reached by bankrupts are outliers in comparison to the values achieved by healthy enterprises. As a result of the comparison of box plots created on the basis of the groups of non-bankrupts, bankrupts, and jointly non-bankrupts and bankrupts, as well as minimum and maximum values in these groups, the following conclusions were drawn. It was observed that the distributions of R_{04}, R_{05}, R_{06}, R_{07}, R_{08}, R_{09} and R_{14} indicators were slightly changed in tails, i.e. in some years minimum or maximum values were reached by bankrupts. These conclusions correspond to the further observations, since it may be observed that the values of some indicators in the group of bankrupts are different from the values achieved in the group of healthy construction enterprises, for example they are definitely low, as in the case of all three liquidity indicators (current liquidity ratio – R_{01}, quick liquidity ratio – R_{02} and cash ratio – R_{03}) and two performance indicators (accounts receivable turnover – R_{11} and fixed asset turnover – R_{12}). It was also observed that there are untypical values of financial indicators in the group of bankrupts (i.e. outliers against the background of values reached by healthy enterprises), e.g.

in the case of all three debt indicators (total debts to assets – R_{04}, debt to equity – R_{05} and long-term debt to equity – R_{06}) and the three profitability indicators (gross profitability – R_{07}, net profitability – R_{08} and ROE – R_{09}). Such values were observed only for some bankrupts and only in some years. However, it has to be emphasised that each bankrupt enterprise has different values of the considered financial indicators in the analysed period. The impact of individual differences among enterprises may be too strong to define general tendencies.

4. Detection of outliers

For the purpose of detecting outliers, the one-dimensional method (quantiles) and the multi-dimensional methods (projection depth function) were used in the pilot research.

The main objective of the analysis is to identify outlying objects. As regards the one-dimensional analysis, the additional objective is to determine which financial indicators have a huge discriminatory power (Yu et al., 2014). The number of bankrupt enterprises having the values of the given indicator in the range of outliers for healthy enterprises was adopted as the criterion. The higher the criterion value, the greater is the discriminatory power of the given indicator.

The analysis of the distributions of financial indicators demonstrated, for example, that in some cases the values observed for bankrupts are different from the values observed for healthy enterprises. In particular, in case of bankrupts the indicator values are much higher or lower as compared with the typical range of values for healthy enterprises, i.e. they are in the tails of the distribution. That is the reason why the one-dimensional analysis related to the areas determined by quantile $q_{0,1}$ or, separately, quantile $q_{0,9}$. Objects with outliers were identified as follows. For a given financial indicator relative quantiles were determined in the group of healthy enterprises. Objects of strongly higher or strongly lower values than the given quantile were considered to be outlying objects. Next, the number of bankrupts that achieve the values from the determined range for healthy enterprises was verified. If values reached by bankrupts occur in both tails of the distribution for healthy enterprises, the area determined by quantiles $q_{0,05}$ and $q_{0,95}$ was analysed and next the number of bankrupts was redetermined. Table 2 contains a specification of the number of bankrupts of the values of the given indicator from the range of outliers for healthy enterprises. The indicators for which no bankrupts were detected in tails were omitted.

The determined numbers of bankrupts of the values of the given indicator from the range of outliers for healthy enterprises allowed us to identify seven financial indicators of a higher discriminatory power, i.e. R_{03}, R_{06}, R_{07}, R_{08}, R_{09}, R_{10} and R_{11} than the other analysed indicators.

Object	**Indicator**										
	$\boldsymbol{R_{03}}$	$\boldsymbol{R_{04}}$(*)	$\boldsymbol{R_{05}}$(*)	$\boldsymbol{R_{06}}$(*)	$\boldsymbol{R_{07}}$	$\boldsymbol{R_{08}}$	$\boldsymbol{R_{09}}$(**)	$\boldsymbol{R_{10}}$	$\boldsymbol{R_{11}}$	$\boldsymbol{R_{12}}$	$\boldsymbol{R_{13}}$
NB	35	32	36	36	36	36	36	36	36	36	36
B	2	1	1	2	3	3	3	3	3	1	1

Table 2. Maximum numbers of bankrupts with the given indicator from the range of outliers for “healthy” enterprises identified with the use of quantile $q_{0,1}$, quantile $q_{0,9}$ (*) or quantiles: $q_{0,05}$ and $q_{0,95}$ (**).

In the next stage, samples including all bankrupts and healthy enterprises that are not outliers for the given method were created. Methods Q.14 and Q.7 were applied in the one-dimensional analysis based on quantiles and methods D.14 and D.7 were applied in the multi-dimensional analysis based on the depth function. In these methods, the basis for rejecting the outlying objects from among healthy enterprises includes all 14 financial indicators or only seven indicators selected on the basis of the discriminatory power.

As a result of the application of the one-dimensional analysis based on quantiles with the use of the values of all 14 financial indicators (Q.14 method), 190 enterprises were indicated among healthy enterprises as outlying objects. Based on seven selected indicators (Q.7 method), the number of indicated enterprises reached 128. Among them 87 healthy enterprises were identified as outlying objects with the use of both methods (approx. 46% and 68% of the total number of outlying objects according to the given method, respectively).

When using the projection depth function as many as 27 out of 36 outlying objects indicated in the 14-dimensional space (D.14 method) or the 7-dimensional space (D.7 method) were repeated in both data sets (75%). However, 18 enterprises, i.e. 9 enterprises in each group, were considered outlying objects only in one of the analysed multi-dimensional spaces. Thus, one may expect differences in logit models estimated on the basis of the knowledge of outliers obtained from another centre of data set.

What should be remembered when using the projection depth function for the identification of outliers is that the method indicates objects far from the centre of the data set without taking into account the direction of the ‘outlying’ (i.e. the afore-mentioned outlying enterprises may include both enterprises with a very good financial standing and the ones facing serious financial problems).

Outlying objects identified among healthy enterprises were applied for the construction of logit models used for the classification of enterprises into bankrupts and non-bankrupts. The purpose of the analysis presented in the following point is to verify if the knowledge of outliers may be useful for improving the effectiveness of classification of the estimated models.

5. Estimation and evaluation of the effectiveness of the logit model classification

In the first stage of the analysis, two logit models were estimated with the use of the backward stepwise regression method on the basis of the whole database (371 enterprises). The input set of explanatory variables was either 14 financial indicators ($M_{O.14}$ model) or seven selected indicators ($M_{O.7}$ model) of the highest discriminatory power in compliance with the criterion provided in Table 3. The classification effectiveness of the received models was evaluated with the use of sensitivity and specificity measures on a training data set, i.e. a set comprising 371 objects (Table 4).

Model	Explanatory variables
$M_{O.14}$	R_{01} R_{04} R_{06} R_{09} R_{11} R_{13} R_{14}
$M_{O.7}$	R_{07} R_{09} R_{11}
$M_{Q.14}$	R_{06} R_{07} (*without the absolute term*)
$M_{Q.7}$	R_{06} R_{14}
$M_{D.14}$	R_{03} R_{09} R_{14}
$M_{D.7}$	R_{06} R_{09} R_{11} R_{14}
$M_{T.14}$	R_{02} R_{04} R_{11}

Table 3. Explanatory variables in estimated logit models.

In the second stage of the analysis logit models were built, also with the use of the backward stepwise regression method, on the basis of the samples obtained from the input database by means of removing healthy enterprises from the data base, as they were considered outlying objects. The analysis was carried out separately for the results generated with the use of the one-dimensional method (M_Q models) and the multi-dimensional method (M_D models) of outliers identification. The input set of explanatory data contained 14 ($M_{Q.14}$ and $M_{D.14}$ models) or 7 ($M_{Q.7}$ and $M_{D.7}$ models) indicators (Table 3). The evaluation of the classification effectiveness of the estimated models was conducted (also with the use of sensitivity and specificity measures) on the whole data base (option I: 371 objects), on the training sample (option II: 181 objects indicated with the use of *Q*.14 method, 243 objects indicated with the use of *Q*.7 method or 335 objects indicated with the use of *D*.14 and *D*.7 methods) and on the control sample (option III: 197 objects indicated with the use of *Q*.14 method, i.e. 190 healthy enterprises indicated as outlying objects and 7 bankrupts, 135 objects indicated with the use of *Q*.7 method, i.e. 128 healthy enterprises indicated as outlying objects and 7 bankrupts, or 43 objects indicated with the use of *D*.14 and *D*.7 methods, i.e. 36 healthy enterprises indicated as outlying objects and 7 bankrupts) (Table 4).

Model	***N***	**Option 1**		**Option 2**		**Option 3**	
		sensitivity	**specificity**	**sensitivity**	**specificity**	**sensitivity**	**specificity**
$M_{O.14}$	371	0.1428	1.0000	X	X	X	X
$M_{O.7}$	371	0.0000	1.0000	X	X	X	X
$M_{Q.14}$	181	0.4286	0.8956	0.4286	0.9483	0.4286	0.8474
$M_{Q.7}$	243	0.4286	0.9011	0.4286	1.0000	0.4286	0.7188
$M_{D.14}$	335	0.2857	0.9835	0.2857	1.0000	0.2857	0.8333
$M_{D.7}$	335	0.2857	0.9753	0.2857	1.0000	0.2857	0.7500
$M_{T.14}$	197	0.1429	0.9945	0.1429	0.9885	0.1429	1.0000

Table 4. Evaluation of classification effectiveness of estimated logit models.

In the third stage of the analysis, an attempt was made to estimate logit models based on sets of objects, including healthy enterprises, indicated as outlying enterprises, and bankrupt enterprises. A training sample for these models is the sample marked as option III. In both considered multi-dimensional cases (*D*.14 and *D*.7) and in one-dimensional case *Q*.7 the backward stepwise regression method led to logit models containing the absolute term only. In the one-dimensional analysis with the use of *Q*.14 method $M_{T.14}$ model was received (Table 3). The evaluation of the classification effectiveness of this model was carried out as before, i.e. on the whole data base (option I), on the control sample (option II – 181 objects) and on the training sample (option III – 197 objects) (Table 4).

The second and third stages of the analysis constitute an attempt to use the knowledge of outlying objects by classic elimination of objects from the data base. The consideration of the case of the input set of explanatory variables, based on seven indicators, indicated as the ones of a high discriminatory power, is an attempt to use the knowledge of outlying objects in the one-dimensional aspect without a classic elimination of objects from the data base.

On the basis of the results presented in Table 4 a conclusion can be drawn that the highest value of sensitivity measure was observed in case of logit models estimated with the use of the knowledge of outliers, gained on the basis of the one-dimensional method for the identification of outliers ($M_{Q.14}$ and $M_{Q.7}$). As regards the afore-mentioned two models, the model based on seven financial indicators demonstrated a higher value of specificity measure in options I (the whole data base) and II (training sample). In this model the knowledge of outlying objects in the one-dimensional aspect was used in an unconventional way. However, it has to be emphasised that all values of sensitivity measure do not exceed 0.5. Thus, the estimated models are not good classification tools for the analysed group of enterprises.

Conclusion

The undertaken research constituted an attempt to respond to the following questions: Does the selection of a method of detecting outlying objects impact the classification effectiveness of the logit model in the event of analysing the financial standing of construction enterprises in Poland? How to take into account outliers in the prediction of the bankruptcy risk of construction enterprises in Poland?

The paper presents the results of the pilot research carried out based on the financial data of construction enterprises in Poland in 2009. The results suggest that the knowledge of outliers is useful in the prediction of bankruptcy. The calculations for 2009 indicate greater usefulness of the one-dimensional method for the identification of outliers (based on quantiles) than the applied multi-dimensional method (i.e. the projection depth function). They also demonstrate the usability of the unconventional use of the knowledge of outlying objects in the one-dimensional aspect.

The authors plan to carry out similar analyses for the same set of objects for the years 2005-2008. Next, they plan to repeat the research on the data base in which the set of bankrupt enterprises will be increased at the cost of the time span. The following options are also considered: extending the research with the option of taking into account all years together, increasing the set of methods for detecting outliers, taking into consideration other methods for the estimation of parameters in the logit model (including methods robust to outliers (Hauser and Booth, 2011)), taking into account additional measures of effectiveness classification of the logit model.

Acknowledgements

Cracow University of Economics, Faculty of Management, Department of Statistics, Research No. 031/WZ–KS/04/2014/S/4227 and Research WZ–KS 2015.

References

Aczel, A. D. (2000). *Statystyka w zarządzaniu*. Warszawa : Wydawnictwo Naukowe PWN.

De Andrés, J., Sánchez-Lasheras, F., Lorca, P., & De Cos Juez, F. J. (2011). A Hybrid Device of Self Organizing Maps (SOM) and Multivariate Adaptive Regression Splines (MAR) for the Forecasting of Firms' Bankruptcy. *Accounting and Management Information Systems, 10*(3), 351-374.

Domański, Cz., & Pruska, K. (2000). *Nieklasyczne metody statystyczne*. Warszawa: PWE.

Hauser, R. P., & Booth, D. (2011). Predicting Bankruptcy with Robust Logistic Regression. *Journal of Data Science, 9*, 565-584.

Kosiorowski, D. (2008). *Wstęp do wielowymiarowej analizy statystycznej zjawisk ekonomicznych. Kurs z wykorzystaniem środowiska R*. Kraków: Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie.

Kosiorowski, D. (2012). Statystyczne funkcje głębi w odpornej analizie ekonomicznej. *Seria specjalna: Monografie, nr 208*. Kraków: Wydawnictwo Uniwersytetu Ekonomicznego w Krakowie.

Pociecha, J. (ed.), Pawełek, B., Baryła, M., & Augustyn, S. (2014). *Statystyczne metody prognozowania bankructwa w zmieniającej się koniunkturze gospodarczej*. Kraków: Fundacja Uniwersytetu Ekonomicznego w Krakowie.

Shumway, T. (2001). Forecasting Bankruptcy More Accurately: A Simple Hazard Model. *The Journal of Business*, *74*(1), 101-124.

Spicka, J. (2013). The financial condition of the construction companies before bankruptcy. *European Journal of Business and Management*, *5*(23), 160-169.

Tsai, Ch-F., & Cheng, K-Ch. (2012). Simple instance selection for bankruptcy prediction. *Knowledge-Based Systems*, *27*, 333-342.

Wu, Y., Gaunt, C., & Gray, S. (2010). A comparison of alternative bankruptcy prediction models. *Journal of Contemporary Accounting & Economics, 6*, 34-45.

Yu, Q., Miche, Y., Séverin, E., & Lendasse, A. (2014). Bankruptcy prediction using Extreme Learning Machine and financial expertise. *Neurocomputing*, *128*, 296-302.