Estimation of the logistic regression model for company bankruptcy

Jarosław Kaczmarek[*]

Keywords: bankruptcy, bankruptcy prediction, logistic regression model

Słowa kluczowe: upadłość, predykcja upadłości, model regresji logistycznej

Synopsis: Estimated on Poland's largest known sample, Firth's logistic regression model which is used for predicting bankruptcy constitutes a unique and specific model which is highly effective in predicting the level of threat to company bankruptcy when compared to other models, used not only in Poland but also abroad.

Introduction

The standard instruments for predicting the threat to company bankruptcy include discriminatory models and logistic regression models. Compared to the new generation methods – such as neural networks – these are less expensive, more communicative, transparent, and their results are easier to interpret and compare. Empirical studies have demonstrated that there are virtually no differences in the predictive capacity of both classes of models. However, the logistic regression model is more favourable because of the absence of assumptions made in reference to the probabilistic nature of explanatory variables and the more natural interpretation of the assessments of the parameters of the model. Its defect is the more complex process of designating the assessment of the parameters of the model.

Concerning small samples – this is characteristic of the Polish models – one should be particularly careful to make best use of the data, to ensure that conclusions contain minimum systematic error and that parameter assessment uncertainty be measured thoroughly. This requires unbiased estimation and, indirectly, the building of confidence intervals maintaining nominal level of coverage. Bearing in mind the above, and for the purposes of research carried out on the threat to company bankruptcy conducted over a number of years, use has been made of the estimated, fully operational, logistic regression model of Firth. This article presents the findings of the research (the main components of the processing model are considered).

Conditions for construction of the model

In order to meet the requirement that the model, as far as possible, relate to the true conditions under which the researched companies operate, focus has been placed on Polish bankruptcy prediction models. These models undoubtedly provide considerable knowledge about the impact of given variables which describe company standing in terms of probability of bankruptcy, however, a common trait of Polish bankruptcy models involves the application of very limited teaching sets (when pairing "1 to 1" – bankrupt, not bankrupt). Therefore, one should treat these findings with extreme care; this is because their high level of prediction capacity, as indicated by the authors of given models, may be considerably overestimated as a result of using small research samples.

[*] Dr Jarosław Kaczmarek, Cracow University of Economics.

Tab. 1. Dimensions of learning samples of Polish bankruptcy prediction models

Authors	Year	Sample	Authors	Year	Sample
M. Pogodzińska, S. Sojak	1995	10	J. Janek, M. Żuchowski	2003	50
D. Hadasik	1998	44	Hamrol, Czajka, Piechocki	2004	100
D. Wierzba	2000	48	D. Wędzki	2004	80
A. Pogorzelski	2000	48	P. Stępień, T. Strąk	2004	36
K. Michaluk	2000	80	D. Appenzeller, K. Szarzec	2004	68
A. Hołda	2001	80	B. Prusak	2005	80 + 78
S. Sojak, J. Stawicki	2001	58	T. Korol	2005	78
J. Gajdka, T. Stos	2003	40	E. Mączyńska, M. Zawadzki	2006	80
M. Gruszczyński	2003	46			

Source: own study based on: [Prusak, 2005, p. 129–172; Hołda, 2006, p. 153–160; Antonowicz, 2007, p. 32–39; Juszczyk, 2010, p. 713–726].

Numerous research has demonstrated that in small samples parameter assessments of the logistic regression model obtained by classic i.e. popular means – i.e. the highest probability method – are characterised by considerable burden. Furthermore, the classic confidence intervals (based on the large sample theory) rarely reach the nominal level of confidence [Firth, 1993, p. 27–38; Heinze, 2006, p. 4216–4226]. The referred to problems almost completely eliminate the application of Firth's logistic regression model which may be treated as a relatively small modification of the classic logistic regression model. Parameter assessments in this model are barely burdened, which is particularly clear in very small samples, whilst confidence intervals are characterised by better probabilistic properties [Firth, 1993, p. 27–38, Heinze, Schemper, 2002, p. 2409–2419]. Bearing in mind the referred to benefits of Firth's model, it is legitimate to presume that it should become one of the basic tools in the modelling of company bankruptcy [Fijorek, 2011, p. 6–9].

Firth's logistic regression model

In the classic logistic regression model it is accepted that the dependent variable $y_i \in \{0,1\}$ $(i = 1, \ldots, n)$ is subject to Bernoulli's decomposition with success probability of $F(x_i'\theta)$, where function F is the distribuant of logistic distribution in the following form:

$$F(x_i'\theta) = \frac{1}{1 + exp[-x_i'\theta]}$$

where x_i is the p–dimensional vector of explanatory variables, and $\theta \in \mathbb{R}^p$ (containing the intercept) is the p–dimensional vector of structural parameters [Long, 1997, p. 56–68].

In order to estimate the model parameters a credibility function and its natural logarithm are designated; next, calculation is made of the partial derivatives of the credibility function logarithm in relation to the model parameters. The solution of simultaneous equations $U(\theta) = 0$is equivalent to finding the parameter assessments vector of θ_{NW} which maximise the credibility function. The θ_{NW} vector is obtained by means of a defined iterative procedure, in that the $U(\theta)$ function is replaced by modifying it somewhat. One may say that Firth's logistic regression model, despite the fact that it was developed on the basis of classic statistical inference, contains a Bayesian equivalent. It is equivalent to the classic logistic regression model with Jeffreys prior distribution imposed on the parameters [Fijorek, 2011, p. 8–9].

Wald's method is most often used for designating the confidence intervals of the model parameters. However, concerning small samples Wald's confidence intervals rarely reach nominal parameter coverage probabilityθ_j.Confidence intervals designated by means of the profile likelihood method are said to contain better properties, particularly when parameter assessment distribution is far from normal [Stryhn, Christensen, 2003, p. 64–72]. This method, however, is characterised by considerable computational complexity. It involves inverting the likelihood ratio test for the parameter constituting the object of interest. Furthermore, it is possible to designate the profile likelihood confidence intervals for individual elements of the structural parameters vector of the θ_j model with use of the iterative method [Venzon, Moolgavkar, 1988, p. 87–94]. Confidence intervals for the likelihood of success of $F(x_i'\theta)$ stemming from the model may also be attained by means of the profile likelihood method. This, however, requires the use of a refined approach [DiCiccio, Tibshirani, 1991, p. 59–64]. However, in this case one may reformulate the problem; this permits the application of a less complex algorithm [Venzon, Moolgavkar, 1988, p. 87–94].

Continuing the theoretical analysis approach, simulation research was carried out. This led to two conclusions. First of all, the greater the number of cases the closer the coverage probability to the nominal level for both types of confidence intervals (Wald's and the profile likelihood method). Secondly, the confidence intervals of the profile likelihood method attain probability coverage which is closer to nominal levels in comparison to Wald's intervals in almost all of the considered simulation scenarios. Nonetheless, the differences between both types of confidence intervals are not large (these are considerably smaller than those observed in the case of the confidence intervals of both types for individual parameters).

For this reason, one may ultimately say that when very small samples are considered and the probabilistic properties of the applied statistical methods are fulfilling the highest standards, it is recommended to apply profile likelihood method confidence intervals, despite their considerable computational complexity. In the remaining cases one may apply Wald's confidence intervals [Fijorek, 2011, p. 4–7].

Model construction – set of indicators

The first stage involves defining the set of metrics which describe in a synthetic but multi–dimensional manner, company standing and economic and financial results. Indicators are most frequently employed in such cases. In consideration of the number of indicators, there exists the possibility of setting up models which differ in terms of sets of variables and weighted coefficients, but despite this, they demonstrate similar classification capacity. In turn, the number of variables has an impact on the analytical capacity of the model (ability to perform factor analyses).

Tab. 2. Economic and financial indicators used in the construction of the model

Area	Specific indicators
Liquidity	Current ratio (W8), quick ratio (W9), cash flow (W10)
Financing	General financial standing (W2), self–financing (W3), covering of fixed assets with fixed capital (W7), ability to service debt (W21), covering of liabilities through financial surplus (W22), debt payment periods (W23), credit worthiness/debt volume (W24)
Profitability	Sales profitability (W16), sales operating profitability (W17), asset profitability (W18), asset operating profitability (W19), equity profitability (W20)
Debt	General debt (W4), short–term debt (W6), long–term debt (W5)
Productivity/efficiency	Asset productivity (W1), inventory conversion cycle (W11), receivables (W12), liabilities (W13), cash (W14), net working capital cycle (W15)

Source: own work.

In order to construct the threat to bankruptcy prediction model [Kaczmarek, 2011, p. 93–123] use was made of a set of 24 indicators belonging to the following areas: productivity, liquidity, financing, profitability, debt and efficiency (tab. 1). The above selection was made on the basis of analyses, specialist literature research and acquired substantive knowledge.

Apart from the explanatory variables in their basic form, consideration was also given to their non–linear function and interaction of a higher order.

Defining the training set

A commonly used procedure involves the collection of data on bankrupt companies, followed by the matching of these companies to those companies which have not gone bankrupt. Concerning small data sets, matching usually is based around expert knowledge and a thorough analysis of each item observed. This kind of approach, however, is not possible when working on larger data sets, which is the basic characteristic when performing work under the *Early Warning System* [Kaczmarek, 2011, p. 27–28].

When creating the training set an application is made of the standard statistical approach – the collection from the population of a random sample of companies, followed by a description for each company and the class it belongs to (bankrupt or not bankrupt).This approach is virtually unknown when predicting bankruptcy, however, it is widely used in medicine. The fundamentally applied method for matching companies involves the case–control technique. This involves defining a number of key characteristics of ~~the~~ statistical units and the matching of each unit with a distinguishing trait to a unit without such a trait, but which is most similar to it in terms of the variables used for matching. In this manner it was accepted that each bankrupt company would be accompanied by companies which have not gone bankrupt but which are similar in terms of the value of assets and net revenue on sales; matching would take place in consideration of PKD (*Polish Classification of Activities*) compliance and the legal and organisational form of the company. Furthermore, the economic and financial data of these companies would derive from the same year. The "1 to 1" matching approach is most often used, but from a theoretical point of view it is justified to even perform matching on a "1 to 5" basis [Hosmer, Lemeshow, 1989, p. 145–162]; this approach was in fact used, in that each non–bankrupt company received, during the bankruptcy model estimation process, a weighted value equivalent to 1/5.

Appropriate data, which served the purpose of creating a training set, was collected over a two–year research period focusing on generally available company data (15 thousand non–bankrupt companies and about 2 thousand bankrupt companies). Following the elimination of incomplete data and after taking into account the criteria for matching, the final teaching set amounted to 426 bankrupt companies and 1,936 non–bankrupt companies.

Tab. 3. Training set characteristics of the estimated model

Type	Ba	Nba	PKD section	Ba	Nba	Year	Ba	Nba
Total	426	1,936						
Production	207	916	Industrial processing	201	900	1998	2	8
Trade	65	318	Power generation and supply	5	16	1999	19	85
Services	154	702	Water supply and sewage (…)	22	95	2000	41	168
			Construction	12	54	2001	63	277
			Trade	65	318	2002	45	201
			Transport and warehousing	77	378	2003	53	241
			Information and communication	20	87	2004	33	153
			Professional activities	20	85	2005	18	73
						2006	14	69
						2007	45	214
						2008	74	361
						2009	19	86

Source: own work. Comment: Ba – bankrupt company, Nba – non–bankrupt company.

Analyses carried out on available specialist literature indicates that this is one of the largest hitherto drawn up company data sets considered in the context of modelling the level of bankruptcy threat (other deficiencies and limitations of the hitherto applied models are not mentioned at this point).

Analysis of correlation of variables in the model

The calculated values of the correlation coefficients served the purpose of presenting these coefficients in the form of a matrix for all indicators used during the analysis, calculated for all companies included in the analysis. The colour red indicates negative correlations and the colour blue indicates positive correlations (the more intensive the colour the stronger the correlation).

The colour white indicates the lack of a correlative relationship between the pair of indicators.

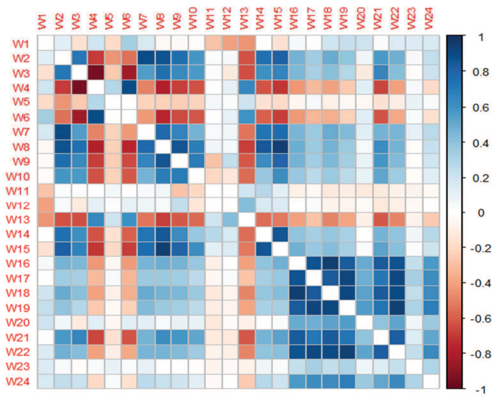


Fig. 1. Matrix of correlation coefficients between given economic and financial indicators
Source: own work.

A detailed analysis of this correlation graph leads to numerous significant conclusions. For example, one may indicate that the profitability indicators, the ability to service debt indicator and the covering of liabilities through financial surplus indicator demonstrate a very strong positive correlation (oscillating around a value of 0.9). The observed strong positive dependencies, however, primarily stem from indicator structure (repetitive elements).

The general debt indicator, the short–term debt indicator and the liabilities conversion cycle indicator are strongly positively correlated with one another and demonstrate significant negative correlation with the majority of the remaining analysed financial indicators. Positive dependencies stem from similar indicator structure, whilst negative correlation with the remaining indicators follows on from the concept of their construction – higher values indicate worse company standing, whilst for the majority of the remaining indictors reverse interpretation is seen as being true.

Indicators relating to asset productivity, long–term debt, inventory conversion cycle and receivables barely demonstrate any correlative connection with any of the remaining indicators under consideration. It appears that the reason for this state of affairs should be considered in the untypical manner in which these indicators may signalise the condition of threat to bankruptcy – company difficulties may be the outcome of both their radically high and radically low values.

On the other hand, concerning the remaining indicators, it is their extremely low or extremely high level which is seen as alarming. In order to explain this mechanism in detail it is necessary to perform a one–dimensional analysis of the economic and financial indicators.

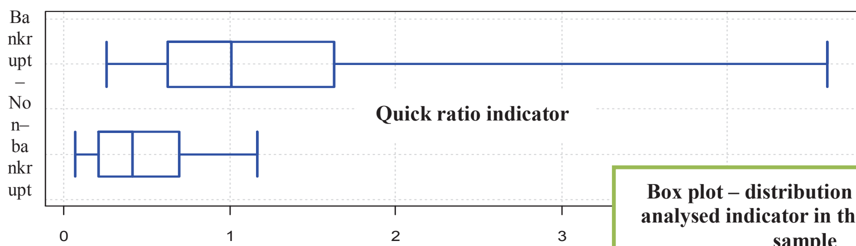
It is also interesting to compare the coefficient values of correlatives indicated separately for the group of bankrupt companies and the group of non–bankrupt companies.

For example:

– the return on equity (ROE) indicator's correlation with other indicators (in particular other profitability indicators) changes significantly in the group of: a) bankrupt companies – it is relatively low, often negative, b) non–bankrupt companies – it is considerably higher and in the majority of cases positive,
– it is not rare that differences in the value of appropriate coefficients of correlation in these groups exceed 0.5. The largest observed difference was 1.02 for the correlation between ROE and ROA (return on assets),
– coefficients of correlation between self–financing indicators, long–term debt and the covering of fixed assets with fixed capital are about 0.3–0.5 higher in the group of bankrupt companies,
– the debt payment period indicator in the group of bankrupt companies is clearly more strongly correlated with the remaining indicators than in the non–bankrupt group.

Estimation of the logistic regression model

The prediction model construction stage was preceded by an analysis of one–dimensional distributions and analysis of the correlation of all 24 potential explanatory variables. With this purpose in mind use was made of box–and–whisker plots and tables were drawn up with chosen distribution percentiles (10, 25, 50, 90), separately for bankrupt companies and non–bankrupt companies.



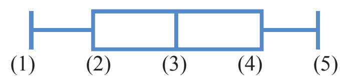
Selected indicator value distribution quantiles

	Bankrupt (428)	Non–bankrupt (1949)
10%	0.095	0.386
25%	0.207	0.629
50%	0.417	1.007
75%	0.695	1.626
90%	0.967	2.983

Content and effectiveness of the classification rule based on the analysed indicator

Indicator < 0.7542 => bankrupt (discovered bankrupts: 81.07%, discovered non–bankrupts: 66.14%)

Box plot – distribution value of the analysed indicator in the researched sample



(1) Level of indicator below which there is a 5% sample.
(2) Level of indicator below which there is a 25% sample.
(3) Level of indicator dividing the sample into two equal sub–samples.
(4) Level of indicator above which there is a 25% sample.
(5) Level of indicator above which there is a 5% sample.

Fig. 2. Diagram depicting the interpretation of results of single–dimension analyses of explanatory variable distributions
Source: own work.

Additionally, a description was given of the predictive capacity of decision rules elaborated for given explanatory variables – decision rules in the following form: if the indicator is larger (smaller) than the threshold value then the company is bankrupt. The threshold value is a figure value which maximises the ability to differentiate between the bankrupt and non–bankrupt, whilst a bankrupt company is prediction stemming from the decision rule when the logical condition is met.

The predictive capacity of decision rules was measured by taking into account their sensitivity (percentage of bankrupts recognised as having gone bankrupt) and specificity (percentage of non– bankrupts recognised as not being bankrupt).

The results of one–dimensional analyses indicate the possibility of distinguishing the following three groups of economic and financial indicators as a criterion for differentiating bankrupt and non–bankrupt companies:

– indicators which permit the best results to be obtained (sensitivity and specificity at a level of 75% or higher, almost 1/3 of researched indicators: overall financial standing, self–financing, current ratio, sales profitability, assets profitability, conversion cycle indicator, cash and net working capital),
– use of the subsequent 11 indicators leads to good results, but each of these has at least one "weakness" (i.e. for one group of researched companies a level of specificity or sensitivity considerably higher than 75% is noted),
– indicators with low levels of specificity and sensitivity (50–60%). These indicators should not be used as the only criterion for assessing a company's economic and financial standing (asset productivity, long–term debt, return on equity, inventory conversion and receivables).

On the basis of the results of the analyses an assessment was next carried out of the parameters of Firth's logistic regression model. In order to define the optimal set of explanatory variables constituting the logistic regression model use was made of the best sub–set method (models with a maximum of 8 explanatory variables were considered). The classification error level method was used as the assessment criterion for matching the model to data.

At this stage the assessment of parameters of the logistic regression model constitutes an important measure in the construction of the bankruptcy prediction model. Of basic importance is the meeting of the commonly binding principles relating to the economics of company functioning – expressed by means of parameter assessment indications. For example, the negative value of the coefficient corresponding to the asset productivity indicator signifies that increase (decrease) translates as decrease (increase) in the level of threat of bankruptcy. It is appropriate, consistent with the pattern of behaviour, that increase in the rate of circulation of capital expressed as growth in sales volume in terms of invested capital is a factor which impacts improvement of company economic and financial standing, and at the same time reduces the risk of threat of bankruptcy.

On the other hand, concerning the value of the short–term debt indicator, or general debt indicator, these should be given a positive assessment – increase in level of debt (which includes additional financial costs of a fixed costs nature), leads to an increase in the risk of engaging in activities which, after exceeding a given level, may lead to the company becoming insolvent, followed by bankruptcy.

In keeping with the approved methodology of procedure, the estimated level of threat to bankruptcy model ($4K_1$) for companies in general is as follows:

$$4K_1 = \frac{1}{1 + exp[-(-0{,}70 - 0{,}42\, Z_1 - 0{,}93\, Z_2 + 0{,}65\, Z_3 - 0{,}73\, Z_4)]}$$

This measure assumes a value of (0,1), in that its higher values indicate a higher probability of bankruptcy (one year prior to this standing) and, in principle, the possibility of bankruptcy, where chance is defined as the likelihood ratio of bankruptcy to the likelihood ratio of non–bankruptcy.

Tab. 4. Parameters of the estimated logistic regression model

Name of indicator	Symbol of indicator	Transformation of indicator	Parameter assessment
Intercept	–	1	– 0.70
Asset productivity indicator	W_1	$Z_1 = (W_1 - 1.89)/1.09$	– 0.42
Self–financing indicator	W_3	$Z_2 = (W_3 - 0.39)/0.31$	– 0.93
Short–term indicator	W_6	$Z_3 = (W_6 - 0.47)/0.27$	+ 0.65
Asset operating profit margin	W_{19}	$Z_4 = (W_{19} - 2.94)/13.46$	– 0.73

Source: own work.

This allows, in quantity terms, the scale of bankruptcy changes to be described dynamically. It is also possible to state whether aspects of bankruptcy will become more or less intense, and also the degree of this change. This also permits a comparison to be made of the degree of threat between various classes and groups of companies. All of the previously constructed models do not display such properties.

Efficiency in detecting threat to bankruptcy

Optimal cut–off points for the level of bankruptcy threat has been designated by means of the ROC (Receiver Operating Characteristic). The ROC curve is a two–dimensional graph which presents sensitivity (percentage of bankrupts recognised as being bankrupt) and 1 – specificity (percentage of non–bankrupts recognised as not being bankrupt), calculated for various values of the cut–off point. As a result, the rule which defines company affiliation to the class of companies in danger of going bankrupt was accepted, providing that the value of the level of bankruptcy threat is greater than 0.5 (high level of bankruptcy threat).

The presented ROC curve demonstrates decision rule behaviour in the event of accepting other cut–off point values. The general rule states that the lower the cut–off point the more the number of detected bankrupt companies; however, this takes place at the cost of recognising an increasing number of non–bankrupt companies as being bankrupt.

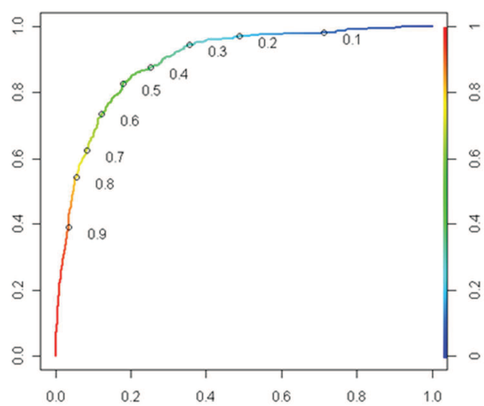


Fig. 3. The ROC curve of the estimated $4K_1$ model
Source: own work.

The predictive capacity of Firth's logistic regression model was calculated with the use of *sensitivity* and *specificity*. Additionally, the quality of models has been measured with the use of the area under the ROC curve (*AUC – Area Under Curve*). The *AUC* measure assumes values ranging [0,1], in that the higher the values the better the given model is assessed.

Tab. 5. Manner of interpreting *AUC* value for the assessed model

AUC value	Model quality	AUC value	Model quality
AUC = 1	excellent model	0.7 ≤ AUC < 0.8	weak model
0.9 ≤ AUC < 1	very good model	AUC = 0.5	random model
0.8 ≤ AUC < 0.9	good model	AUC < 0.5	incorrect model

Source: own work.

Taking into account the values for efficiency of Firth's logistic regression model it is necessary to state that this model is characterised by a high level of capacity to predict the state of threat to company bankruptcy, irrespective of the type of business activities being performed (production, trade, service).

Tab. 6. Measure of efficiency of the estimated logistic regression model

Model	Number of bankrupts	Number of non–bankrupts	Sensitivity	Specificity	AUC
Companies in general	426	1.936	82.4%	82.1%	0.894

Source: own work.

Uncertainty connected with estimating the level of bankruptcy threat

The role of each indicator describing a company's economic and financial standing in shaping the level of bankruptcy threat was checked in terms of scenario analysis. In order to guarantee the comparability of results relating to the given scenario only the value of an individual economic and financial indicator was manipulated, whilst the values of the remaining indicators remained at a determined level.

On the basis of an analysis of the development of curves describing the level of threat to bankruptcy (centrally placed curves) was it possible to assess what the role of a given economic and financial indicator was in the shaping of threat. For example, the role of the self–financing indicator and the short–term debt indicator in the shaping of the level of threat to bankruptcy is considerably greater than the role of the remaining two indicators i.e. the asset productivity indicator and the sales operating profitability indicator. Curves located above and below the central curve constitute, respectively, the upper and lower limits of the 95% confidence interval for the level of threat of bankruptcy. In the figures the continuous line indicates Wald's confidence intervals, whilst the broken line indicates the intervals attained by means of the *profile likelihood* method. The constructed confidence intervals demonstrate a very high level of uncertainty connected with the estimated level of threat of bankruptcy. It must be assumed that in the smaller samples which are so frequent in the Polish threat to bankruptcy models, the uncertainty of estimates will be at an even higher level.

The next conclusion stemming from the analysis of graphs is the fundamental lack of differences between Wald's confidence intervals and intervals obtained by means of the *profile likelihood* method. By this virtue, in order to attain a confidence interval for level of threat to bankruptcy stemming from the proposed Firth's logistic regression model, one should recommend the application of Wald's intervals as methods which are far less complex in terms of computation.

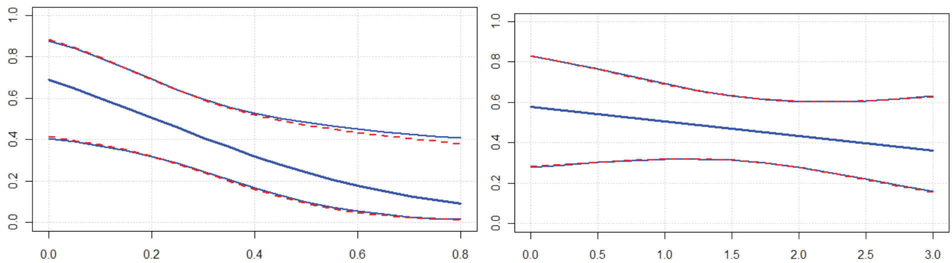


Fig. 4. The level of threat of bankruptcy indicator as a self–financing and asset productivity indicator value
Source: own study based on: [Kaczmarek, 2011, p. 93–123].

Conclusion

The object of considerations involved the use of Firth's logistic regression model as a tool for describing the connection between the multi–dimensional condition of company economic and financial standing indicators, and the level of threat to company bankruptcy. In comparison to the discriminant model, Firth's model is more favourable because of the absence of assumptions in relation to the probabilistic nature of the explanatory variables and the more natural interpretation of the assessments of the parameters of the model; its defect is the more complex process of designating the assessment of the parameters of the model.

Firth's logistic regression model which describes the level of threat of company bankruptcy was elaborated on the basis of a set of 426 bankrupt companies and 1,936 non–bankrupt companies.

In terms of Polish specialist literature this is one of the largest hitherto drawn up company data sets considered in the context of modelling the level of bankruptcy threat.

The estimates made on such a comprehensive bankruptcy prediction model sample constitute a unique and undoubtedly more effective manner of predicting the level of company threat of bankruptcy in relation to previously used models, not only in Poland but also abroad. This model is characterised by a high capacity to predict the state of company threat of bankruptcy, and for this reason it should be widely used in a practical sense, in keeping with its purpose i.e. as an *Early Warning System*.

References

1. Antonowicz P., (2007), *Metody oceny i kondycji ekonomiczno–finansowej przedsiębiorstw,* ODDK, Gdańsk.
2. DiCiccio T., Tibshirani R., (1991), *On the implementation of profile likelihood*, Technical Report 9107, University of Toronto, Department of Statistics.
3. Fijorek K., (2011), *Aproksymacja modelu regresji logistycznej Firtha za pomocą ważenia obserwacji*, materiały z XLVII Konferencji Statystyków, Ekonometryków i Matematyków, Uniwersytet Ekonomiczny w Krakowie, Osieczany.
4. Fijorek K., Fijorek D., (2011), *The effectiveness of best subset variable selection in Firth logistic regression*, "Applied Informatics Methods".
5. Fijorek K., Sokołowski A., (2011), *Separation–resistant and bias–reduced logistic regression*, "Statistica macro".
6. Firth D., (1993), *Bias reduction of maximum likelihood estimates*, "Biometrika", Vol. 80/1993.
7. Heinze G., (2006), *A comparative investigation of methods for logistic regression with separated or nearly separated data*, "Statistics in Medicine", Vol. 25/2006.
8. Heinze G., Schemper M., (2002), *A solution to the problem of separation in logistic regression*, "Statistics in Medicine", Vol. 21/2002.

9. Hołda A., (2006), *Zasada kontynuacji działalności i prognozowanie upadłości w polskich realiach gospodarczych*, Wydawnictwo Akademii Ekonomicznej w Krakowie, Kraków.
10. Hosmer D.W., Lemeshow S., (1989), *Applied Logistic Regression*, Wiley, New York
11. Juszczyk S., (2010), *Prognozowanie upadłości przedsiębiorstw,* „Ekonomista", Vol. 5/2010
12. Kaczmarek J., (2011), *Przedsiębiorstwa w okresie kryzysu. Retrospektywna analiza zmian stopnia zagrożenia*, „Przegląd Organizacji", Vol. 10/2011.
13. Long Scott J., (2005), *Regression models for categorical and limited dependent variables*, SAGE, 1997.
14. Prusak B., (2005), *Nowoczesne metody prognozowania zagrożenia finansowego przedsiębiorstw,* Difin, Warszawa.
15. Stryhn H., Christensen J., (2003), *Confidence intervals by the profile likelihood method, with applications in veterinary epidemiology*, ISVEE X, Chile.
16. *The Report of the assessment of enterprises bankruptcy threat – micro economy component*, Kaczmarek J. (red.), authors: Fijorek K., Kaczmarek J., Kolegowicz K., Krzemiński P., MSAP UEK, June 2011 (duplicated typescript).
17. Venzon D.J., Moolgavkar S.H., (1988), *A method for computing profile–likelihood based confidence intervals*, "Applied Statistics", Vol. 37/1988.

Estymacja modelu regresji logistycznej zagrożenia upadłością przedsiębiorstw

Standardowymi narzędziami w zakresie predykcji zagrożenia upadłością przedsiębiorstw są modele dyskryminacyjne oraz modele regresji logistycznej. Na tle metod nowszej generacji (np. sieci neuronowe) są one mniej kosztowne, bardziej przejrzyste, a ich wyniki łatwiejsze do interpretacji i porównań. Przeprowadzone badania empiryczne wykazują niemal całkowity brak różnic w zdolnościach predykcyjnych obu klas modeli. Za modelem regresji logistycznej przemawia natomiast brak założeń czynionych w odniesieniu do probabilistycznej natury zmiennych objaśniających oraz bardziej naturalna interpretacja ocen parametrów modelu. Wadą jest bardziej złożony proces wyznaczania ocen parametrów modelu.

W małych próbach – co charakteryzuje polskie modele – należy zadbać, aby dane zostały wykorzystane maksymalnie efektywnie, wnioskowanie było obarczone jak najmniejszym błędem systematycznym, a niepewność ocen parametrów była mierzona rzetelnie. Oznacza to postulat estymacji nieobciążonej oraz pośrednio postulat budowania przedziałów ufności utrzymujących nominalny poziom pokrycia. Mając powyższe na względzie, dla celów prowadzonych od kilku lat badań zagrożenia przedsiębiorstw upadłością, wykorzystano estymowany, w pełni funkcjonalny model regresji logistycznej Firtha.

Do budowy modelu predykcji zagrożenia upadłością wykorzystano zbiór 24 wskaźników z obszarów: produktywności, płynności, finansowania, rentowności, zadłużenia oraz sprawności (tab. 1). Ich doboru dokonano na podstawie analiz i studiów literaturowych oraz nabytej wiedzy merytorycznej. Oprócz zmiennych objaśniających w podstawowej formie, rozważone zostały dodatkowo ich nieliniowe funkcje oraz interakcje wyższych rzędów.

Klasyczne podejście w tworzeniu zbioru uczącego, polegające na dobraniu do przedsiębiorstw upadłych metodą ekspercką przedsiębiorstw, które nie upadły, nie może być zastosowane na większych zbiorach danych, co jest podstawową właściwością prac w ramach krajowego Systemu Wczesnego Ostrzegania. Stąd podstawową, zastosowaną metodą dobierania przedsiębiorstw była technika *case–control*. Dobieranie odbywało się na poziomie zgodności działu PKD oraz formy prawno–organizacyjnej. Zastosowano dobieranie „1 do 5", przy czym każde przedsiębiorstwo nieupadłe otrzymało w procesie estymacji modelu upadłości wagę równą 1/5. Zbiór uczący liczył 426 przedsiębiorstw upadłych oraz 1.936 nieupadłych. Analiza dostępnej literatury przedmiotu wskazuje, że jest to jeden z największych jak

dotychczas zbiór danych o przedsiębiorstwach rozważany w kontekście modelowania stopnia zagrożenia upadłością.

W dalszej kolejności obliczone wartości współczynników korelacji zmiennych modelu posłużyły do przedstawienia, w formie rysunku, macierzy tych współczynników dla wszystkich wskaźników wykorzystanych podczas analiz, obliczonych dla wszystkich uwzględnionych w analizie przedsiębiorstw. Szczegółowa analiza tego wykresu korelacji dostarcza wiele istotnych wniosków wykorzystanych w estymacji modelu. Analiza jednowymiarowych rozkładów oraz analiza korelacji wszystkich 24 potencjalnych zmiennych objaśniających została przeprowadzona z wykorzystaniem wykresów typu „ramka–wąsy" oraz tabel z wybranymi percentylami rozkładu (10, 25, 50, 90), osobno dla przedsiębiorstw, które upadły oraz dla tych, które nie upadły (rys. 2). Zdolności predykcyjne reguł decyzyjnych mierzono za pomocą ich czułości (odsetek upadłych uznanych za upadłych) oraz specyficzności (odsetek nieupadłych uznanych za nieupadłych).

Następnie dokonano oceny parametrów modelu regresji logistycznej Firtha. W celu określenia optymalnego zbioru zmiennych objaśniających tworzących model, wykorzystano metodę najlepszego podzbioru (rozważano modele liczące do ośmiu zmiennych objaśniających). Za kryterium oceny dopasowania modelu do danych przyjęto metodę poziomu błędu klasyfikacji.

W wyniku oszacowania modelu zagrożenia upadłością uzyskuje się miarę, która przyjmuje wartości z przedziału (0,1), przy czym wyższe jej wartości wskazują na wyższe prawdopodobieństwo upadłości (na jeden rok przed tym stanem) – a zasadniczo, na szansę upadłości, gdzie szansa jest definiowana jako stosunek prawdopodobieństwa wystąpienia upadłości do prawdopodobieństwa nie wystąpienia upadłości. Pozwala ona w sposób ilościowy opisywać skalę zmian zjawiska upadłości w ujęciu dynamicznym oraz porównywać stopień zagrożenia między różnymi klasami i grupami przedsiębiorstw. Wszystkie dotychczas skonstruowane modele nie posiadają takich właściwości.

Optymalny punkt odcięcia dla stopnia zagrożenia upadłością wyznaczono za pomocą krzywej ROC (*Receiver Operating Characteristic*). Przyjęto regułę definiującą przynależność przedsiębiorstwa do klasy przedsiębiorstw zagrożonych upadłością jeśli wartość stopnia zagrożenia upadłością jest większa od 0,5 (klasa wysokiego zagrożenia upadłością). Zdolności predykcyjne modelu regresji logistycznej Firtha zostały zmierzone za pomocą czułości oraz specyficzności. Dodatkowo jakość modeli zmierzono za pomocą pola pod krzywą ROC (*AUC – Area Under Curve*).

Oszacowany na rozległej próbie model predykcji upadłości jest zatem unikatowym w zakresie wielkości zbioru uczącego będącego podstawą jego estymowania, jak i zastosowania innowacyjnych narzędzi i technik szczegółowych. Charakteryzuje się wysokimi zdolnościami przewidywania stanu zagrożenia przedsiębiorstwa upadłością, niezależnie od rodzaju prowadzonej działalności gospodarczej (produkcyjna, handlowa, usługowa). Daje to podstawy skuteczniejszego przewidywania zagrożenia upadłością przedsiębiorstw w stosunku do modeli dotychczas stosowanych, nie tylko w Polsce. Cechy te przemawiają za jego szerokim zastosowaniem praktycznym, zgodnie z celem dla którego powstał, tj. dla potrzeb Systemu Wczesnego Ostrzegania.