PROBABILITY OF DEFAULT USING THE LOGIT MODEL: THE IMPACT OF EXPLANATORY VARIABLE AND DATA BASE SELECTION

Concepción BARTUAL
Universitat Politècnica de València
E-mail: conbarsa@esp.upv.es

Fernando GARCIA
Universitat Politècnica de València
E-mail: fergarga@esp.upv.es

Francisco GUIJARRO
Universitat Politècnica de València
E-mail: fraguima@esp.upv.es

Agustín ROMERO-CIVERA
Universitat Politècnica de València
E-mail: aromero@cegea.upv.es

Abstract. The Spanish economy is suffering a severe financial crisis which is affecting all Spanish savings banks as well as some major banks. One of the triggers of the crisis is the high companies’ default rate experienced in the last years due to a deficient credit risk management by financial institutions. Credit risk analysis is mainly undertaken using the logit model to calculate the probability of default of the companies. In this work we describe some problems that arise when using this model and that can have a negative impact on the quality of the results obtained.

Keywords: Credit risk, Default, Logit Model

JEL classification:
G21 - Banks; Other Depository Institutions; Micro Finance Institutions; Mortgages
G32 –Financial Risk and Risk Management
G33- Bankruptcy; Liquidation
D81 – Criteria for Decision-Making under Risk and Uncertainty

Introduction

Credit risk analysis is one of the most important tasks to be undertaken by financial institutions. The lack of a correct methodology to calculate the probability of default of the clients may lead to high losses in the banks, create systemic risk, and affect the whole economy of a country. An example of such an event can be seen in the Spanish case, where the economy is suffering because of the high default rates of the credits and the huge losses of the credit institutions since 2010. It has become obvious that the credit risk management undertaken by the Spanish banks in the previous decade has been inadequate. The aim of this paper is to make clear that the models employed to calculate the probability of default, as most models, have some *caveats* that must be considered when making use of the models. If the models are not correctly used, the results obtained can lead to inaccurate understanding of the situation and to bad decisions that can affect a whole country.

The origin of the study of the probability of default is attributed to Beaver and Altman. Using univariate analysis on 30 different ratios, Beaver (1966) showed that the value of certain ratios varied significantly between healthy companies and those in financial difficulties. Altman (1968) used linear discriminant analysis on various financial ratios in a multivariant context to develop insolvency prediction models. This study encouraged other researchers to look for new statistical and econometric techniques that would provide a method of predicting defaults, including the famous Z-score created by de Altman et al. (1977). Without being exhaustive, among the pioneering papers we can include: Jensen (1971), Gupta & Huefner (1972), who used cluster analysis; Vranas (1992) with a linear probability model; the work of Martin (1977), Ohlson (1980), Zavgren (1985), Peel (1987), Keasey et al. (1990) and Westgaard & Wijst (2001) on logit models; Zmijewski (1984), Casey et al. (1986) and Skogsvik (1990) with probit models; Luoma & Laitinen’s study (1991) based on survival analysis and the work of Scapens et al. (1981) on catastrophe theory. In the last two decades, researchers have focused on artificial intelligence and non-parametric methods, including: mathematical programming, expert systems (Elmer y Borowski, (1988); Messier & Hansen, (1988)), machine learning (Frydman et al. (1985)), rough sets (Slowinski & Zopounidis (1995); Dimitras et al. (1999), McKee (2000), neural networks (Wilson & Sharda (1994); Boritz & Kennedy (1995)) and multicriteria decision analysis – MCDA (Andenmatten (1995); Dimitras et al. (1995); Zopounidis & Doumpos (2002)). In many of these studies a high degree of precision was achieved in classifying and predicting business defaults.

It should be pointed out that although there are many methods of estimating the probability of a default, as stated above, at the present time the traditional methods, especially those based on the logit model, are still preferred by professionals in the field. However, one must be aware of certain robustness problems that can arise when using logit, especially in relation to the composition of the sample used to estimate the model. Researchers should pay close attention to three factors: the choice of variables to be used in the model, the influence of the sample on the model results and the cutoff point. Financial variables, especially accounting ratios, are normally used in this type of works and it is not usually advisable to mix absolute and relative variables. Since a choice can be made from a wide range of variables, a factor analysis is normally carried out to reduce their number, keep the degrees of freedom high and avoid multicollinearity problems, while ensuring that the principal financial dimensions (profitability, liquidity, solvency, etc.) are represented. Evidently, it is highly probable that the result of the factor analysis will be influenced by the sample of companies used; if these companies are changed, the variables selected after the factor analysis will also vary.

Whatever the variables used, the logit model finally obtained will depend on the sample on which the model is based. This means that only some of the preselected variables will actually be used in the model, since both the selection and the weighting of the variables will depend on the sample of companies.

Furthermore, whatever cutoff point is chosen, even though it will not modify neither the selected variables nor their weights, this cutoff point will affect the discrimination process and thus also the percentage of correct and incorrect predictions.

In the present study we will use the logit model to analyse credit risk on a sample of Spanish companies using financial information. Throughout the model estimation process we will see how the estimated models will in fact vary as the sample is modified. This point is of great importance for researchers and professionals, since it shows the high degree of dependence of the models on the sample used.

The rest of the paper is structured as follows: Section 2 describes the data base and the selection of the independent variables. In Section 3 the logit model is calculated on two different subsamples and the changes on the models obtained are commented. Finally, Section 4 concludes.

Selection of the Companies in the Data Base and the Independent Variables

In order to analyse the probability of default, the companies in the data base must be separated into two groups: defaulted and not defaulted companies. Identifying which companies have defaulted may be the first challenge. For this study we have considered both the legal situation (being subject to court proceedings) and net negative worth (technical bankruptcy) when classifying the financial situation of the companies in the sample.

The data base for our study consisted of Spanish firms belonging to Group A (agriculture, stock-farming and forestry), Group C (manufacturing, food processing and soft drinks) as classified by the Spanish National Classification of Economic Activities. The firms had total assets between €2m and €50m in 2007, the year with the lowest default rate in the last decade. The information for financial year 2007 was obtained from the SABI data base.

Out of the 622 companies analysed, 49 companies were defined as insolvent.

A set of financial and accounting ratios were selected from the firms’ accounts as independent variables. These ratios belong to different categories such as liquidity, solvency, profitability and economic structure, and usually appear in the models mentioned in the literature.

Table 1: Ratios used in the empirical analysis

ROA	Operating income / Total assets
RAI	Pre-tax profits / Total assets
ORA	Ordinary profits / Total assets
FRA	Financial results / Total assets
ORS	Ordinary results / Sales
EC	Equity /Creditors
C1	Total assets / Creditors
C2	Assets / Creditors – Cash – Temporary investments)
L1	Cash / Short term creditors
L2	Cash / Assets
L3	Current assets / Short term creditors
L4	Operating income / Current liabilities
L5	Creditors / Short term creditors
OIFE	Operating income / Financial expenses
SA	Sales / Total assets
P1	Operating income / Sales
P2	Sales / Personnel expenses
P3	Sales / Financial expenses
P4	Pre-tax profits / Financial expenses
P5	Sales / (Financial expenses + Personnel expenses)
PFE	Profits before tax and interest / Financial expenses
CRSD	(Cash + Realizable assets) / Short term debt
EA	Equity / Total assets

Resource: Authors

When working with a list of interrelated ratios, normally a preliminary step is undertaken to reduce the number of variables and avoid statistical problems. In our study, principal components analysis and the Kaiser criterion were used. Table 2 gives the varimax orthogonal rotation of the factor matrix. Nine factors or groups were extracted and the first element of each group was selected as representative. So, the selected explanatory variables to be used in the models are: RAI, L3, P5, PFE, P1, FRA, L5, C2 and SA.

Table 2: Rotated component matrix

	Component								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
RAI	0.935	0.004	-0.076	0.030	0.062	0.230	-0.044	0.003	-0.092
ORA	0.930	-0.012	-0.056	0.032	0.069	0.224	-0.084	0.023	-0.062
ROA	0.912	-0.025	-0.037	0.047	0.063	-0.183	-0.043	-0.071	-0.135
L4	0.594	-0.137	-0.160	0.068	0.086	-0.401	0.101	0.306	0.208
EA	0.472	0.298	0.092	-0.013	-0.002	0.247	-0.089	-0.235	0.307
L2	0.345	0.283	0.007	0.088	0.008	0.297	0.009	-0.069	0.010
L3	0.055	0.907	0.013	-0.011	-0.009	-0.086	0.060	0.277	0.100
CRSD	0.057	0.905	0.010	-0.009	-0.006	-0.089	0.052	0.281	0.094
L1	-0.036	0.800	0.070	0.025	0.051	0.196	0.044	-0.216	-0.059
C1	-0.069	0.797	0.241	-0.052	-0.014	0.194	-0.099	-0.322	-0.093

Table 2: Rotated component matrix

P5	-0.058	0.090	0.987	-0.043	-0.034	-0.013	0.021	-0.026	0.024
P2	-0.035	0.134	0.844	-0.125	-0.016	0.130	-0.040	-0.150	-0.053
P3	-0.066	0.003	0.817	0.075	-0.044	-0.186	0.088	0.136	0.109
PFE	0.066	-0.008	-0.056	0.992	0.047	-0.010	-0.007	0.020	-0.007
OIFE	0.058	-0.002	-0.026	0.992	0.038	-0.003	-0.012	0.003	-0.015
P1	0.086	-0.024	-0.021	0.042	0.991	-0.022	0.000	0.022	0.015
ORS	0.091	0.047	-0.063	0.043	0.989	0.026	-0.011	0.005	0.003
FRA	0.193	0.059	-0.090	-0.028	0.009	0.860	-0.009	0.148	0.070
L5	-0.150	0.062	0.068	-0.016	0.000	0.008	0.739	0.064	0.079
EC	0.050	-0.017	-0.025	0.003	-0.014	-0.036	0.733	-0.089	-0.128
C2	-0.041	0.061	-0.006	0.013	0.019	0.103	-0.046	0.813	-0.106
SA	0.156	-0.100	-0.181	0.060	-0.084	-0.173	-0.297	0.057	-0.656
P4	-0.012	-0.032	-0.051	0.018	-0.034	-0.064	-0.203	-0.043	0.651

Resource: Authors

Using the Logit Model to predict Business Failures with different samples

Once the variables to be introduced in the model are selected, the logit model can be applied on the sample. The logit technique provides a linear combination of independent variables that makes it possible to estimate the likelihood of a firm belonging to either of two previously defined groups (not default/ default). Each firm can only belong to one group. The model calculates the probability “p” of the firm belonging to the insolvent subpopulation by expression (1):

$$p=\frac{1}{1+e^{-\left(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k\right)}} \qquad (1)$$

Xi being the selected ratios and β the estimated coefficients for each of the ratios used. If the probability is equal to or greater than 0.5, the firm is assigned to the group of not defaulted companies. If not, it is placed in the group of defaulted firms.

To obtain the prediction model to calculate the probability of default, two different analyses were carried out by logistic regression. First all the companies in the sample were used. Second, a balanced sample of defaulted and not defaulted companies was employed. When calculating the model, different cutoff points were considered. This is due to the fact that around 8% of the firms in the original sample were insolvent (49 out of 622), so different initial probabilities of belonging to one group or the other needed to be taken into account.

The Forward method (a new variable is introduced in each step) and the Wald statistic were used in all the models to select the subset of variables to be included in the model, being statistically significant.

Table 3 shows the model estimated by logit regression for the total sample of 622 firms from the Spanish food and agriculture sector using financial information as of 2007.

Table 3: Summary of the model of the complete sample (622 firms)

$p=\frac{1}{1+e^{-(-2,941-12,007RAI-19,496FRA+0,020L3)}}$				
Variable	**Coefficient β**	**Wald statistic**	**Significance**	**Exp (β)**
RAI	- 12.007	28.860	0.000	0.000
FRA	- 19.496	4.806	0.028	0.000
L3	0.020	6.891	0.009	1.021
Constant	- 2.941	153.805	0.000	0,053
0.5 cutoff point: % of correctly classified cases: 92.80% Non insolvent firms: 99.50% Insolvent firms: 14.30%				
0.2 cutoff point: % of correctly classified cases: 91.80% Non insolvent firms: 96.30% Insolvent firms: 38.80%				

Resource: Authors

As can be seen on Table 3, the calculation was repeated, changing only the cutoff point from 0.5 to 0.2. In this case, when the model assigns a value greater than 0.2 the firm is classified as insolvent. Another result to underlined is that the probability of being in the first group (non insolvent firms) is reduced, but on the other hand the probability of correctly predicting insolvency increases. In both cases, the models do not correctly detect most of the insolvent firms (14.30% and 38.80% for a cutoff point of 0.5 and 0.2 respectively).

The same analysis is repeated again, with a balanced sample, including an equal number of insolvent and non-insolvent firms. To do this analysis, as the sample of companies is different, new independent variables are selected. Using principal components analysis and the Kaiser criterion again, eight variables are selected: CRSD, EA, P1, PFE, RRF, L1, L5 and P4. The results for a 0.5 cutoff point is shown in Table 4. The new model has improved greatly the capacity for identifying the insolvent firms, up to 92.70%.

Table 4: Summary of balanced simple model (82 firms: 41 insolvent and 41 non insolvent)

$p=\frac{1}{1+e^{-(1,368-6,081EA)}}$				
Variable	**Coefficient β**	**Wald statistic**	**Significance**	**Exp (β)**
EA	- 6.081	19.100	0.000	0.002
Constant	1.368	12.119	0.000	3.926
0.5 cutoff point: % of correctly classified cases: 89.00% Non insolvent firms: 85.40% Insolvent firms: 92.70%				

Resource: Authors

The model obtained only includes one independent variable, equity over total assets (EA).

It is noteworthy to observe that the selected variables in the two models are different. The variable EA was not even selected as independent variable after the principal components analysis when applied on the sample of 622 firms. Nevertheless, the second model, applied on a reduced sample of selected firms, can predict better than the first one.

Conclusions

In recent years the need for banking institutions to undertake a correct risk evaluation has become evident. Among the most important risks to consider, credit risk appears in a preferent place. A mistaken credit risk analysis can have a very negative impact on the balance sheets of the financial institutions and may lead to big economic problems. An example of this can be recognized in the case of the present Spanish economic crisis, which is affecting the whole Euro area. As the number of defaulted credits is increasing, the number of bankrupt banks increases as well, creating a wave that affects the whole economy and the welfare of the citizens.

Credit risk has been a subject of study for many decades. There exist many different models to calculate the probability of default of the companies, such as those based on artificial intelligence, mathematical programming, expert systems, neuronal networks etc. The most widespread method nevertheless remains the analysis by the logit regression model, wich is used, for example, by the most important rating agencies. Furthermore, this methodology is used as a benchmark in many of the studies in the literature. The use of the logit model is not completely free of difficulties, specially the correct selection of the explanatory variables and the appropriate sample to estimate the model. This issues should be borne in mind by researchers and investors who very often do not give it the attention it deserves. This lack of attention can lead to mistakes when interpreting the outcomes of the models, which results in bad investment decisions.

The present paper describes the credit risk analysis of a number of Spanish business companies by means of the logit model. After carrying out a factor analysis to select the explanatory variables, the logit model was estimated by Wald’s forward method on two different samples of business companies. The first sample contained the entire population of selected companies and the second a balanced sample made up of one half insolvent and one half solvent companies.

The main conclusion that can be drawn from this work is the influence that the researcher can have on the models obtained. The researcher must pay special attention when selecting the database, as this database will conditionate all the results, such as the selection of the independent variables, the cutoff point or the final model obtained. Or, in other words, the researcher can modify the results just by changing the database, the sample to be used, the way of selecting the independent variables or the cutoff point.

Together with these problems, there are other important issues researchers must be aware of, as the quality and reliability of the data, and the survival bias.

References

Altman, E.I. (1968). Financial ratios, dicriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23(4): 589-609.

Altman, E.I., Hadelman, R.G., and Narayanan, P. (1977). Zeta analysis: A new model to identify bankruptcy risk of corporations. *Journal of Banking and Finance*, 1(1): 29-54.

Andenmatten, A. (1995). *Evaluación du risque de défaillance des emetteurs d'obligations: Une approche par l'aide multicretère á la décision*. Presses Polytechniques et Univertitaires Romandes, Lausanne.

Beaver, W.H. (1966). Financial ratios as predictors of failure. *Journal of Accounting Research*, 4: 71-111.

Boritz, J.E., and Kennedey, D.B. (1995). Effectiveness of neural network types for prediction of business failure. *Expert Systems with Applications*, 9(4): 503-512.

Casey, M., McGee, V., and Stinkey, C. (1986). Discriminating between reorganized and liquidated firms in bankruptcy. *The Accounting Review*, 61(2): 249-262.

Dimitras, A.I., Zopounidis, C., and Hurson, C. (1995). A multicriteria decision aid method for the assessment of business failure risk. *Foundations of Computing and Decision Sciences*, 20(2): 99-112.

Dimitras, A.I., Slowinski, R., Susmaga, R., and Zopounidis, C. (1999). Business failure prediction using rough sets. *European Journal of Operational Research*, 114(2): 263-280.

Elmer, P.J., and Borowski, D.M. (1988). An expert system approach to financial analysis: The case of S&L bankruptcy. *Financial Management*, 17(3): 66-76.

Frydman, H., Altman, E.I., and Kao, D.L. (1985). Introducing recursive partitioning for financial classification: The case of financial distress. *The Journal of Finance*, 40(1): 269-291.

Gupta, M.C., and Huefner, R.J. (1972). A cluster analysis study of financial ratios and industry characteristics. *Journal of Accounting Research*, 10(1) Spring: 77-95.

Jensen, R.E. (1971). A cluster analysis study of financial performance of selected firms. *The Accounting Review*, 16(1) January: 35-56.

Keasey, K., Mcguinnes, P., and Short, H. (1990). Multilogit approach to predicting corporate failure: further analysis and the issue of signal consistency. *Omega*, 18(1): 85-94.

Luoma, M., and Laitinen, E.K. (1991). Survival analysis as a tool for company failure prediction. *Omega*, 19(6): 673-678.

Martin, D. (1977). Early warning of bank failure: a logit regression approach. *Journal of Banking and Finance*, 1(3): 249-276.

McKee, T. (2000). Developing a Bankruptcy prediction model via rough sets theory. *International Journal of Intelligent systems in accounting, Finance and Management*, 9: 159-173.

Messier, W.F., and Hansen, J.V. (1988). Inducing rules for expert system development: An example using default and bankruptcy data. *Management Science*, 34(12): 1403-1415.

Ohlson, J.A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 18(1): 109-131.

Peel, M.J. (1987). Timeliness of private company reports predicting corporate failure. *Investment Analyst*, 83(January): 23-27.

Scapens, R.W., Ryan, R.J., and Flecher, L. (1981). Explaining corporate failure: a catastrophe theory approach. *Journal of Business Finance and Accounting*, 8(1): 1-26.

Skogsvik, R. (1990). Current cost accounting ratios as predictors of business failures: the Swedish case. *Journal of Business Finance and Accounting*, 17(1): 137-160.

Slowinski, R., and Zopounidis, C. (1995). Application of the rough set approach to evaluation of bankruptcy risk. International *Journal of Intelligent Systems in Accounting, Finance and Management*, 4: 24-41.

Vranas, A.S. (1992). The significance of financial characteristics in predicting business failure: an analysis in the Greek context. *Foundations of Computing and Decision Sciences*, 17(4): 257-275.

Westgaard, S., and Wijst, N. (2001). Default probabilities in a corporate bank portfolio: a logistic model approach. *European Journal of Operational Research*, 135(1): 338-349.

Wilson, R.L., and Sharda, R. (1994). Bankruptcy prediction using neuronal networks. *Decision Support Systems*, 11(5): 545-557.

Zavgren, C.V. (1985). Assessing the vulnerability to failure of American industrial firms. A logistic analysis. *Journal of Business Finance and Accounting*, 12(1): 19-45.

Zmijewski, M. (1984). Methodological issues related to the estimation of financial distress prediction models. Studies on Current Econometric Issues in Accounting Research. *Journal of Accounting Research*, 22: 59-86.

Zopounidis, C., and Doumpos, M. (2002). Multicriteria classification and sorting methods: A literature review. *European Journal of Operational Research*, 138(2): 229-246.