Modelling corporate bond rating with the use of market–based indicators

Martina Novotná[1]

Abstract. The paper presents an estimation of corporate bond rating models based on both financial and market company indicators. The analysis will be carried out for companies in the oil and gas industry having a rating assessment from Moody's rating agency. The paper aims to offer a more detailed understanding of the relationship between company market indicators and bond rating classification. Bond rating models will be estimated by multivariate statistical methods such as discriminant analysis and logistic regression. The contribution of the paper is to identify variables with a significant impact on corporate bond rating in the selected industry. The models being derived allow classifying bond rating of companies with relatively high accuracy, even when a limited set of input variables is considered. The practical use of models lies in the area of management decision process and managing credit risk.

Keywords: Discriminant analysis, estimation, logistic regression, rating model.

JEL Classification: C35
AMS Classification: 62H12

1 Introduction

The aim of this study is to examine and quantify relationships among rating and other relevant data. The primary question is whether financial and market variables affect bond rating. If the answer is positive, the next question is what the nature of their relationship is. The study is based on cross-sectional data of a variety of companies from oil and gas industry, mostly from the United States. The whole sample covers 155 companies with Moody's rating; for the purposes of validation, it was split into two sub-samples. Experimental sample (approximately 75 %) will be used for model estimation and the remaining part (test sample) will be used for validation of models. Two methods will be used to estimate bond rating models, multinomial logistic regression and multivariate discriminant analysis. The next paragraph describes the methodology; results and classification ability of models will be assessed in the following chapters of this paper.

2 Overview of the methodology

This chapter is focused on a brief overview of two methods that will be used to estimate bond rating models, discriminant analysis and logistic regression analysis. The latter method became one of the most used methods to estimate bond rating or default prediction, see for example Altman, Sabato and Wilson [2], Waagepetersen [9], or Westgaard and Wijst [10]. An alternative and traditional approach to predict bond ratings is discriminant analysis introduced for example by Pinches and Mingo [6], Ang and Patel [3], or Altman and Eisenbeis [1].

Discriminant analysis

Discriminant analysis is a common statistical method used for separation of groups, and hence a suitable method for bond rating modelling. Discriminant functions are linear combinations of variables that best separate groups, for example the k groups of multivariate observations. In the following part of this subchapter, the explanations and definitions were taken from Rencher (2002, p. 277 – 286) [7].

For k groups with n_i observations in the ith group, we transform each observation vector y_{ij} to obtain $z_{ij} = a'y_{ij}$, i = 1, 2,..., k; j = 1, 2,...,ni, and find the means $\bar{z}_i = a'\bar{y}_i$, where $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$. We seek the vector a that maximally separates $\bar{z}_1, \bar{z}_2, \dots \bar{z}_k$. The separation criterion among $\bar{z}_1, \bar{z}_2 \dots, \bar{z}_k$ can be expressed in term of matrices,

$$\lambda = \frac{a'Ha}{a'Ea} \tag{1}$$

[1] Technical University of Ostrava, Department of Finance, Sokolská třída 33, Ostrava, martina.novotna@vsb.cz.

where matrix H has a between sum of squares on the diagonal for each of the p variables, and matrix E has a within sum of squares for each variable on the diagonal. Another expression of the separation criterion is

$$\lambda = \frac{\mathrm{SSH}\ (z)}{\mathrm{SSE}\ (z)} \tag{2}$$

where SSH (z) and SSE (z) are the between and within sums of squares for z. The main task of the discriminant analysis is to find a set of weights (a values) for the outcome variables to determine a linear composite:

$$\mathrm{Z} = \mathrm{a}_1\mathrm{Y}_1 + \mathrm{a}_2\mathrm{Y}_2 + \cdots + \mathrm{a}_\mathrm{p}\mathrm{Y}_\mathrm{p} \tag{3}$$

so that the ratio (2) is maximized. The discriminant analysis follows by assessing the relative contribution of the y's to separation of several groups and testing the significance of a subset of the discriminant function coefficients. The discriminant criterion (1) is maximized by λ_1, the largest eigenvalue of $E^{-1}H$; the remaining eigenvalues correspond to other discriminant dimensions. The test of significance is usually based on the Wilks' lambda, Λ, the most widely used criterion. The test statistic at the mth step is

$$\Lambda_{\mathrm{m}} = \prod_{\mathrm{i=m}}^{\mathrm{s}} \frac{1}{1+\lambda_\mathrm{i}}, \tag{4}$$

which is distributed as $\Lambda_{\mathrm{p-m+1,k-m,N-m+1}}$. The statistic,

$$\mathrm{V_m} = -\left[\mathrm{N} - 1 - \tfrac{1}{2}(\mathrm{p}+\mathrm{k})\right]\ln\Lambda_\mathrm{m} = \left[\mathrm{N} - 1 - \tfrac{1}{2}(\mathrm{p}+\mathrm{k})\right]\sum_{\mathrm{i=m}}^{\mathrm{s}} \ln(1+\lambda_\mathrm{i}), \tag{5}$$

has an approximate χ^2-distribution with (p-m+1)(k-m) degrees of freedom.

Logistic regression analysis

Hosmer and Lemeshow (2000, p. 31) [5] define the multiple logistic regression model as follows. A collection of p independent variables is denoted by the vector $\mathbf{x'} = (x_1, x_2, \dots, x_p)$, assuming that at each of these variables is at least interval scale. The conditional probability that the outcome is present is denoted by $P(Y=1|\mathbf{x}) = \pi(\mathbf{x})$. Then, the logit of the multiple logistic regression model is given by the equation,

$$g(\mathbf{x}) = \beta_0 + \beta_1.x_1 + \beta_2.x_2 + \cdots + \beta_p.x_p \tag{6}$$

and the logistic regression model is expressed by the following formula,

$$\pi(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1+e^{g(\mathbf{x})}}. \tag{7}$$

Based on De Laurentis (2010, p. 54 – 55) [4], the g (.) function (6) is known as a link function, which links variables xj and their coefficients βj with the expected value E(Yi) = πi of the ith observation of Y. The link function can be defined as the logarithm of the ratio between the probability of event (e.g. default) and the probability of non-event (e.g. non-default). This ratio is known as "odds" and can be formulated as follows:

$$g(\mathbf{x}) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1.x_1 + \beta_2.x_2 + \cdots + \beta_p.x_p\ . \tag{8}$$

The logit function associates the expected value of the dependent variable to a linear combination of the independent variables. The relationship between independent variables and the probability of default π is nonlinear, while the relationship between logit (π)and independent variables is linear.

Consider we have a sample of n independent observations $(\mathbf{x_i}, y_i)$, i=1,2,...*n*. Fitting the model requires to estimate vector $\boldsymbol{\beta}' = (\beta_0, \beta_1, \dots, \beta_p)$ by the maximum likelihood method. The likelihood function can be described by the following formula, according to Hosmer and Lemeshow (2000, p. 8) [5]:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)^{1-y_i}], \tag{9}$$

where $\pi(\mathbf{x})$ is defined as (7).

Assume $\widehat{\boldsymbol{\beta}}$ is the solution to the likelihood equations, then the fitted values for the multiple regression model are $\hat{\pi}(\mathbf{x_i})$, the value of the expression (8) computed using $\widehat{\boldsymbol{\beta}}$ and $\mathbf{x_i}$.

The multinomial logistic regression allows predicting membership of more than two categories. In this case, it breaks the outcome variable down into series of comparisons between two categories. In the analysis below, bond rating is a dependent variable, which has four possible outcomes. The existence of four categories requires three logit functions and determination of the baseline category, which is then compared with other logits.

3 Sample description

Companies with Moody's rating assessment have been considered in this study, the relevant data come from Moody's official websites[2], companies' annual reports and Yahoo! Finance websites[3] of business finance, stock market, quotes and news. After checking the data and adjustment for companies without all data available, the final sample consists of 155 companies. For the reasons of calculations[4], original rating categories have been re-coded as presented in the Table 1. The first three highest categories have been merged together because of a small number of representative companies, which could negatively affect results and stability of models.

Rating category	Rating code	Number of cases	Marginal percentage
Aaa, Aa, A	4	21	13.5 %
Baa	3	59	38.1 %
Ba	2	30	19.4 %
B	1	45	29.0 %
Total		**155**	**100%**

Table 1 Sample structure

The selection of independent variables should be thoroughly considered, because the set of input variables can substantially affect results, specifically predictive ability and stability of final models. The analysts usually stand on their previous results, experience and other research studies. Basically, most models are estimated based on financial statements of companies. Many studies prove that relatively simple rating models containing basic financial indicators provide good classification ability and can be used as a tool to assign a rating classification.

There are many possible **financial indicators** that can be used in the analysis. The selected indicators should reflect profitability, activity, liquidity and capital structure of companies and all of them should have a relationship with rating. To use some variables in the analysis, main assumptions should be met. First, the variables should have a normal distribution; secondly, multicolinearity should be avoided. In this study, the following financial variables are considered initially:

1. Total assets (*TA*);
2. Equity to total assets ratio (*Equity_to_TA*);
3. Long term debt to total assets ratio (*LTD_to_TA*);
4. Short term debt to total assets ratio (*STD_to_TA*);
5. Return on assets (*ROA*);
6. Return on equity (*ROE*);
7. Return on capital employed (*ROCE*);
8. Interest coverage (*Int_cov*);
9. Current ratio (*Curr_ratio*);
10. Total assets days outstanding (*Days_TA*).

The relationship between each variable and rating should have an economic rationale. For example, we can assume that the higher the size of total assets, the higher the protection of company's creditors, and the higher the rating category. Some variables had to be transformed to approach a normal distribution, such as TA (*LogTA*), Int_cov (*LogInt_cov*), Curr_ratio (*LogCurr_ratio*), Days_TA (*LogDays_TA*).

The main task of this study is to investigate the relationship among rating and selected **market-based variables** such as *beta, earnings per share, enterprise value and market capitalisation*. The paper should answer the question if these market indicators are related to rating. If so, the next step would be to investigate this relationship and use the market indicators to estimate bond rating models. To approach a normal distribution, some of these variables have been transformed (*LogMarketCap*, *LogEV*, *LogBeta*).

4 Bond rating models

Discriminant analysis (DA) and multinomial logistic regression (MLR) will be carried out to identify variables most relevant to rating classification. Two approaches will be used, the method in which all independent variables are included in the model (full), and stepwise method (step), which aims to include only the most significant variables in the model.

[2] http://www.moodys.com/
[3] http://finance.yahoo.com/
[4] PASW Statistics 18

4.1 Estimation of models

First, bond rating models will be estimated from financial data only. Then, only market-based data will be used and finally, results will be compared and a combination of both previous approaches will be applied.

Estimation of models with financial variables

The original set of independent variables was modified and some financial variables (3, 7, 10) were removed for the reasons of high correlations with other variables. The results (Table 2) show that by using only one financial variable (LogInt_cov), it is able to achieve similar classification ability as in the case of a model with seven variables. Overall, multinomial logistic regression provides better classification ability than discriminant analysis.

Model	Approach	Number of input variables	Number of variables in the model	Variables included	Classification ability
(A)	DA Full	7	7	All	46.8 %
(B)	DA Step	7	1	LogInt_cov	44.4 %
(C)	MLR Full	7	7	All	56.8 %
(D)	MLR Step	7	3	Equity_TA, LogInt_cov, LogCurr_ratio	52.3 %

Table 2 Models with financial variables

Estimation of models with market-based variables

Analogically to the previous case, both discriminant analysis and multinomial logistic regression were used to estimate the models and find the most significant indicators for classification. The results (Table 3) show that considering companies' market data only, model with much better classification ability can be obtained. The most relevant variables are EPS and LogEV.

Model	Approach	Number of input variables	Number of variables in the model	Variables included	Classification ability
(E)	DA Full	4	4	All	64.3 %
(F)	DA Step	4	2	EPS, LogEV	62.7 %
(G)	MLR Full	4	4	All	79.7 %
(H)	MLR Step	4	1	LogEV	70.9%

Table 3 Models with market-based variables

Combination of financial and market-based variables

When all the independent variables enter the analysis, the overall classification ability gently rises, especially in the case of MLR. By using all 11 variables, classification ability of 89.6 % can be achieved. By applying stepwise methods, the final models contain only two indicators, LogCurr_ratio and LogMarketCap (Table 4).

Model	Approach	Number of input variables	Number of variables in the model	Variables included	Classification ability
(I)	DA Full	11	11	All	62.7 %
(J)	DA Step	11	2	LogCurr_ratio LogMarket_Cap	66.7 %
(K)	MLR Full	11	11	All	89.6 %
(L)	MLR Step	11	2	LogCurr_ratio LogMarket_Cap	76.1 %

Table 4 Combination of financial and market-based variables

Modifications and adjustments

Based on the results above, it is evident that some variables contribute to classification more than the others. The final models would stand on the previous results and use only four predictors with the most significant discriminating power on rating, such as *LogInt_cov, EPS, LogEV* and *LogMarketCap*. Classification ability of the adjusted models is in Table 5.

Model	Approach	Number of input variables	Number of variables in the model	Variables included	Classification ability
(M)	DA Full	4	4	All	64.3 %
(N)	MLR Full	4	4	All	76.3 %

Table 5 Modification of models

The overall results suggest that market indicators contribute to the discrimination more than financial ratios. By adding market data to the original set of financial ratios, the total classification ability of models increases. Both methods, the discriminant analysis and multinomial logistic regression, provide similar results, however models estimated by MLR achieve higher classification ability. The best model from this point of view was estimated by MLR and uses all 11 financial and market variables (Model K). Good classification results are then achieved by MLR models using either 4 market indicators (Model G), or 4 combined variables (Model N). The overall results are surprising because they suggest that earnings per share, enterprise value, market capitalization and beta can give a good signal of a bond investment quality.

4.2 Verification and validation

Based on the criterion of classification ability on the original sample, the following three models, (G), (K) and (N) will be examined in more detail. All these models have been estimated by multinomial logistic regression, which allows simpler comparing of results and overall fit of models.

Criterion	Model G	Model K		Model N
Number of predictors:	4	11		4
Predictors included: (*Likelihood ratio tests of parameters*)	EPS LogMarketCap LogEV*** LogBeta***	LogTA Equity_to_TA*** STD_to_TA** ROA* ROE* LogInt_cov	LogCurr_ratio EPS LogMarketCap LogEV** LogBeta**	LogInt_cov EPS LogMarketCap LogEV*
Model fitting: *Chi-Square*	133,286*** (df=12)	150,528*** (df=33)		132,122*** (df=12)
Goodness-of-Fit: *Pearson* *Deviance*	 184.708 (df=219) 78.954 (df=219)	 28.304 (df=165) 30.030 (df=165)		 157.824 (df=264) 114.580 (df=264)
Measures of R^2: *Cox and Snell* *Nagelkerke*	 0.815 0.875	 0.758 0.816		 0.894 0.959

****p<.001, ** p<.01, * p<.05*

Table 6 Verification

To assess the fit of models, we use a log-likelihood statistic, which is based on summing the probabilities associated with the predicted and outcome variables, Tabachnik and Fidell (2007, p. 446) [8]. The statistic indicates how much unexplained information there is after the model has been fitted. The larger the value, the more unexplained observations there are. The chi-square test tests the decrease in unexplained variance from the baseline model to the final model. All the final models explain a significant amount of the original variability, so they better fit than the original model. The next test tests whether the models predicted values are significantly different from the observed ones. If the statistics (Pearson and Deviance) are not significant, than predicted and observed values are not different, and the model is a good fit. All three models are a good fit based on this test. The significance of predictors to the models was assessed by the likelihood ratio tests. In all models, variable LogEV has a significant main effect on rating category classification; it is even the only significant predictor in Model N. Due to a large number of derived models in this study, parameter estimates and odds ratios are not included in this paper, however they can be provided on demand.

The three selected models (G, K, N) were used to predict bond rating of companies other than that used for estimation of models. As the test sample covers only 25 companies, results of the validation will not likely be accurate and can be misleading.

Model	Correct classification 4-rating	Correct classification 2-rating
Model G	16 %	64 %
Model K	28 %	60 %
Model N	32 %	84 %

Table 7 Validation

As expected, the ratio of correctly classified companies is very low, which is likely the result of relatively small control sample. When classifying into four rating groups, all three models give bad results. However, all models contribute significantly to the classification in case of just two rating groups, investment and speculative category. Validation proved that that the Model N provides very accurate predictions.

5 Conclusion

The overall results suggest that market-based indicators contribute to the discrimination more than financial ratios. By adding market-based data to the original set of financial ratios, the total classification ability of models increases. Both methods, discriminant analysis and multinomial logistic regression, provide similar results, however models estimated by MLR achieve higher classification ability. The best model from this point of view was estimated by MLR and uses all 11 financial and market variables (Model K). Good classification results are then achieved by MLR models using either 4 market-based indicators (Model G), or 4 combined variables (Model N). The overall results are surprising because they suggest that earnings per share, enterprise value, market capitalisation and beta can indicate the bond rating category of companies relatively accurately. The most significant variable for bond rating prediction is LogEV. Thus, the enterprise value can be a very important indicator of creditworthiness and can give a good signal of a bond investment quality.

Acknowledgements

The paper is based on research activities sponsored through SGS research project No. SP2012/19. The support is greatly acknowledged.

References

[1] Altman, E. and Eisenbeis, R.: Financial applications of discriminant analysis: A clarification. *Journal of Financial and Quantitative Analysis* **13** (1978), 185 – 195.
[2] Altman, E.I., Sabato, G. and Wilson, N.: The value of non-financial information in small and medium-sized enterprise risk management. *The Journal of Credit Risk* **6** (2010), 95 – 127.
[3] Ang. J.S. and. Patel, K.: Bond rating methods: Comparison and validation. *The Journal of Finance* **30** (1975), 631-640.
[4] De Laurentis, G., Maino, R. and Moletni, L.: *Developing, validating and using internal ratings. Methodologies and case studies*. John Wiley & Sons Inc., Chichester, 2010.
[5] Hosmer, D. W. and Lemeshow, S.: *Applied logistic regression*. John Wiley & Sons Inc., New York, 2000.
[6] Pinches, G. and Mingo, K. A.: A multivariate analysis of industrial bond ratings. *The Journal of Finance* **28** (1973), 1-18.
[7] Rencher, A. C.: *Methods of multivariate analysis*. John Wiley & Sons Inc., New York, 2002.
[8] Tabachnik, B. G.: and Fidell, L. S. *Using multivariate statistics*. Pearson Education, Inc., Boston, 2007.
[9] Waagepetersen, R.: A statistical modeling approach to building an expert credit rating system. *The Journal of Credit Risk* **6** (2010), 81 – 94.
[10] Westgard, S. and Wijst, N.: Default probabilities in a corporate bank portfolio: A logistic model approach. *European Journal of Operational Research* **135** (2001), 338 – 349.