Credit Risk Modeling using Multiple Regressions

Maria Dimitriu¹, Ioana-Aurelia Oprea²*, Marian-Albert Scrieciu¹

¹Academy of Economic Studies, Bucharest, Romania.
²BCR, Erste Group
*Corresponding Author: E-mail: ioanaaurelia.oprea@bcr.ro

Abstract

In classical theory, the risk is limited to mathematical expectation of losses that can occur when choosing one of the possible variants. For banks, risk is represented as losses arising from the completion of one or another decision. Bank risk is a phenomenon that occurs during the activity of banking operations and that causes negative effects for those activities: deterioration of business or record bank losses affecting functionality. It can be caused by internal or external causes, generated by the competitive environment.

Keywords: Banking system, Credit risk, Multiple regression.
JEL Classification: G21, G32

Introduction

The experience of developed countries underlines the need to develop portfolios of information on all bank customers and a database for information processing as a main way to strengthen the bank's position in its relations with customers, regardless of their size, and as a weapon the prevention and avoidance of credit risk [1]. By using credit derivatives, banks keep the loan on their balance sheet. Transferring credit risks using derivatives means risks that credit risk transfers with loan sales or securitizations do not have. Banks using these derivatives have to bear associated counterparty, operational, and legal risks [2].

In studies of the economy or companies, using data and statistical methods for analysis of information is inevitable, therefore perception objectively and effectively of economic reality recommend the use of quantitative analysis methods. Methods exploiting the information collected via the econometric models are based on specific concepts of logic and mathematics. Using the scoring, the lender can appreciate quickly, objectively and consistently the previous loans, and can calculate the probability that the loan will be repaid according to the contract. [3]. Econometric tools available in the investigation of an economic phenomenon are considering the following: [4].

- Identifying features of the phenomenon studied. The first step to be taken in this regard is the choice of economic theories that will guide the research of this phenomenon: it is defined the quantitative relationships between different statistical variables used to characterize this phenomenon.
- Testing the statistical assumptions on some specific aspects of the phenomenon studied.
- Making predictions for a specific time horizon. Companies are considering these forecasts to anticipate and apply a correction to future developments of the economic phenomenon.

Using multiple regression can determine the impact of several independent variables on certain variables (called dependent variable). The general form of multiple regression equation is:

\[ Y_t = a_0 + a_1 * X_{1t} + a_2 * X_{2t} + a_3 * X_{3t} + ... + a_k * X_{kt} + \varepsilon_t, \]

where:
- \( t = 1, 2, ..., n \) – observations of the sample
- \( Y_t \) - observation \( t \) of the dependent variable
- \( X_{jt} \) - independent variables, explanatory, \( j = 1, 2, 3, ..., k \)
- \( X_{kt} \) - observation \( t \) of independent variables \( X_j \)
- \( a_0 \) - constant, free term of equation
- \( a_1, ... k \) - coefficients of independent variables
- \( \varepsilon_t \) - error term of equation.

The coefficient of independent variable reflects how dependent variable \( Y_t \) changes when the independent variable, \( X_j \), changes by one unit, while the other independent variables remain
The main activity of commercial banks is lending activity, so a special importance is given to credit risk management. [5].

This study has the starting point the economic and financial analysis of 337 companies for 3 consecutive years, taking into account 12 quantitative factors. The econometric model proposed considers that the probability of default of loans by a bank is directly dependent on the 12 factors analyzed. Each customer's probability of default is determined by the category of risk to which it belongs. So we divided the customers into five classes of risk. Depending on the values of the indicators analyzed each firm is classified into the following classes of risk, thus giving a value for the probability of default:

- Risk class a with a probability of default 0.3%
- Risk class b with a probability of default 0.5%
- Risk class c with a probability of default 1.5%
- Risk class d with a probability of default 5%
- Risk class e with probability of default 10%

In this study the dependent variable is the probability of default (PD) and independent variables are: evolution of total turnover (ca), commercial return or profit margin (margin), return on equity (ROE) , return on investment (ROI), net cash flow (cf), intensity of investments (imob), investment ratio (inv), equity ratio (KPR), quick liquidity (lich), overall net indebtedness (indat), average accounts receivable (cl) and average accounts payable (fz).

The 12 endogenous variables were calculated by following formula:

\[ PD = a_0 + a_1*ca + a_2*margin + a_3*roe + a_4*roi + a_5*cf + a_6*imo + a_7*inv + a_8*kpr + a_9*lish + a_{10}*indat + a_{11}*cl + a_{12}*fz \]

It is worth mentioning that this case study involves the definition of three different models, one for each year, and at the end of the analysis the best model is recommended.

Determination of the Optimal Regression Model

First we made a descriptive analysis of data, analysis that helped us to identify if data series are stationary. To check the stationarity of variables we used the Phillip Perron test and noticed that all independent variables are stationary, hence their fluctuations around a trend occurring in parallel with the abscissa, and the probability shown by Phillip Perron test is equal to 0. As example we test for the probability of default.

Finally Jarque-Bera test confirms that the variable is normally distributed, as associated test probability 0.00. The assumption of normality of the variable can be checked and rejected also by specific tests Eviews: simple hypothesis test and quantile-quantile graph.

The next step was to determine the coefficients of multiple regression model using the least square method that shows the following values for the variables.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA</td>
<td>0.01154</td>
<td>0.00074</td>
<td>1.551000</td>
<td>0.1219</td>
</tr>
<tr>
<td>MAR/A</td>
<td>0.003239</td>
<td>0.001987</td>
<td>1.629710</td>
<td>0.1041</td>
</tr>
<tr>
<td>ROE</td>
<td>0.014659</td>
<td>0.002988</td>
<td>0.617565</td>
<td>0.0000</td>
</tr>
<tr>
<td>ROI</td>
<td>-0.036641</td>
<td>0.012140</td>
<td>-2.524275</td>
<td>0.0121</td>
</tr>
<tr>
<td>CF</td>
<td>-5.45085</td>
<td>6.81560</td>
<td>-0.793590</td>
<td>0.4280</td>
</tr>
<tr>
<td>IMOB</td>
<td>0.091545</td>
<td>0.004556</td>
<td>2.025906</td>
<td>0.0000</td>
</tr>
<tr>
<td>INV</td>
<td>0.000446</td>
<td>0.000431</td>
<td>0.086339</td>
<td>0.9357</td>
</tr>
<tr>
<td>KPR</td>
<td>0.016979</td>
<td>0.004097</td>
<td>0.454334</td>
<td>0.6500</td>
</tr>
<tr>
<td>LICH</td>
<td>0.001557</td>
<td>0.001936</td>
<td>0.590510</td>
<td>0.2807</td>
</tr>
<tr>
<td>INDAT</td>
<td>0.010230</td>
<td>0.005653</td>
<td>1.862405</td>
<td>0.0697</td>
</tr>
<tr>
<td>CL</td>
<td>0.001382</td>
<td>0.001822</td>
<td>0.795683</td>
<td>0.4486</td>
</tr>
<tr>
<td>FZ</td>
<td>6.152e-05</td>
<td>0.005160</td>
<td>0.699426</td>
<td>0.4919</td>
</tr>
<tr>
<td>C</td>
<td>3.056120</td>
<td>0.297992</td>
<td>10.832400</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

As previously mentioned the proposed model to be used in lending decisions will be chosen as the best model of the three analyzed.

Regression equation for Model 2 is as follows:

\[
PD = -0.001838769405 \times CA - 0.05770658157 \times MARGIN - 0.00087878188804 \times ROE - 0.00953228523 \times ROI - 1.39035012e-05 \times CF + 0.008825282231 \times IMOB - 0.0001422659158 \times INV - 0.03281036515 \times KPR + 0.0003608745752 \times LICH + 0.008793179356 \times INDAT - 0.000203597285 \times CL + 0.009775815716 \times FZ + 2.750100729 
\]

Analyzed indicator have values displayed in the table below:

For Model 3 has produced the following values:

With form equation:

\[
PD = -0.005379910878 \times CA - 0.05169358257 \times MARGIN - 0.004833026573 \times ROE - 0.02375204693 \times ROI - 1.046135381e-05 \times CF + 0.01552840358 \times IMOB - 0.000193004128 \times INV - 0.0417891334 \times KPR + 0.008371068961 \times LICH + 0.0003458981285 \times INDAT + 0.00274293772 \times CL + 0.002757301955 \times FZ + 3.074346603 
\]
As a criterion for choosing between the three competing models we used Akaike and Schwartz tests. According to them the best performing model is one that has the minimum value for one of the two indicators, because these two indicators decrease their values with decreasing adjustment errors. At the same time the quality of estimation is improved by increasing the size of data series used to estimate parameters. [6]

Analyzing the output of each regression we concluded that the optimal model is:

\[
PD = 0.00115354791 \times CA + 0.00323832969 \times MARGIN - 0.01496891484 \times ROE - 0.0305444 \times ROI - 5.401135453 \times CF + 0.001496178241 \times IMOB - 0.003446 \times INV + 0.001792029 \times LICH + 0.001381941557 \times CL + 6.1483895576e-05 \times FZ + 3.055119835
\]

For this equation, Figure 6, the test has the lowest Akaike value, respectively 3.73400 compared to 4.191457 for the corresponding Model 2, and 4.241265 for the Model 3. At the same time Schwartz criterion displays the minimum value also for Model 1: 3.881363, compared to 4.339143 for Model 2 and respectively 4.388627 for Model 3.

Source: own calculations

Fig. 5: Estimation parameters for 2010

The model shows satisfactory statistical results. \( F \) statistic value and its associated probability value 0.000, suggesting the model is correct. Dependence of exogenous variable regression with regression factors is given by the coefficient of determination Adjusted R-squared, noted R2. Mathematically it is calculated as:

\[
R^2 = \frac{SPT}{SPE} = 1 - \frac{SRE}{SPT}, \text{ where}
\]

- \( SPT = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \) quantifies dispersion of endogenous variable below the action of endogenous factors in the model, and factors unregistered

- \( SPR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \) measures the influence of exogenous variables in the total endogenous variable of series size.

- \( SPE = \sum_{i=1}^{n} (\epsilon_i)^2 \) is the sum of squared errors adjustment and measure the influence of unregistered factors of multiple regression model. [4]
The value of coefficient of determination must belong to interval \([0, 1]\) and increases with increasing number of endogenous variables used to define the regression model. If the indicator is different from 0, then the endogenous variable is explained, largely, by endogenous variables. Value of the coefficient R-squared indicates that over 34% of the variability of the probability of default is explained by the evolution of total turnover (ca), commercial return or profit margin (margin), return on equity (ROE), return on investment (ROI), net cash flow (cf), intensity of investments (imob), investment ratio (inv), equity ratio (KPR), quick liquidity (lich), overall net indebtedness (indat), average accounts receivable (cl) and average accounts payable (fz); the rest is the contribution of factors that are not included in the analysis. Durbin Watson statistic (DW) is a statistical test which tests the serial correlation of errors. If the errors are not correlated, the value of DW will be around 2. Value 2.17 of Durbin-Watson test in Fig. 7, suggests the autocorrelation of the first order residues, which has a negative impact on model validation, even if the value of F statistics and the associated probability value is 0.000 suggesting the correct model specification.

### Hypothesis Testing for Multiple Regression Model

So far, the results of econometric regression analysis show that the model can be validated. To make sure that the estimated parameters are effective, homebound and linear, the multifactorial model implies complying with the assumptions outlined above, where we have defined multiple regression, so that the next step is hypothesis testing.

#### Autocorrelation residues (errors)

The value of Durbin Watson test (DW) different from 2 shows that it is possible residue autocorrelation, which must be confirmed by the Breusch-Godfrey test Figure 8. Statistical value of 1.69 and R-squared of 3.51 suggests rejection of the null hypothesis, i.e., the lack of residual values correlation.

### Heteroscedasticity

According to this hypothesis, variant of residues must be constant; otherwise the estimators are no longer effective. As above, we proceed to make a statistical test. F-statistic value of 3.56 and R-squared and their associated probability of 0.00 accept null hypothesis, so there is no heteroscedasticity.

### Normality of Residues

The verification for normality residues is done by residues histogram and Jarque-Bera test, Figure 10. The test measures the difference between the coefficient of asymmetry and kurtosis for the distribution analyzed with the normal distribution. In our study, coefficients of asymmetry (skewness) and flatness (kurtosis) are significantly different from 0, respectively 3, and the histogram is not symmetric. Kurtosis indicator has a value greater than 3, so we have a leptokurtosis distribution. [6].
The value 3.56 for F-statistic and R-squared and also their associated probability of 0.00 accept null hypothesis, so there is no heteroscedasticity.

These results induce the failure of assumption of normality of residues. Also, Jarque-Bera test suggests that errors are not normally distributed, since the probability is 0.00.

Predictions

To test the quality of the model again, we wanted to do an estimate based on recorded data. We see that the forecast model fails to mimic almost all series of actual values, which is another added advantage to validate the model.

Conclusions

The main problem in creating an external credit assessment institution is to build a statistical model to quantify the probability of default in accordance with the requirements of Basel II. In
terms of methodology, credit risk modeling includes:[7].

- Comprehensive evaluation of characteristics of the borrower and facility that he wishes to access;
- Meaningful differentiation of risk, namely granularity grading scale;
- Reasonable accuracy and consistency over time of estimates of quantitative credit risk

After analysis and testing, we concluded that there is a strong correlation between the probability of default and endogenous variables: evolution of total turnover (ca), commercial return or profit margin (margin), return on equity (ROE), return on investment (ROI), net cash flow (cf), intensity of investments (imob), investment ratio (inv), equity ratio (KPR), quick liquidity (lich), overall net indebtedness (indat), average accounts receivable (cl) and average accounts payable (fz).

Probability of Default - PD is determined, largely, by the evolution of the 12 indicators as well as other company specific factors. It is important to note the superiority of the factorial model credit risk management. This is confirmed firstly by the high value associated coefficient of determination, \( R^{2} = 34.05\% \), and the tests based on information criteria.

This is consistent with modern theories that minimize credit risk, specific for loan portfolio and can be achieved through diversification. Because the model assumptions were checked and met, we validate the proposed model for analysis because of good results of statistical tests performed. The scoring system developed with multiple regression is an important element for the commercial banks for credit risk management. This tool for credit risk management provides information on the quality of bank debtor company and thus the bank can make a correct decision as regards the amount, the guarantees to be made, and not least: interest and cost of the loan. Over time it has been shown that the onset of financial crises was based on an inadequate management of credit risk, therefore should be noted that credit risk is the most important factor that influencing the performance of a banking company. Credit risk management should not be restricted only to solve problems arising after its manifestation but rather to prevent such crises.

**References**