UNIVERZITA KARLOVA V PRAZE

FAKULTA SOCIÁLNÍCH VĚD

Institut Ekonomických Studií

Matej Kosturák

The prediction of corporate bankruptcy and credit risk

Master thesis

Praha 2013

Author of the thesis: **Matej Kosturák**

Consultant: **PhDr. Petr Gapko**

Opponent:

Date of the defense: 2013

Evaluation:

Bibliographic citation

KOSTURÁK, Matej. *The prediction of corporate bankruptcy and credit risk*. Praha, 2013. 74 pg. Master thesis (Mgr.) Univerzita Karlova, Fakulta sociálních věd, Institut Ekonomických Studií. Supervised by PhDr. Petr Gapko

Abstract

This thesis present concise but comprehensive overview of most important paper dedicated to prediction of corporate bankruptcy, as well as overview of the theory behind the employed models and crucial indicators for quality assessment and comparison of the estimations. Manually collected data includes financial statement, identification information and especially specifications of management and responsible persons. From this point of view, data collected are of high quality and in Czech Republic relatively unique. Noticeable is also multiple imputation method used, current "state-of-the-art" technique for missing data treatment. Practical part concentrates on models estimation for various data setting, when contrasting models on raw and truncated datasets. By smoothing data, significantly better model can be estimated with superior discriminating power on the same data points. Inclusion of macroeconomic variables as well as even more significant governance indicators according to current stage of research, improved estimated models.

Anotace (abstrakt)

Táto práce prináša výstižný a súhrny prehľad najdôležitejších vedeckých článkov a statí venovaných téme predikcie firemných bankrotov, ako aj prehľad teórie použitých modelov spolu s hlavnými kvalitatívnymi indikátormi a ukazovateľmi na vyhodnotenie a porovnanie odhadov. Manuálne zbierané dáta obsahujú účtovné uzávierky, identifikačné údaje and špeciálne údaje o osobách vo vedení podniku. Z tohto uhla pohľadu sa jedná o výnimočný data súbor v prostredí Českej republiky. Technika imputácie chýbajúcich hodnôt, momentálne najnovšiu a najviac odporúčanú metódu na riešenie problému nevyplnených hodnôt. Praktická časť sa koncentruje na odhad

modelov v prostredí rozdielnych dát a porovnáva odhadnuté modely v upravených a originálnych dátach. Značne lepší model je možné odhadnúť po vyhladení dát, s lepšou diskriminačnou schopnosťou na rovnaké pozorovania ako v originálnom data súbore. Zahrnutie makroekonomických ukazovateľov ako aj ešte dôležitejšie, podľa momentálneho výskumu, ukazovatele kvality správy podniku malo signifikantný efekt vo výslednom modely.

Keywords

bankruptcy prediction, financial stability, corporate sector risk, corporate rating

Klíčová slova

predikcia bankrotov, finančná stabilita, riziko firemného sektoru, rating spoločnosti

Declaration

1. Hereby I declare that this master thesis has been written by me, and me only and that I have only used resources that are identified in reference list.

2. I agree with using this thesis for the academic purposes.

In Prague 31.7.2013 Matej Kosturák

Acknowledgement

I would like to thank to my consultant PhDr. Petr Gapko for his patience, advice, resources and all the help he provided me with. It was an enriching experience to write my thesis under his supervision.

Master Thesis Proposal

Institute of Economic Studies
Faculty of Social Sciences
Charles University in Prague

Author:	**Bc. Matej Kosturák**	Supervisor:	PhDr. Petr Gapko
E-mail:	mkosta@centrum.cz	E-mail:	petr.gapko@seznam.cz
Phone:	776 537 107	Phone:	
Specialization:	Finance, Financial Markets and Banking	Defense Planned:	June 2013

Proposed Topic:

The prediction of corporate bankruptcy and credit risk

Topic Characteristics:

Credit risk models are inherent part of risk analysis in financial institutions as well as other credit grantors. Default on meeting financial obligations causes higher write-offs of bad debt consequently reflected in higher reserves which decrease amount of total capital available in credit market. Undervaluation of credit risk from financial institutions can lead to bigger structural problems, which can pose serious threat to financial stability. Non-financial institutions are exposed to credit risk as well. In case of important partner default, serious operational problems can withstand.

In my work I intend to verify several hypotheses which could help improve present comprehension of corporate bankruptcy probability models. Most past studies predict corporate bankruptcy on yearly basis. Nowadays, data are collected more frequently and higher volume of more frequent data is available, which allows more elaborated verification. Confirmation of the first hypothesis can change convention of institutions which use credit risk models and encourage them to employ more frequent data and therefore enhance model prediction power. Opposite result can help same institutions in simplification of their models and reducing amount of calculations which can save time and costly resources.

In case of incorporation of macroeconomic estimation of probability of bankruptcy into a microeconomic probability model, it might afford us to control model for changing macroeconomic conditions and cyclical bankruptcy. Target of second hypothesis is primarily to verify time stability of bankruptcy probability model and in case of affirmation. Secondary target would be to fit the best estimation of time dependent model.

Relatively newly observed topic in field of credit risk models is research of industry specific effects on probability of bankruptcy or default. Quality of data did not allow researchers wider variability of research questions, and multitude of industry specific data was not sufficient for expansion of probability models in individual characteristics. In case of confirmation of industry dependent variables as significant, more accurate models could result from such an adjustment.

Data are crucial and most important building block of this thesis. Optimal data would be with as high frequency as possible which would mean monthly frequency. Since such a frequent data would be accessible very unlikely, quarterly data are a reasonable solution. Number of time periods should be sufficient for ability to enquire time stability of probability model, as well as control for different business cycles in order to catch bankruptcy occurrence in time series. As a requirements for a data quality are relatively higher, in this stage of work I do not intend to specialize on concrete geographic area. Data should be composed from accounting statements (balance sheets and income statements), alternatively from financial indicators derived from accounting reports. Accounting statements should be sufficient to categorize into 4 main groups: solvency indicators, liquidity indicators, profitability indicators and activity indicators.

Hypotheses:

1. Hypothesis #1: More frequent data improve quality of predictions of corporate bankruptcy and credit risk
2. Hypothesis #2:Incorporation of macroeconomic indicators of bankruptcy probability would enhance prediction power of the model
3. Hypothesis #3: Predictions of bankruptcy and credit risk is industry dependent

Methodology:

As an adequate method for research of probability of corporate and credit risk I intend to utilize credit rating method. Credit rating is standard approach in risk estimation process. More specifically financial rating will be employed, when data from accounting statements should be sufficient to categorize into 4 main groups: solvency indicators, liquidity indicators, profitability indicators and activity indicators. Using econometric model, which will be selected according to available data and their quality, proper indicator should be selected to construct rating function, which express corporation creditworthiness. Model will be built as a predictive model, when data with historic information, are used to construct rating function, which determinate score of a individual company by insertion its financial indicators into developed rating function. Score can be regarded as probability, that particular company will default on meeting their obligations, or bankrupt at all. There are plenty of methods to build rating function, as linear regression, decision trees, neural networks, expert systems or logistic regression. Most likely logistic regression will be used, as it is one of the most commonly used approaches even in non-academic environment, what would make results of my thesis widely applicable.

Outline:

1. Introduction
2. Literature overview
3. Scoring models with higher focus on practically used model
4. Financial indicators construction and explanation
5. Data description
6. Resulting estimated model
7. Application of the results of the model
8. Summary

Content

Introduction

Predicting probability of corporate bankruptcy or different characteristics and signals of companies' deterioration is an important part of modern financial world. Various techniques are used daily in credit risk departments of banks due to regulatory requirements from bank supervision as well as their employment for internal credit assessment of companies. Information from accounting statement serves as a base for financial indicators construction as original building block of bankruptcy prediction models. Nowadays, also other indicators and variables assist in model estimations, such as stock market data or macroeconomic indicators. Several models can be employed for model estimation. There is no definite conclusion about the best model for bankruptcy prediction in academic community, however the most popular are binary outcome models as Logit or Probit and Multivariate Discriminant Analysis (MDA). The very first listed model is employed for model estimation also in this thesis, since it is widely used in practice, meaning that potential results could be employed in real-world setting.

The most influential and important literature is described in the first chapter. Starting with cornerstone of modern bankruptcy prediction analysis raised by univariate analysis method presented in Beaver's work, through one of the first attempts of multivariate analysis combined with univariate one by Deakin to one of the most influential works in this area conducted by Altman and his MDA method accompanied by contributional discussion about application of MDA proposed by Joy and Tollefson. Methodology of ratio selection as well as construction of datasets and models are briefly but sufficiently explained. Following part is dedicated to binary outcome models in pioneering influential work of Olhson with Logit model and Zmijevsky with weighted probit model. To the end of literature overview chapter, there are also some other interesting articles concerning modern utilization of alternative indicators and models for small enerprises.

The next part deals with binary outcome model theory with description of logic behind model building. Necessity of using binary models is explained on the background of limitation of classical linear regression model. Logit and Probit models are presented alongside with their function forms and major difference consisting from use of different distribution functions. Subsequently, major indicators and ratios for accessing fit and quality of models are listed and briefly explained. Methodological chapter follows with data collection description, financial ratio construction and characterization. Macroeconomic indicators were collected and explained altogether with new and modern governance indicators constructed from firms'

provided representation and contact public information. Modern missing data technique of multiple imputation was employed and is described as well, thus closing methodological chapter.

The last chapter is dedicated to practical estimation of Logit model. Various techniques are examined and evaluated. Two models are presented and contrasted on the ground of different dataset processing and adaptation, mainly connected with neglecting or smoothing of influential observations and outliers. At the end, the final model is more closely described and explained altogether with recommendations for practical use of Logit models and related dataset treatment techniques.

1. Literature overview

1.1 Beaver

One of the first major studies dedicated to prediction of corporate bankruptcy was carried out by Beaver [Beaver 1966]. However, there are some previous studies carried out by Fitz Patrick [Patrick 1932], Winakor and Smith [Winakor, Smith 1935] and Merwin [Merwin 1942][1]. Unfortunately, they were not available for this paper. Nevertheless, Beaver sums up main results of their works.

Beaver´s study was conducted on sample from Moody's industrial manual, which contained data for industrial, publicly owned companies. Author identified 79 failed firms in the time period from 1954 to 1964. Firms were classified according to their activity/industry and assets' size. Dataset was heterogeneous, since 79 bankrupted firms were conducting business in 38 different industries, while 18 industries contained just one company.

Selection of non-failed firms was carried on via paired-sample method. Author looked for the most similar companies to the failed ones, according to chosen criteria of asset size and industry. Financial reports of defaulted companies were collected for five years prior to a failure, with additional condition - financial statement could not be older than six months before the date of failure. Financial report data of non-failed firms were distributed according to years before failure, corresponding to the years of their paired firms.

Author calculated up to 30 financial ratios from all possible combinations according to three criteria: popularity (usage in literature), good performance in previous studies and ratio defined in cash-flow concept. Ratios are divided into six sections[2] according to a common component. Only one ratio from group was scrutinized in the analysis, which was performed in three steps:

(1) comparison of mean values

(2) dichotomous classification test

(3) an analysis of likelihood ratios

[1] All cited in [Beaver 1966]

Comparison of mean values

Mean values of the ratios were computed for both, failed as well as non-failed firms for each year in the time horizon. Selected ratios were cash-flow to total debt, net income to total assets, total debt to total assets, working capital to total assets, current ratios, and the non-credit interval[3]. In the analysis of evidence author states: "The difference in the mean values is in the predicted direction for each ratio in all five year before failure. Failed firms not only have lower cash flow than nonfailed firms but also have a smaller reservoir of liquid assets. Although the failed firms have less capacity to meet obligations, they tend to incur more debt than do the nonfailed firms." [Beaver 1966: 80] Author points out also a trend for failed and non-failed firms in their ratios. Such trend seems to be constant with small deviations for non-failed firms. On the other hand, departure from trend in case of failed firms is obvious, with magnifying effect to the proximity of default. Data also indicated a strong consistency, suggesting that there is a significant difference in ratios between failed and non-failed corporations.

Beaver's results are consistent with previous studies. Fitz Patrick [Patrick 1932 cited in Beaver 1966] indicated significant difference in the ratios between failed and non-failed firms up to three years before default. Winakor and Smith [Winakor, Smith 1935 cited in Beaver 1966] scrutinized mean ratios for ten years before failure and found out an increasing difference in means as the date of default approached. Also Merwin [Merwin 1942 cited in Beaver 1966] verified difference in increase of means to the end of companies' life.

Author himself identified several problems with his methodology. When focusing solely on means´ difference, analyst could neglect possibly important question – How large is the difference? Mean analysis relies solely upon one point of ratio distribution. Different distribution of ratios, even symmetrical, could overlap, if dispersion of means is high enough. Such case poses threat to predictive ability. On the other hand, in case of skewed distribution, extreme observations could cause majority of the difference in means.

Dichotomous Classification Test

Seeing that means could not been used on their own, Beaver employed Dichotomous Classification Test, which tries to predict corporate default exclusively by knowledge of

---

[2] Cash-flows ratios, net-income ratios, debt to total-assets ratio, liquid-assets to total assets ratio, liquid-assets to current debt ratios, turnover ratios

[3] "immediate assets (current assets excluding inventory and prepaid expenses) minus current liabilities,

[4]divided by total operating expenses excluding depreciation [Palepu at al. 2007 : 419]

financial ratio. Test makes dichotomous prediction – firm is in default or is not. After arranging ratios in ascending order, author visually identified break-even point, at which incorrect predictions would be minimal. If the firm is above or below (depending on used ratio), firms is classified as non/failed. Subsequently after classification of every firm, the predictions were contrasted to actual default states and proportion of incorrect predictions was calculated. The proportion of wrong predictions could be taken as indicator of prediction ability.

Author defined as a strongest failure predictive ratio cash-flow to total-debt ratio, since one year before default the error was only 13%. Obviously, not all ratios´ prediction power was equally accurate. Net-income to total-debt ratio (very high correlation with cash-flow to total-debt ratio) was identified as second best ratio, followed by total-debt to total-assets. Liquid assets ratios[4] were considered as the least feasible ratios.

Beaver also tried to identify industry effect. After dividing companies into two groups, predictions were compared by classification test. No significant difference was noted, even though distribution of industries differed in both subsamples. Therefore another indirect test was used for identification industry effect. Comparison of paired and unpaired analysis percentage error showed small, but statistically significant higher error in unpaired analysis, what could signalize residual effect of industry.

Analysis of Likelihood Ratios

Classification test also has its drawbacks. Taking ratio as dichotomous, even after admitting real life decision as dichotomous, conditional error may be dependent on the magnitude of the ratios. Beaver used histograms to assess the likelihood ratios from financial ones. For example, in the case of the cash flow to total debt ratio, 28 % of the ratios of the non-defaulted companies fall in the interval 0.1 to 0.2 and 21 % of defaulted companies fall in the interval -0.1 to -0.2.

Conclusion

Beaver did not use financial ratios as a predictor as bankruptcy per se, but his work can be regarded more as a scrutiny of accounting figures, which employs ratios in order to assess financial health of the company. Beaver verified, that distribution of the ratios of non-

[4] cash to current liabilities, quick assets to current liabilities, current ratio (current assets to current liabilities)

defaulted companies were stable through observed time horizon, while ratio distribution of the defaulted companies exhibit deterioration closer to failure.

Main observations of Beaver´s study are valid even in these days. He stated:

1) Not all ratios predict equally well.
2) The ratios do not predict failed and non-failed firms with the same degree of success [Beaver 1966: 91]

One of the real life lesson from these remarks can be that investor will never be absolutely able to eliminate chance of financing a company, that will default.

Beaver himself identifies several problems in his study. He found out, that all of the observed ratios do not conform normality assumption. Even after a simple transformation (log and square root) data were as badly skewed as the original distribution of ratios. Beaver used a univariate model because most of the multivariate techniques in 1960's relied upon assumption of normality.

Beaver also identified common problem in failure prediction models. "The sample of failed firms will include those firms whose illnesses were not detectable through ratios. This is biased sample for investigating the usefulness of the ratios. Important information is missing – how many firms were saved from failure because their problems were detected in time through the use of ratios?" [Beaver 1966: 101]

1.2 Deakin

Deakin [Deakin 1972] in his study firstly replicated an approach of Beaver, when he used same ratios. Secondly he tried *linear combination* of the Beaver's best fourteen financial indicators predicting potential failure. Deakin however used different definition of failure, including firms, which filled for insolvency, bankruptcy, or were liquidated in order to satisfy creditors. Beaver used broader definition, as loan obligation default or missed preferred dividend payment.

Sample of thirty-two failed companies between 1964 and 1970 was selected and matched according to industry, year of financial information and asset size to nonfailed companies. Most of the ratios estimated by Deakin were basically consistent with Beaver's measurement. There was significant difference in the cash/sales ratio predictive ability. Deakin explains "One possible explanation is that corporations tended to invest more of their cash reserves during the late 1960's when interest rates were high. Thus a low cash/sales ratio may have been due to good money management rather than to general company mismanagement." [Deakin, 1972: 171]

Using the discriminant analysis, Deakin looked for the linear combination of financial indicators with best discriminating power between the classified groups. Important premise of discriminant analysis procedure is a random groups-categorizing from independent samples. Therefore, a second sample of thirty two non-failed companies was chosen randomly. Moody's industrial manual data, from the 1962 to 1966 was employed. The data of the companies corresponded to five-year time horizon in the original sample. The fourteen ratios defined by Beaver were applied to the discriminant analysis program, with discriminant weights output, indicating linear combination maximizing difference between the categories, altogether with vector signaling the relative importance of each variable. Summation of the product of every variable, consequently multiplied by corresponding ratio, produces a score maximizing difference between the two categories.

"The application of statistical techniques, particularly discriminant analysis, can be used to predict business failure from accounting data as far as three years in advance with a fairly high accuracy." [Deakin, 1972: 178] On the other hand, it is necessary to point out that Deakin´s study consisted only from small sample and discriminant analysis prevents to catch changing conditions of companies, when prohibit classifying them into distinct groups over time.

1.3 Altman

Altman [Altman 1968] in his paper "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy" tried to assess quality of ratio analysis as an analytical technique. Even though this particular article is regarded as the first multivariate (more specifically multiple discriminant analysis - MDA) approach to predict firm failure, Altman states: "Prediction of corporate bankruptcy is used just as an illustrative case" (for assessing quality of ration analysis as an analytical technique) [Altman 1968: 589].

Altman prefers multivariate analysis because of a major handicap of a univariate one, which is an ambiguity. Using only one ratio as a benchmark, or employing several ratios individually, can pose significant bias on decision-making. Company, in solvency problems, may record above average liquidity levels in the same time. In that case, prediction or identification of company as defaulted lies solely on arbitrary decision of examiner. Generally, majority of companies in financial distress reports mixed financial ratios, what undermine even more the usefulness of a univariate analysis.

MDA

Multivariate discriminant analysis uses classification of the observation into groups based on a prior selection of specific individual properties. This technique is suitable for qualitative dependent variables. MDA's aim is derivation of linear combination of best discriminating characteristics between different groups. In case of a prediction of corporate bankruptcy, MDA establishes discriminatory values of financial ratios for the best distinction between bankrupt and non-bankrupt companies.

MDA in comparison to a univariate analysis possesses two main advantages. The first improvement is a consideration of complete range of information available for the firms as well as interaction of multiple inputs. The second advantage is a decline in space dimensionality, when original dimension is reduced because of discriminant inclusion.

Since analysis of bankruptcy requires only two groups of companies, failed and non-failed, Alman used simplest form of discriminant functions in only one dimension:

$$Z = v_1 x_1 + v_2 x_2 + \cdots + v_n x_n$$

where:

$v_1, v_2, \ldots v_n$, are discriminant coefficients, which MDA calculates
$x_1, x_2, \ldots x_n$, are independent variables with actual values

Result of discriminant function is single score, also known as Z value, which can be used for categorization of companies. Using financial ratios as independent variables poses a threat of collinerality, since ratios with the same denominator will have high correlation. This obstacle can be overcome by choosing smaller number of utilized ratios, what also offers another benefit of simple model with fewer explanatory variables.

Data sample

Data sample in Altman's study consisted of 66 companies, divided into 2 equal groups. Bankrupted companies´ group consists of firms, which filed for bankruptcy under Chapter X of the US national Bankruptcy Act between 1946 and 1965 with average assets of $6.4 million. Altman, similarly as Beaver, used paired sample technique. Collected data for non-bankrupted companies were from similar time horizon and matched in industry and asset size. Twenty-two variables in form of financial ratios were selected into five ratio categories, including profitability, liquidity, leverage, solvency and activity ratios. Ratios were chosen based on the previous popularity in studies and potential relevancy. Moreover Altman used

completely new ratios, meaning not only ratios derived from financial statement, but also information on market value of equity.

From the twenty-two starting ratios, five ratios were chosen as the most suitable for prediction of corporate bankruptcy by the MDA computer program[5]. "In order to arrive at a final profile of variables the following procedures are utilized:

- Observation of the statistical significance of various alternative functions including determination of the relative contributions of each independent variable;
- evaluation of inter-correlations between the relevant variables;
- observation of the predictive accuracy of the various profiles;
- judgment of the analyst" [Altman 1968: 594]

In the final function, the most significant ratios from a univariate analysis were not included, because function was determined based on its overall prediction power. Because numerous combinations of variables are possible, it is hard to claim optimal solution. The final function derived by MDA in Altman's work was as follows:

$$Z = .012X_1 + .014X_2 + .033X_3 + .006X_4 + .999X_5$$

where:[6]

X_1 = Working capital/Total assets
X_2 = Retained Earnings/Total assets
X_3 = Earnings before interest and taxes/Total assets
X_4 = Market value equity/Book value of total debt
X_5 = Sales/Total assets
Z = Overall Index

Empirical results

Z – score, after identifying individual coefficients, is possible to compute for every company in the sample, hence identify the firm within either bankrupt, or non-bankrupt group according to value of the score. Since coefficients in Z-score equation function possess positive signs, the higher a likelihood of company's bankruptcy is, the lower its discriminant score should be.

[5] The MDA program developed by W. Cooley and P. Lohnes as stated in [Altman 1968]
[6] For detailed description and rationale behind used ratios see Altman, [Altman 1968: 594-596]

While function is built upon past observations, model can be identified as an explanatory one. However, with new additional companies, also predictive ability of the function can be evaluated. Generally, matrix classifying either correctly identified companies according to Z-score, or misclassifications including Type I[7] and Type II error, can be constructed as follows:

Table no.1 - Accuracy Matrix

	Predicted Group Membership	
Actual Group Membership	Bankrupt	Non-Bankrupt
Bankrupt	H	M_1
Non-Bankrupt	M_2	H

Source: Altman 1968: 599

where:

H – correctly classified companies, (H as a Hits)
M_1 – Type I error (M as a Misses)
M_2 – Type II error

The sum of correctly classified firms divided by total number of companies in the sample can be used as a measurement of success of the model, when one would get percentage of correct classifications. Such a measure can be employed comparably as a coefficient of determination in regression analysis (R^2), where main difference lays in the explanation of percentage of variation.

The original model of classifying initial sample was very good, since 95 per cent (63 out of 66 companies) were identified correctly. Type I error was just 6 per cent and Type II only 3 per cent. However, one should be aware of the fact that model based on initial sample analysis was derived by estimation of coefficients of the same companies, what poses bias on such an analysis. When taking into account prediction power of the test two years before bankruptcy, only 72 per cent successfully identified companies could be found, with 28 per cent Type I error and only 6 per cent Type II error.

Longer time horizon predictive ability of the model was examined by Altman as well. Beaver [Beaver 1966] found out, that deterioration of the companies can be detected by the univariate analysis even five years before actual bankruptcy. Altman however states another important

[7] Type I - companies bankrupted in reality, but classified as non-bankrupt, Type II is defined vice versa

question: “Is it enough to show that a firm’s position is deteriorating or is it more important to examine when in the life of a firm does its eventual failure, if any, become an acute possibility?” [Altman 1968: 604]

In order to answer this question, data for thirty-three firms from original sample were collected. Difference in this analysis was a sobserved period. Third, fourth and fifth year prior to default were taken into account. It is rational to suppose that relative predictive power of the model would decrease with the increasing lead time, as one could observe in previously stated study by Beaver. Table no.2 summarizes the predictive power for the overall observed five year period according to Altman.

Table no.2 - Five Year Predictive Accuracy of the MDA Model (Original Sample)

Year Prior to Bankruptcy	Hits	Misses	Per cent Correct
1st n=33	31	2	95
2nd n=32	23	9	72
3rd n=29	14	15	48
4th n=28	8	20	29
5th n=25	9	16	36

Source: Altman [Altman 1968: 604]

As can be observed from the Table no.2, relative predictive accuracy of the model behaves according to hypothesis of deteriorating predicting power with increase in time horizon of prediction prior to bankruptcy. However, in 4th and 5th year, an inconsistency can be seen, when a model became less reliable with more years, therefore relative change has little meaning as to explaining predictive power of the model.

Two important conclusions from Altman study can be derived [Altman 1968: 604- 606]:

- All of the observed ratios show a deteriorating trend as bankruptcy approached
- The most serious change in the majority of these ratios occurred between the third and the second year prior to bankruptcy

One of the main advantages of the Altman´s model was overall simplicity for that time. Even though chosen methodology required quite demanding computer computation power (for time horizon given) for estimating variables in Z-score function, final model could be used by professionals and financial institutions even without computers. After establishing correct

function and particular coefficients by Altman, for getting Z-score, it was enough to compute financial ratios from reports.

While model as whole worked relatively well, comparing to previous attempts to predict potential bankruptcy, "zone of ignorance" was the most obvious drawback of the model. Altman tried to identify a point where minimum number of misclassifications could be found. "The best critical value conveniently falls between 2.67-2.68 and therefore 2.675, the midpoint of the interval is chosen as the Z value that discriminates best between the bankrupt and non-bankrupt firms." [Altman 1968: 607]

Altman`s model, or more generally methodology, could be practically used in several areas. The most common area for use would be B*usiness Loan Evaluation*, since prediction of corporate bankruptcy is made mainly in order to minimize bad loans. However, model on its own does not offer absolute control of identification of potential threat, should be used as an addition to qualitative and individual analysis of companies. Resulting, other possible application of derived model comes, particularly *Internal Control Considerations and Investment Criteria*. Model can be used not only for prediction of default, but also for identification of problems of company, and according to previous results, up to two years before bankruptcy. Hence change to corporate strategy or merger with another company could be suggested to the firm to avoid bankruptcy.

1.3.1 Zeta analysis

After various attempts to construct bankruptcy prediction models for manufacturers[8] or for specific industry sectors[9], authors [Altman, Haldeman, Narayanan 1977] decided to construct new bankruptcy classification model. The sample is composed of 53 bankrupt companies and paired sample of 58 non-bankrupt ones. Matching is set according to industry, size and time. Average asset size of bankrupted companies group is almost $100 million. Manufacturer and retailer group is almost equally numerous in all the sample. Up to 27 variables were counted and were classified into groups of profitability, coverage and other earning relative to leverage measures, liquidity, capitalization, earnings variability and few miscellaneous measures.

[8] Beaver (1967), Altman (1968),Deakin (1972) and Edmister (1972 cited in Altman, Haldeman, Narayanan 1977)

[9] Altman on railroads (1973), Sinkey on commercial banks (1975), Korobow and Stuhr (1975) and with Martin (1976), also on commercial banks, Altman and Lorris on broker/dealers (1976) and Altman on savings and loan associations (1977a) all cited in Altman, Haldeman, Narayanan 1977

Reporting adjustments

Authors adjusted data for recent important changes in accounting and reporting. First major change was detected in capitalization of non-cancellable operating and finance leases. Amounts attributed to capitalization were added to assets and liabilities, interest costs of lease were implied as well. Secondly reserves of contingency nature were, if applicable, included in equity and net profit was adjusted for change in reserves. In case of reserves related to revaluation of assets, they were netted against them. Thirdly, minority interests and other liabilities on the balance sheet were netted against other assets, because it allowed authors for comparison of earnings with assets generating them. Fourth, non-consolidated entities were consolidated using pooling of interest method. Fifth, goodwill and intangibles were deducted from assets and equity, because of the problems with identification of true economic value. Sixth, capitalized R&D and interest costs, as well as other deferred charges were expensed, not capitalized.

Multivariate statistical technique, for example discriminant analysis, was used with scrutiny of linear and quadratic structures. Using an iterative process, authors reduced variables up to seven, which proved to be statistically most reliable: *return on assets, (*EBIT/Tot. As.); *stability of earning*s (normalized measure of the standard error of estimate around 10-years trend); *debt service* (interest coverage ratio); *cumulative profitability* (retained earnings/Tot. As.; by authors identified as most important); *liquidity* (current ratio), *capitalization* (equity/total capital) and *size* (Tot. As.)

Classification accuracy based on data from one year prior to failure using the linear model, was 92.8% in total (96.2% for the bankrupt and 89.7% for the non-bankrupt group of companies). Linear model proved to be superior after validation and “holdout” tests of the bankrupt group, since quadratic model misclassification rate was over fifty percent of future bankrupt companies five year prior. New ZETA model, compared to Altman’s 1968-model, is better in classification of bankruptcy from 2 to 5 years prior, while the initial year’s accuracy is almost the same.

The new optimal cutoff score $ZETA_c$ was defined as[10]:

$$ZETA_c = \ln \frac{q_1 C_I}{q_2 C_{II}}$$

where:

q_1,q_2 = prior probability of bankrupt (q_1) or non-bankrupt (q_2)
C_I,C_{II} = costs of type I and type II errors [Altman, Haldeman, Narayanan 1977: 43]

Efficiency of the ZETA method of classification can be alternatively compared by expected cost of ZETA (EC_{ZETA}):

$$EC_{ZETA} = q_1(M_{12}/N_1)C_I + q_2(M_{21}/N_2)C_{II}$$

where:

M_{12},M_{21} = observed type I and type II errors
N_1,N_2 = number of observation in the bankrupt (N_1) and non-bankrupt (N_2) groups

Authors in their test anticipated the same prior probabilities and the same costs of errors, what resulted in a cutoff score of zero. However potential bias could be involved in such a method. Costs of Type I and Type II error are analogous to previous cases, when type I is accepted loan that defaults and Type II is rejected loan that would pay off. Commercial bank loan function is utilized to specify cost of errors in classification as follows[11]:

$$C_I = 1 - \frac{LLR}{GLL}, \qquad C_{II} = r - i$$

where:

LLR = amount of loan losses recovered
GLL = gross loan losses (charged-off)
r = effective interest rate on the loan,
i = effective opportunity cost for the bank

According to different estimations of Type I and Type II costs as well as probabilities of bankruptcy, various cutoff scores as well as model efficiency can be determined. The

[10] Authors explain advantage of new score as follows: "Note that if one sets equal prior probabilities of group membership, the linear model will result in a cutoff or critical score of zero. All firms scoring above zero are classified as having characteristics similar to the non-bankrupt group and those with negative scores similar to bankrupts. The same zero cutoff score will result if one desired to minimize the total cost of misclassification." [Altman, Haldeman, Narayanan 1977: 43]

development of new ZETA model for bankruptcy classification was quite accurate when right classification one year prior to bankruptcy was over 90% (70% up to five years). Including retailers into a model, did not negatively affected results. ZETA model according to authors outperformed these days alternative bankruptcy classification strategies in terms of expected cost.

1.3.2 Discussion on the financial applications of discriminant analysis

Joy and Tollefson scrutinized discriminant analysis from design and interpretation perspective. The first concern lies in classification of entities with *m* attributes into a priori categories. Authors states: "If the *m* attribute measurements arise from multivariate normal populations such that the categories have identical variance-covariance matrices, but different mean values for the attributes, then linear multiple discriminant analysis (LMDA) provides an optimal solution to the classification problem. When the measurements arise from multivariate normal populations, but the variance-covariance matrices are not identical, quadratic rather than linear multiple discriminant analysis yields the optimal solution." [Joy, Tollefson 1975: 723] Tests to determine stated conditions were not provided in the majority of previous studies.

Criticism of sample design is directed towards inconsistency between population of loan applicants, to which model should be discriminating, and loan acceptances, from which sample was built, including denied loans applicants. Another part of critics contains difference between cross-validation and intertemporal validation tests, since many studies used hold-out sample from the original sample period, but interpreted results as predictive. An assumption of stationarity of population over time supports stated interpretation. Researchers however did not validated model to establish existence of stationarity.

Altman in the paper *Predicting Financial Distress of companies: Revisiting the Z-Score and ZETA® Models* [Altman 2000] alongside with recapitulation of his previous clarifications and recommendations, explained use of Z-score model on private firms[12], since the biggest problem of original model was use of stock price data. Altman states: "Rather than simply insert a proxy variable into an existing model to calculate z-scores, I advocate a complete reestimation of the model, substituting the book values of equity for the market value in

[11] For a more detailed discussion of this investigation see Altman [1977b].
[12] Not publicly traded

X_4[13]." [Altman 2000: 20] Model coefficients subsequently change as well as cut off score and classification criterion. New model is defined as:

$$Z' = 0.717X_1 + 0.847X_2 + 3.107X_3 + 0.420X_4 + 0.998X_5$$

Overall performance of the model is little bit worse than original one, with wider zone of ignorance, lower Type I accuracy (-3%) and lower Z' score (4.14 vs. 5.02). Altman further developed scoring model for emerging markets. Initial analysis is identical with classical analysis used for US companies. After a quantitative modeling and qualitative evaluation of currency, industry or other risk factors can be incorporated to the model. Procedure using bond-rating equivalents and quality of servicing foreign currency bonds can approximate country with the lack of credit rating experience to mature credit markets.

1.4 Logit

One of the first works using stochastic process technique integrating the probability distribution on risk measurement of the bankruptcy was performed by Santomero and Vinso [Santomero and Vinso 1977], as well as Martin [Martin 1977], using directly logit methodology. Their research was executed on over two hundred commercial banks. Banking industry, however, substantially differs from manufacturing industry in capital structure, as well as in operational aspect. Financial ratios used for prediction of bankruptcy are therefore nontransferable to other industry research.

1.4.1 Ohlson

The main findings of Ohlson study can be summarized at first as identifying four basic factors, which were statistically significant in affecting the probability of failure (in one year horizon). These four factors are [Olhson 1980: 110]:

- the size of the company
- measures of the financial structure
- measures of performance
- measures of current liquidity

[13] Market value of equity/book value of total liabilities

Secondly, he is pointing out overstatement of the predictive (forecasting) power of previous models. Main concern is driven by the fact, that previous models employed predictors derived from statements, which were (could be) release after the date of the bankruptcy.

There is one relatively important disadvantage of the model. It does not use any price data of the firms generated by the stock market. However, in the case of non-listed companies this feature can be regarded as an advantage.

Olhson chose conditional logit analysis, compared to multivariated disriminant analysis (MDA) used by i.e. Altman. There are several problem with MDA[14] as certain statistical conditions were set upon the distributional properties. "Variance-covariance matrices of the predictors should be the same for both groups (failed and nonfailed); moreover, a requirement of normally distributed predictors certainly mitigates against the use of dummy independent variables." [Olhson 1980: 112] These shortcomings mainly limit possibility of statistical testing of variables for significance. Another problem of the MDA is a score itself, which is basically an ordinal ranking device, what makes it more difficult for decision problem as a misclassification. MDA also does not allow identification of probability of failure for an individual firm, it rather shows higher or lower likelihood of failure. There are "matching" problems related to data selection procedure, since variables used for matching are quite similar to predictors.

Conditional logit analysis according to author avoids all of the problems associated with MDA. Strong advantage of logit approach is that no assumptions about prior probabilities of bankruptcy or special distribution of predictors have to be imposed.

Definition of failure was purely legal, when failed firms were regarded as ones who filed for bankruptcy procedure under Chapter X or XI, possibly other notification signaling bankruptcy. Data population is of period from 1970 to 1976. Shares of companies were traded eigher publicly or privately (OTC) and companies were categorized as industrial. The final sample consisted of 105 bankrupt firms.

[14] See also Eisenbeis [1977] and Joy and ToUefson [1975] for extensive discussions, cited in Olhson 1980

A Probabilistic Model of Bankruptcy: Logit

The logarithm of the likelihood of potential outcome, which is indicated in the binary sample of bankrupt and nonbankrupt companies is given by:

$$l(\beta) = \sum_{i \in S_1} \log P(X_i, \beta) + \sum_{i \in S_2} \log(1 - P(X_i, \beta))$$

where:

X_i = vector of predictors for the i^{th} observation
β = vector of unknown parameters
$P(X_i, \beta)$ = probability of bankruptcy for any given X_i and β, when $0 \leq P \leq 1$ is some probability function

The maximum likelihood estimate of β_1, β_3, β_4,…, is reached by solving: $max_\beta l(\beta)$

Ratios and basic results

Olhson did not used any new ratios and the criterion for ratios selection was simplicity. Following variables were used, with suggested sign of coefficients:

- SIZE = log[15](total assets/GNP price-level index), GNP=100 for 1968, Negative
- TLTA = Total liabilities divided by total assets, Positive
- WCTA = Working capital divided by total assets, Negative
- CLCA = Current liabilities divided by current assets, Positive
- OENEG[16] = One if total liabilities exceeds total assets, zero otherwise, Indeterminate
- NITA = Net income divided by total assets, Negative
- FUTL = Funds provided by operations divided by total liabilities, Negative
- INTWO = One if net income was negative for the last two years, zero otherwise, Positive
- CHIN = $(NI_t - NI_{t-1})/ (|NI_t| - |NI_{t-1}|)$, NI_t is net income, Negative

Three logit models were estimated with stated predictors. Model 1 predicts bankruptcy within one year, Model 2 within two years (company did not default in subsequent year), Model 3 predicts within one or two years. Likelihood ratio index is used as a measure of goodness of fit.[17] Four statistically significant factors in assessing the probability of failure are size, financial structure reflected by leverage (TLTA), performance measure or combination of

[15] [Olhson 1980, pg. 118]: "The log transform has an important implication. Suppose two firms, A and B, have a balance sheet date in the same year, then the sign of $P_A - P_B$ is independent of the price-level index."
[16] Serves like correction of discontinuity for TLTA, in case of negative book value
[17] 1 – log likelihood at convergence/log likelihood at zero. See McFadden[1973] for further details.

several performance ratios (NITA, FUTL), measure of liquidity or combination of liquidity factors (WCTA, CLCA).

1.5 Probit

Typical research of bankruptcy prediction use nonrandom sample, what can eventually result in parameter bias. Zmijevsky [Zmijevsky 1984] describes two estimation biases from nonrandom sample. The first type of bias arises from *choice-based* sample. The second one is product of "complete data" criterion what can be described as *sample selection* bias. Both of these problems are however natural in financial distress prediction, since there is a low number of financially distressed firms compared to whole firm population. Data for financially distressed companies are either incomplete or unavailable.

Choice-based samples bias results from sampling procedure. Firstly, distressed and healthy companies are identified, and then selection of firms from both populations separately is executed. Such a data sample can be described as nonrandom, since firm occurrence in the sample is based upon financial distress attribute of the company. "For most specifications of selection probability type models (i.e., logit and probit model specifications), the constant and all of the coefficients are asymptotically biased." [Zmijevsky 1984: 65] As an appropriate solution for choice-based sample bias, author suggests weighted exogenous sample maximum likelihood model (WESML), which weights the log-likelihood function by the proportion of the population and sample frequency rate of separate groups:

$$L = \left[\frac{POP}{SAMP}\right]\sum_{i}(B)\ln[\Phi(H)] + \left[\frac{1-POP}{1-SAMP}\right]\sum_{i}(1-B)\ln[1-\Phi(H)]$$

where:

POP = the proportion of bankrupt firms in the population
SAMP = the proportion of bankrupt firms in the sample
B = 1 if bankrupt, 0 otherwise
ROA = net income to total assets
FINL = total debt to total assets
LIQ = current assets to current liabilities
Φ = cumulative density function for a standard normal variable
$H = \alpha_0 + \alpha_1 ROA + \alpha_2 FINL + \alpha_3 LIQ$

Reduction in the bias with unweighted probit was recognized as probability of bankruptcy in the sample approached probability in the population. Zmijevsky reports that "firms with high

bankruptcy probabilities have low probabilities of having complete data. Thus, it is those firms with high bankruptcy probabilities that are excluded from estimating the model." [Zmijevsky 1984: 77]

1.6 Other important writings

Kallber and Udell investigate effect of private information sharing and exchange on quality of business payments. Empirical examination is performed on the lending decision problem. "Findings indicate that exchange-generated information provides significant explanatory power in failure prediction models controlling for other credit information that is easily available to lenders." [Kallber, Udell 2003: 449]

Three questions were examined by the authors. At first, wheter sharing of credit information is beneficial mechanism for overcoming information asymmetry problem when lending. The second: Is information provided by private exchange more efficient than otherwise collection of publicly available information? An finaly: Can private information exchange enhance credibility of information intermediary. In case private information exchange can be helpful, or even enhance predicting distress of companies, it could bring significant value. These questions were examined on business exchange-generated information[18]. "Our methodological approach is to examine whether D&'B's[19] payment experience information adds power in a failure prediction model controlling for information contained in standard financial statements and other sources available to creditors." [Kallber, Udell 2003: 459]

Authors used ratios derived from balance sheet. Sample consisted of 241 failed and 2482 non-failed companies, with failure rate 8.8%. Two groups of data in the sample were used to examine ability to discriminate between failed and non-failed firms. In the first group, payment information was included in form of PAYDEX[20]. Dummy variable PAYDUM representing payment information was created in case this information is not available for company. The second group of data consists of financial statements and other descriptive information about companies.

Logistic regression with dependent variable being probability of the firm non-failing was used with following results [Kallberg, Udell, 2003: 464]:

[18] Consumer exchange generated information not included

[19] World's largem priváte informatik exchange. D&B possess informatik on over 70 mil. Businesses in over 200 countries. More on http://www.dnb.com

[20] Summarizes the payment history information based on D&B's data

Table no.3 - Failure prediction model (Panel A: Model estimates)

Parameter	Estimate	T- statistics
Constant	-3.213	-6.368
Paydex	0.075	12.593
Paydum	5.737	8.519
Leverage	-0.116	-4.658
Quick	0.246	2.782
NLB	1.056	2.454
Neglife	-1.987	-5.302
Collat	-0.503	-2.442
Age	0.046	3.894
Sic59	0.625	2.896

Source: Kallberg, Udell, 2003: 464

Model $\chi^2 = 447.7$

Log likelihood = -373.6

AIC = 382.6

Pseudo $R^2 = 0.736$

Results show that information about payment ability from D&B has significant predicting power in company failure. Authors argue that payment information from information exchanges possess significant value comparing to other sources. Moreover results indirectly show that private information exchanges can enhance credibility problems of financial intermediaries.

Vallini [Vallini et al. 2008] looked at the smaal enterprises model estimation problem. Small enterprises have generally face less requirements for data disclosure compared to larger companies, not even mentioning listed ones. Moreover small enterprises´ data, made available, have tendency to be less accurate and less reliable, since external control is less likely, compared to professional analysts, auditors, etc. of large companies. Moreover small business data are harder to interpret due to relatively high specificity.

Data sample consisted of 3 063 failed firms from CERVED database, containing accounts for all Italian limited companies operating in the manufacturing, building and service industry

collected by Chambers of Commerce. Default was defined traditionally in time of formal legal proceeding start, for example companies which became insolvent in 2005.

Control group was created of 3 050 non-defaulted companies using stratified random sampling to achieve same sample structure for control group as for defaulted firms group concerning selected criterias of: size[21], geographical location, business sector. Whole dataset consisted of 6 113 companies. Since 75.2% of companies did not exceeded 1.8 million euros turnover, they can be classified as small enterprises.

Risk predictor variables were chosen on the basis of two criteria:

- Frequency in the research literature
- Ability to describe essential aspects as profitability, leverage and liquidity

Initial set was made from 23 variables supposed to have significant impact as default predictors. Using multicollinearity analysis followed by the variable-reduction stepwise method ratios were reduced to ten:

- cash flow/ total debts
- total debts/ (total debts+equity)
- acid test ratio
- interest charges/
- current ratio
- equity/ long-term materiál assets
- ROI
- net financial position/ turover
- long-term assets/ number of employees
- interest charges/ bank loans

Logistic regression was used on the same sample and with the same economic and financial ratios as in MDA analysis. Overall accuracy 67.2% with 42% Type II and 23.6% Type I errors outperformed MDA. Prediction accuracy for individual groups is higher in case of size (68.5%) and geographical area (68.4%). Regarding business sector analysis with overall accuracy of 67.4%, we can talk about the same effect as in case of overall model. Results in size group individualy followed pattern from MDA with higher accuracy with larger companies. For other individual groups results were comparable in terms of pattern.

From other conclusions, lower prediction capacity of smaller sized firms should be noted, with 9% difference compared to largest group. Therefore assumption of lower quality and reliability of the smaller firms reports can be affirmed. Different results were recorded for different regions in Italy, what offers application for banks in credit evaluation. Finally authors state that: “Above all, decisional functions should be based on a reasonably

homogeneous sample. Pooling different business sectors or geographical areas tends to reduce a model's prediction accuracy." [Vallini et al. 2008: 21]

2. Metodolody

2.1 Logit, Probit

A basic task of economic models or hypotheses is to understand observed behavior or system and find out relations between underlying processes. Such an effort is rather complicated, since one is not able to control and observes whole scale of factors interfering with processes and also because processes themselves influence behavior or action through the instruments of experience and expectations. Model between individual entity and population of individuals' data sample is built to catch statistical inference between them.

Special type of model is the one with qualitative alternatives. McFadden [McFadden 1973] outlined a general procedure for econometric models´ formulation, which describes population choice behavior from distribution of individual decision-making rules.

In finance, default of a client on loan can be expressed by dependent variable, as dummy variable with values 1 and 0, where one of them means default. Difference from classical models is that binary variable appears as explained variable on the left hand side of the equation. Probabilistic interpretation is being used, meaning quantification of relation between individual values of explanatory variables and probability of occurrence of event described by binary variable.

Linear regression model is not suitable for predicting probability, since probability of default could exceed interval <0,1>, therefore normalization to desired range is required. The second problem of linear model would be in homoscedasticity assumption, which is usually violated in real world.

Let y_t be explanatory variable with binary values: 1 when event occurred and 0 when not.

According to Cipra [Cipra 2008], in probabilistic notation one can write:

$$P(y_t = 1 | x_{t\cdot}, \beta) = 1 - F(-x_{t\cdot}\beta), \qquad t = 1, \dots, T$$

where F(·) would be appropriate probabilistic distribution function, with values from interval (0;1).

---

[21] 2001 turnover

Regression model comes from simple linear model:

$$y_t = x_t.\beta + \varepsilon_t$$

where: ε_t are independent and identically distributed *(iid)* random variables with:

$$E(y_t|x_t.,\beta) = 0, \quad var(\varepsilon_t|x_t.,\beta) = F(-x_t.\beta)[1 - F(-x_t.\beta)]$$

Interpretation of parameters β_i, are different than in a case of classical linear model, where parameter can be explained as marginal effect on explained variable. In our case, i tis possible to write:

$$\frac{\partial E(y_t|x_t.,\beta)}{\partial x_{ti}} = \frac{P(y_t = 1|x_t.,\beta)}{\partial x_{ti}} = f(-x_t.\beta).\beta_i$$

where: f(·) is probability density of distribution function F(·). Therefore:

$$\frac{\partial E(y_t|x_t.,\beta)/\partial x_{ti}}{\partial E(y_t|x_t.,\beta)/\partial x_{tj}} = \frac{\beta_i}{\beta_j}$$

What implies that ratio of marginal effects of regressors, can be found as a ratio of parameters corresponding to the regressors. For numerical estimate in practice for F(·) only several probabilistic distributions are used. Probit built upon distribution function Φ(·) of normal distribution N(0,1):

$$P(y_t = 1|x_t.,\beta) = 1 - F(-x_t.\beta) = 1 - \Phi(-x_t.\beta) = \Phi(x_t.\beta)$$

And Logit built upon logistic distribution:

$$P(y_t = 1|x_t.,\beta) = 1 - F(-x_t.\beta) = 1 - \frac{e^{-x_t.\beta}}{1 + e^{-x_t.\beta}} = \frac{e^{x_t.\beta}}{1 + e^{x_t.\beta}}$$

with density of probabilistic distribution $f(x)=e^x/(1+ e^x)^2$

Estimation of the model with binary variable is done by maximum likelihood (ML) method. After logarithmic transformation likelihood function looks like:

$$L(\beta) = \sum_{t=1}^{T} y_t \ln[F(x_t.\beta)] + \sum_{t=1}^{T} (1 - y_t) \ln[1 - F(x_t.\beta)] \hat{\beta}$$

Maximum likelihood estimation of model parameters $\hat{\beta}$ can be obtained with appropriate algorithmic maximization of function L(β). "These two (logit and probit) cumulative distribution functions (c.d.f.) differ only in the tails, and the logit resembles the c.d.f. of a t-distribution with 7 degrees of freedom, whereas the probit is the normal c.d.f., or that of a t with ∞ degrees of freedom. Therefore, these two forms will give similar predictions unless there are an extreme number of observations in the tails." [Balthagi 2008: 324]

2.2 Goodness of fit

Quality of estimated models with binary explained variable is usually provided with McFadden's R^2 coefficient based on likelihood ratio and is analogical to coefficient of determination. According to McFadden: "One can define an analog of the multiple-correlation coefficient[22]:

$$R^2 = 1 - \frac{L}{L_R}$$

where: L is maximal value of function L(β), defined above, and L_R is defined with the same equation, where the numerators of the residuals are evaluated under the hypothesis that the parameter vector is zero, or is zero except for pure alternative choice effects. [McFadden 1973: 122]

Log likelihood, also called *Deviance,* is used in the logistic regression estimation as a criterion for assessing parameters for inclusion into the model. Statistical software usually states metrics multiplied by -2 because of hypothesis testing purposes. Higher values signalize worse prediction of explained variable. Ratio can be simply defined as:

$$D = -2 \ln \left[\frac{likelihood\ of\ the\ fitted\ model}{likelihood\ of\ the\ saturated\ model} \right]$$

where: saturated model use as many parameters as data points. In case of a binary outcome variable[23], as in our case, the saturated likelihood becomes 1, therefore statistics is simplified to:

[22] Denotation was changed from McFadden to be consistent with previous equations.
[23] Either 0 or 1

$$D = -2ln[likelihood\ of\ the\ fitted\ model]$$

Log likelihood is also a base for another measure for assessing model, more specifically assessment significance of independent variable inclusion into the model. Statistics G defined as:

$$G = -2\ln\left[\frac{likelihood\ without\ the\ variable}{likelihood\ with\ the\ variable}\right] =$$

$$= D(model\ without\ the\ variable) - D(model\ with\ the\ variable)$$

Statistics follow the χ^2 distribution with degrees of freedom depending on numbers of compared models' variables. In case one ratio is added (excluded), statistics have one degree of freedom.

Person's R for contingency tables can be used as well, when employing successful estimates (both positive and negative) combining with Type I and Type II errors. Ratio can reach values form -1 to +1, where higher (closer to +1) means more successful estimations.

Table no.4 – Accuracy Matrix

	Predicted Group	**Membership**
Actual Group Membership	Bankrupt	Non-Bankrupt
Bankrupt	a	b
Non-Bankrupt	c	d

where:
a & d – correctly classified companies, (H as a Hits)
b – Type I error
c – Type II error

This Pearson's R indice of predictive efficiency is calculated as:

$$r = \frac{(ad - bc)}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

Sensitivity of the model is defined as a/(a+c), and specificity as d/(b+d) ROC curve, more exactly area under the ROC curve is an ndicator from same category as Pearson's R. Area ranges from zero to one. One means the best ability to discriminate between subjects, with

positive and negative outcome. The curve generated by all possible cutpoints of sensitivity vs 1 – specificity is called ROC Curve. Values of an area over 0,8 are considered as excellent discrimination [Hosmer Lemeshow 2000].

The Wald test, commonly represented in statistical programs in column headed *z*, can be estimated when maximum likelihood estimation of the slope parameter is compared to its standard error:

$$W = \frac{\hat{\beta}_1}{\widehat{SE}(\hat{\beta}_1)}$$

Under the null hypothesis that $\beta_1=0$, obtained ratio will follow normal distribution and can be simply used for variable selection by two tailed p-value.

Additional criteria for evaluation of models can be Akaike's Information Criterion (AIC) and Bayesian information criterion (BIC) also called Schwarz criterion. Both of these measures are bases on the likelihood function. Log likelihood grows with adding parameters to the model, what if used carelessly can result in model overfitting. Both criteria offer assistance, when penalizing for adding more variables to the model. Generally, when choosing from multiple models with variety of indicators, lower Akaike's or Schwarz criterion means better model.

Akaike's criterion: $AIC = -2(log - likelihood) + 2k$

where: k indicate number of parameters in the model including intercept.

Schwarz criterion: $BIC = -2lnL + k\ ln(n)$

where:
L – maximized likelihood function
n – number of observations in data set

2.3 Data collection

Data were collected through MagnusWeb database provided by Bisnode Czech Republic[24], leading provider of economical, financial, credit and trade information in Czech Republic. Magnusweb database contains information about Czech economic subjects. It is possible to look up identifications of company according to name, legal form, identification number[25], address, date of establishment of a company, quantitative characteristics such as amount of registered capital, number of employees, turnover, industry specification by NACE[26] coding, indication of negative events as executions or insolvency, finally financial statements and many others.

Default definition is essential for model specification. Available data basically determine such a definition. Two relevant possibilities were available. The first one is date of default on payments. Generally speaking, it is a situation when company has delayed payments, usually for invoice or interest. The second option is legal start date of insolvency procedure.

Legal start date of insolvency, which seems more reasonable date for default definition, is based on available data. First of all, default on payment was not clearly defined by data provider and also methodology, for such a classification, was missing in data specification. Claim of default can be proposed by any trade partner of company under suspicion of default on payment. Such a claim can be speculative with aim to damage competitor's reputation, or can be just a result of trade conflict, for example delayed payment because of reclamation of delivered goods. Finally, one can suppose that state of default on payments can be set rather arbitrarily, compared to official insolvency procedure.

Filling for insolvency requires documenting (by petitioner) various creditors whose credit is late, specifically at least 30 days after due date and also requires proving that debtor is not able to pay for the debt.[27] Insolvency is therefore stronger definition of default compared to previous case, since it requires multiple unpaid liabilities as well as multiple creditors, what undermines arbitrariness of default definition. Debtor is also able to defend himself against unreasonable insolvency procedure, and ultimately demand compensation from petitioner. Such an option reduces motivation for vexatious insolvency petitions from creditors' side simultaneously offering argument for preferring insolvency as more relevant definition compared to a default on payment.

[24] Bisnode Česká republika, a.s.
[25] Generall number (IČ) as well as tax identification number (DIČ)
[26] Nomenclature générale des Activités économiques dans les Communautés Européennes
[27] §3 of Czech Insolvency Law

Companies in insolvency procedure and available financial statements were selected from MagnusWeb database. Such companies were extracted in two steps. At first active companies in insolvency[28] were filtered. Date of last financial report and start of insolvency had to be validated in order to include only companies with available financial statements at least 2 years before insolvency petition. List of companies was exported from database, including following information: registered trade name, indicator of negative events, identification number[29], city, date of establishment, telephone number, fax, mobile phone number, e-mail, web address, legal form, stage of insolvency, locality, registered capital, number of employees, category of number of employees, category of turnover, main NACE code and description, year of last financial report, court were company is registered, executive officer, chairman of the board, limited partner, active partner, general director, financial director, trade director economical director, marketing director, human resources director, full name of statutory organ, role in statutory organ, full name of employees representative and his position in company and finally start date of insolvency procedure.

Second stage represented downloading financial statements for selected companies. Financial reports included generally balance sheet and income statement under Czech accounting standards[30]. Various level of detail of reports was available, but most commonly reported and at the same time sufficient for this analysis was level 3 (out of 5) including information in Appendix. Numbers were rounded to thousands. Since various information were collected concerning companies' features at the date of data collection, assumption that such an information has not changed for two previous years was applied.[31]

Initial sample of companies in insolvency consisted of 554 firms. However, not all data did comply with qualitative criteria for analysis. The most usual problem was that an initial date of insolvency was more than two years older than last financial report. Another problem withstood from lack of available financial reports, when at least two subsequent statements prior to insolvency were required.

At first companies, with financial reports maximally 1120 days older than start day of insolvency procedure, were selected. That means three latest statements plus 40 days as correction for possible later availability of report. Even though company reports closing the books to the end of a calendar year, final reports can be available with some delay, as some preparation time is needed. Potential creditor, assessing credit risk, can therefore presuppose

[28] Having financial reports available
[29] IČ
[30] CAS from 2003 for entrepreneurs

little bit longer prediction period than exactly a year. From selected companies, firms with at least 2 consecutive financial statements were chosen and simultaneously firms with filled incomes statements, since some companies filled just zeros to all entries in P&L. Finally 228 companies were selected with 623 entries in data sheet.

Matching group to defaulted companies was subsequently formed. Matching criteria were limited by possibilities of search options in employed database system. Chosen input options allowed to filter companies according to legal form, category of numbers of employees, category of turnover, range of date of establishment, NACE code, and assets range. Filter options also offered alternative to choose only companies without any negative event in the past, what exactly corresponded to the needs of creation of matching group.

For every one of 228 insolvent companies, firms in the same industry were found. Cited matching criteria were prioritized after initial formation of trial non-defaulted group. Top priority category was industry according to NACE code, with maximum deviation to one level higher category selection. Fortunately, almost in all cases exactly same NACE level could be found. The second most important matching criterion was a range of assets. Asset size is basic indicator of company size. Effect of the size of the company on default rate was clearly demonstrated by basically every study in this area. Date of establishment was another measure taken into consideration, since age of the company was mention almost in every study, which accounts for this element as a very significant variable. Category of turnover was initially next priority. However after verification of initial trial sample, this criterion was dismissed, as turnover did not always corresponded to provided income statement numbers. Optional additional criteria, in case when many companies were available after use of priority criteria, legal form and number of employees suited as good pattern to approximate companies to defaulted ones. Finally companies with newest financial report available were preferred.

Even though Czech Republic industry as well as number of companies in MagnusWeb database is limited, even in some specific industries and especially companies with high amount of assets at least two matching firms could be found. In 86% of cases at least three companies were matched to defaulted ones and just for two companies only one suitable counterpart was found. Altogether dataset with 831 non-defaulted companies was composed.

Non-defaulted companies' entries were filtered with two major constraints. Only five latest financial statements, the latest being dated maximally to 2008, were included in dataset. The

[31] Information as companies' location, contact information, registered capital, industry, etc.

earliest date of start of bankruptcy proceeding, included in dataset, is 18th of January 2010, and according to criteria for defaulted companies, the oldest financial statement is from year 2008. Criterion for oldest financial statement is hence self-explaining, as entries with older financial statement would not have defaulted counterpart. In case of defaulted companies, assumption of invariability for four years of stated company specific information was applied. In case of non-defaulted companies, one can suppose even longer period than in case of defaulted, since no negative events were recognized in the firm's history. Therefore there is lower likelihood of changing major information, or more generally scope of provided information, which as show later is more important for model than exact entry itself.

Macroeconomic indicators were collected from Czech statistical office. Unemployment and employment rates were collected from seasonally-adjusted monthly data and averaged for years[32]. Yearly unemployment rates for individual regions of Czech Republic were available[33] as well, and subsequently included into dataset. Location of company was first matched to region according to main activity filled in MagnusWeb database. Afterwards, regional unemployment figures were matched to regions and year of financial statement.

For inflation, the same source and procedure was used[34]. From inflation rate, as an increase in CPI compared with the corresponding month of preceding year, yearly averages were calculated. Resulting numbers correspond to official yearly inflation rates published with Czech statistical office.[35] Real yearly GDP growth figures were collected from cross-sectional publication of main macroeconomic outlook published by Statistical Office.[36]

Interest rates data were acquired from Czech National Bank ARAD Database[37]. Interest rates per annum for newly closed deposits and loans contracts for non-financial institutions were averaged from monthly data. Percentage of bad loans of non-financial firms monthly data[38] were collected from ARAD database as well. Yearly averages were subsequently computed and included to the datasheet. If available, monthly data were preferred to yearly figures, because of potential use in model verification and testing for companies bankrupted in 2013.

---

[32] http://www.czso.cz/csu/csu.nsf/informace/cnez070113.doc

[33] http://vdb.czso.cz/vdbvo/tabdetail.jsp?kapitola_id=15&potvrd=Zobrazit+tabulku&cas_1_21=2007&go_zobraz=1&cislotab=VSPS+507_ro%C4%8Dn%C3%AD&voa=tabulka&str=tabdetail.jsp

[34] http://www.czso.cz/eng/redakce.nsf/i/inflation_rate

[35] For first 6 months of year 2013

[36] www.czso.cz/eng/redakce.nsf/i/macroeconomic_indicators/$File/AHLMAKRO.xls

[37] http://www.cnb.cz/cnb/STAT.ARADY_PKG.PARAMETRY_SESTAVY?p_sestuid=17816&p_strid=AD&p_lang=CS

[38] http://www.cnb.cz/cnb/STAT.ARADY_PKG.PARAMETRY_SESTAVY?p_sestuid=12119&p_strid=AD&p_lang=CS

2.4 Financial indicators

Financial ratios can be described as proportions between chosen values coming from a firm's reported financial statements. The main objective of financial indicators is dependent on the point of analyst´s view. Principal purpose of ratios for shareholders is to evaluate and measure company's performance in terms of profitability (short-term and long-term). Managers also use ratios for estimating operational performance. Both of these objectives are also concerning financial analysts and traders. Creditors' approach to usage of financial ratios is to assess company's capacity to sustain certain level of debt as well as ability to pay interest and repay a principal of granted loans. Creditors' point of view is the closest to the aim of this paper, since default of company in form of bankruptcy significantly undermines firm's ability to fulfill obligations from loan contract. However, not only creditors can be interested in company's default, since it affects also trade partners, stakeholders and more broadly through spillover effect, also society as such. One should not disregard also perspective of profitability and effectivity, since operational problems are generally main reasons leading to firm deterioration.

Countless financial indicators can be constructed from financial statements inputs. Since there are many possible variables with different purpose, classification into four categories helps focus on main objective of the ratios. Liquidity, solvency and leverage, profitability and activity categories of ratios can be formed from financial statements. If inputs of market data are available, investment ratios can be constructed as well.

Initial set of financial indicators in previous studies was generally composed from the most popularly used ratios in previous studies. Altman [Altman 1968] originally selected twenty two potentially working variables according to popularity in literature and potential relevancy to study. Olhson employing Logit model, used directly nine ratios, arguing that "criterion for choosing between different predictors was simplicity" [Olhson 1990: 118]. From more recent studies, Vallini [Vallini et al. 2008: 9] initially determined twenty three financial ratios on the basis of two criteria:

- their frequency in the research literature on company default prediction[39]
- their ability to describe essential aspects of three areas of company economic and financial profile; namely: profitability, leverage, and liquidity.

Table no.5 – List of financial indicators

Notation	Name of Ratio	Definition	Exp. Effect
Liquidity Ratios			
CuR	Current ratio	Current assets/ Current Liabilities	-
QR	Quick ratio	(Cash + Short term receivables)/ Current Liabilities	-
CaR	Cash ratio	(Cash & Equivalents)/Current Liabilities	-
WC	Working capital	(Current Liabilities - Current Assets)/Assets	-
Solvency Ratios			
CapR	Capitalization ratio	Long Term Liabilities /(Long Term Liabilities + Equity)	+
LevI	Leverage I	Debt/Equity	+
LevII	Leverage II	Debt/Total Assets	+
ICR	Interest coverage ratio	EBIT/Interest Expenses	-
ICRII	Interest coverage ratio II	EBIT/(Interest Expenses + Other Financial Costs)	-
CFR	Cash flow I	Cash Flow/ (Debt)	-
CFRII	Cash flow II	Cash flow/ (Debt/ 365)	-
CFRIII	Cash flow III	Cash flow / [(Debt-Reserves)/365]	-
RetER	Retained earning ratio	Retained Earnings from Previous Periods/Assets	-
RetERII	Retained earning ratio II	(Retained Earnings from Previous & Actual Periods)/Assets	-
Profitability Ratios			
Gprof	Gross profit	Operating profit/Revenues	-
EBITDAm	EBITDA margin	EBITDA/Revenues	-
ROA	ROA	Net Income/Total Assets	-
ROE	ROE	Net Income/ Equity	-
Nprof	Net Profit Margin	Net income/Revenues	-
Activity Ratios			
DIO	DIO	Inventories/(Cost of Goods Sold/365)	+
DSO	DSO	Receivables/(Sales/365)	+
DPO	DPO	Payables/(Sales/365)	+
CCC	CCC	DIO+DSO-DPO	+
SLT	sales turnover	Sales/Assets	-

Source: own calculations

Basically the same procedure was used by Jakubik and Teply [Jakubik, Teply 2008: 7] initially determining twenty two financial ratios. According to the same procedure of popularity in previous research and an ability to describe important features of company's financial profile, set of financial ratios was determined.

2.4.1 Liquidity ratios

The first category consists of liquidity ratios, which assess company's ability to meet short-term obligations. The most liquid assets are extracted from balance sheet and compared to other variable, usually of a kind of short-term liabilities. Liquidity of assets is established according to their relative ability to convert into cash. Various ratios can be constructed, when different approach to current assets is employed according to degree of assets convertibility into cash.

[39] (e.g., Altman, 1968, 1993; Altman, Brady, Resti, & Sironi, 2005; Altman, Haldeman, & Narayanan, 1977; Altman, & Sabato, 2005, 2006; Beaver, 1967; Blum, 1969, 1974; Crouhy, Mark, & Galai, 2001; Edmister, 1972);

- **Current ratio** specifies proportion between current assets, which are available for covering current/ short-term liabilities. Current or short-term assets include cash & cash equivalents, marketable securities, short-term receivables and inventory.
- **Quick ratio** also called acid-test ratio uses, compared to previous case, only the most liquid assets, comparing them to current liabilities. Inventory and other current assets are excluded from the ratio, since their convertibility into cash is not that straightforward and fast.
- **Cash ratio** contains even more liquid assets, including cash, cash equivalents and financial assets.
- **Working capital** is composed of working capital, defined as current assets minus current liabilities over total assets.

2.4.2 Solvency and leverage ratios

Solvency ratios are second category of financial ratios expressing ability of firm to meet its long-term obligations. Proportion of debt plays significant role in assessing risk of bankruptcy, since debt predispose overall financial risk, which firm and its owners face. Practically it can be stated, that the higher the debt proportion, the higher the risk of bankruptcy. The main channels of debt risk are interest volatility and agency problem.

- **Capitalization ratio** captures debt part of a firm's capital structure. Long-term liabilities are set against long-term liabilities plus equity. Prolonged perspective of this ratio helps understand company's position in terms of sustaining its operations and potential growth.
- **Leverage ratio I** compares total liabilities and its total equity. Ratio describes basic position between internal and external financing from the point of ownership view. Higher leverage enhance agency problem between creditors and shareholders, since from the same projects, shareholders could relatively profit more than debt holders, when risk mitigation falls through potential default majorly on creditors.
- **Leverage ratio II** balance firm's total liabilities to its overall assets, defining company's capital structure. Lower values signalize lower levels of debt employed by a company, hence lower dependence on external financing. Generally speaking the higher the ratio, the higher the probability of insovency

- **Interest coverage ratio** describes with what effort a company is able to pay interest expenses. Ratio consists of earnings before interest and taxes (EBIT)[40] over firm's interest expenses. In case the ratio is low, company may have problems paying interest and effectively could default on debt obligation which magnifies bankruptcy risk.
- **Interest coverage ratio II** differs from previous ratio only in definition of EBIT and interest expenses, since according to Czech financial statements another solution can be proposed. Account of other financial costs contains generally fees for bank accounts and loans, which could be broadly included into interest expenses. They are directly connected to financial services provided by bank and their nonpayment can lead to breaching of contractual terms and conditions resulting in factual default on debt contract. Other financial costs are therefore imputed to EBIT as well as to interest expenses. Supposed effect remains the same as in previous case.
- **Cash flow ratio I** is another way how to approach solvency measurement. Distinction of this indicator is in application of cash flow, which was derived from net income and adjusted for non-cash changes in income statement such as depreciation and amortization, value adjustments, revenues and costs from overestimation of commercial papers. Cash position as well as operational factors are included, what offers contrasting perspective to application of EBIT.
- **Cash flow ratio II & III** are variations to original ratio. However, turnover approach is used to diversify ratios spectrum. Turnover approach also interprets ratio in time dimension. Indicator II is a turnover version of previous measure, and number III is moreover adjusted debt for reserves, since reserves can be perceived as negative debt – can be used to repay part of debt.
- **Retained earnings ratio I & II.** Since retained earnings compose majority of equity, past profitability enters solvency ratio, when compared to total assets. Retained earnings are balance sheet inputs, and are accounted as retained earnings from the previous period plus net income minus dividends. While ratio I. accounts only for retained earnings from the previous periods, the second one adds also retained earnings from the actual period.

[40] From Czech financial statements it is derived by adding tax and interest expenses to net income.

2.4.3 Profitability ratios

Profitability ratios are a class of a metrics which help estimate firm's performance in terms of profit generation compared to sources used to create earnings from various angles. Higher values mean generally better performance and lower probability of bankruptcy. Profit figures should, however, theoretically exceed cost of capital. From opportunity costs´ view, capital requirement level higher than profit margins effectively leads to value destruction. Low figures of profit ratios for longer period can therefore cause foreclosure of firm as well. Ultimately, overall effect of profitability ratios should be negative on bankruptcy likelihood; however breakeven point can be different than zero.

- **Gross profit ratio** is one of a profitability measures showing the gross profit as a portion of revenues. It is suitable to evaluate the operational performance of the company.
- **EBITDA margin** is close to gross profit ratios. Personal, SG&A and other operational costs are added to gross profit to evaluate operational profitability on different stage.
- **Net Profit Margin** ratio of profitability is computed as net income over revenues, measuring how much out of every koruna of sales reaches it to earnings. Net sales include all incurred costs and deductions, therefore captures overall profitability levels, not just operational.
- **Return on Assets (ROA)** is financial ratio indicating how profitable a company is in relation to assets. ROA offers an idea about management efficiency in employing assets to generate income. Firm's annual earnings are divided by its total assets. Advantage of ROA ratio is that it disregards capital structure of a company hence captures better return on investment.
- **Return on Equity (ROE)** is the share of net income proportioned to equity. ROE measures a company's profitability from investors' point of view. Comparison using return on equity can be, however, distorted by different capital structure of a company. Higher indebted companies can have better ratio figures using leverage effect.

2.4.4 Activity ratios

Activity ratios are accounting indicators which assess a company's ability to transform different accounts from balance sheet to sales. Capitalization of specific assets, leverage or different balance sheet components is being assessed, as company's ability to turn them into cash relatively fast leads to higher sales through better efficiency, what should have negative effect on bankruptcy probability.

- **Days inventory outstanding (DIO)** provides estimation of a firms' performance of how long it takes to turn its inventory into revenues. Practically, the lower (shorter) the ratio, the better.
- **Days sales outstanding (DSO)** describes number of days it takes for a company to collect revenue after a sale has been made. A high DSO figure signalizes problems, since it takes longer to collect money from customers, what can put pressure on cash flow available for meeting obligations to counterparties.
- **Days payable outstanding (DPO)** indicates a length of period a firm is taking till paying back to trade creditors. Effect of this ratio can be ambiguous. On one hand higher number means more time available for repayment, what reduces cash demand, but at the same time it enlarges payables and could sent message that company has problems to pay back in reasonable time.
- **Cash conversion cycle (CCC)** is the amount of time needed by a firm to convert resource inputs into cash flows. The shorter the length of the cycle, the lower tied-up time of the capital in business operations with positive effect on bottom line.
- **Asset turnover ratio** determines a company's ability to its assets effectively to generate revenues. Higher values signalize higher efficiency.

2.4.5 Correlation analysis of financial indicators

Since many financial indicators employ the same variables, high correlation is supposed a priory, mainly in same ratio groups, especially for liquidity ratios and cash flow employing ratios. Correlation analysis affirmed high correlation for following selected pairs of variables[41].

[41] Since table would be too large, rows and columns for variables without problematic correlation were excluded

Table no. 6 - Correlation between variables

	CuR	*QR*	*WCR*	*LevII*	*CFRI*	*CFRII*	*RER*	*RERII*	*EBITDAm*	*ROA*	*DSO*
QR	COR	1	-	-	-	-	-	-	-	-	-
CaR	COR	COR	-	-	-	-	-	-	-	-	-
LevII	-	-	COR	1	-	-	-	-	-	-	-
CFRII	-	-	-	-	COR	1	-	-	-	-	-
CFRIII	-	-	-	-	COR	COR	-	-	-	-	-
RER	-	-	COR	COR	-	-	1	-	-	-	-
RERII	-	-	COR	COR	-	-	COR	1	-	-	-
ROA	-	-	COR	COR	-	-	COR	COR	-	1	-
NetPm	-	-	-	-	-	-	-	-	COR	-	-
CCC	-	-	-	-	-	-	-	-	-	-	COR
SLTUR	-	-	COR	COR	-	-	COR	COR	-	COR	-

Abs. correlation	
COR	>0,8
COR	>0,95
COR	>0,99

Source: own calculation

Only one of highly correlated variables should be subsequently employed in model estimation as it can be problematic to differentiate between the effect of variables if they are highly correlated. Almost perfect correlation[42] was detected between Current ratio and quick ratio, subsequently for Cash flow ratio I and Cash flow ratio II.

2.5 Macroeconomic indicators

Macroeconomic indicators are supposed to have effect on bankruptcy, since external economic conditions determine performance of companies as well. Jakubik examined macroeconomic determinants on executions and bankruptcies in Czech Republic and states: An empirical analysis shows predictive ability of yearly growth rates of GDP, interest rates, inflation rates and indebtedness levels of enterprises on bankruptcies development in Czech Republic. [Jakubik 2007]

Virolainen [Virolainen 2004] focused on macro stress testing with a macroeconomic credit risk model for Finland and claims that the results suggest a considerable relationship between corporate sector default rates and key macroeconomic factors such as GDP, interest rates and corporate indebtedness. Also Liu [Liu 2004] concluded that macroeconomic indicators have effect on bankruptcies.

Employment rate should have a negative effect, since the higher the employment in economy, the better overall conditions can be supposed, therefore lower bankruptcy rates. **Unemployment rates** have an opposite effect. Reason behind inclusion of both, employment and unemployment rates is rather methodological, since unemployment can be biased by number of people who were unregistered from labor office and simultaneously are not

[42] Higher than 0.995 in absolute terms

included into work force. Also figures for **regional unemployment** were included according to locality of main activity of the company.

As Wadhwani states**: "**In the absence of index-linked loans, higher **inflation** implies higher liquidation rates and default premia**."** [Wadhwani 1986: 120] A cause of this effect can be higher growth of nominal interest rates (costs), compared to growth of revenues. The same problem is caused by growth of **interest rates on loans** which enlarges interest expenses and for highly leveraged companies can have liquidating effect.

Gross domestic product (GDP) captures overall economic condition of the economy. Growth of GDP should have negative effect on bankruptcy probability. Growth of **bad loans**, on the contrary, signalizes problems of firms with repayment of their financial obligations. **Interest rates on deposits** can have similar effect as bad loans, since higher deposit rates can signalize insufficient liquidity in financial institutions as well as constraining of monetary policy leading to higher defaults.

2.5.1 Correlation analysis of macroeconomic indicators

In case of macroeconomic indicators, no correlation above 0.99 was detected, however in eight cases correlation over 0.95 was estimated. Variable of bad loans one year lag was the most often correlated and relation with six variables was estimated. Macroeconomic indicators are generally correlated in higher count compared to financial indicators, since a lot of them are just different expressions of same principle, or directly affect each other through financial system or real economy.

Table no. 7 - Correlation between variables

	u	u-1	i	i-1	rdep	rdep-1	rlon	rlon-1	gdp	badlon	badlon-1	ureg-1
e	-	-	COR	-	-	-	-	-	-	-	-	-
e-1	-	COR	-	COR	-	COR	-	COR	-	-	COR	-
u	1	-	COR	-	COR	-	COR	-	-	COR	-	-
u-1	-	1	-	COR	-	COR	-	COR	-	-	COR	-
i	-	-	1	-	COR	-	-	-	-	-	-	-
i-1	-	-	-	1	-	COR	-	COR	COR	-	-	-
rdep	-	-	-	-	1	-	COR	-	-	COR	-	-
rdep-1	-	-	-	-	-	1	-	COR	-	-	COR	-
rlon	-	-	-	-	-	-	1	-	-	COR	COR	-
rlon-1	-	-	-	-	-	-	-	1	-	-	COR	-
ureg	-	-	-	-	-	-	-	-	-	-	-	COR

Abs. correlation	
COR	>0,8
COR	>0,95

Source: own calculation

2.6 Other indicators

Asset size as well as **an age** of a company are one of the most used variables with significant negative effect on bankruptcy likelihood. Larger and older companies have apparently survived competition which allowed them to grow larger and older, therefore internal

processes and ability to sustain fierce market condition are implicitly included in these variables.

Brand new variables, to my knowledge never used before, are *governance indicators*. **Contact score** was constructed from contact information according to magnitude of posted information. In case no contact information was available, company received 1 point. Telephone information[43] was the first category, physical address or e-mail the second and finally web address was the last category worth of point. Subsequently, company could score maximally 4 point for contact information.

The second part of governance indicator, named **representation score**, was comprised of unique entries of responsible positions in company. One point was appointed to companies, for which just one member of official representation was stated. If two positions were filled in, company could get two points, hence three points for three unique names corresponding to position. For four and more unique entries[44], company could receive four points. Position filled the most often was chief executive officer, which usually corresponded to personal name of statutory organ.

Individual scores for contact information and representation were included into dataset as well as overall governance score. Idea behind governance score is that the more open and transparent companies are in mentioned categories, the better governance can be expected, hence lower probability of default.

Table no. 8 - Macroeconomic and other indicators

Notation	Name of Indicator	Exp. Effect
Macroeconomic Indicators		
e	Employment rate	-
u	Unemployment rate	+
ureg	Unemployment in region	+
i	Inflation	+/-
rdep	Interest rate on deposits	+
rlon	Interest rate on loans	+
gdp	GDP	-
badlon	Bad Loans	+
Other Indicators		
As	Assets	-
age	AGE	-
Cscor	Contact Score	-
Rscor	Representation Score	-
Gscor	Governance Score	-

[43] Fixed line or mobile phone

[44] Maximum of unique name in firm's representation was eight.

2.7 Missing data

A problem of missing data was not basically described in any previous studies. However, this problem is inevitably included in financial ratio analysis, since significant numbers of companies do not necessarily use at least one of tools, reflected in financial statements i.e. short-term bank loans.

Many methodological approaches to treat missing values, hence financial ratios, are available. "Missing data mechanism" defined by Rubin [Rubin 1976] and colleagues [Little, Rubin, 2002] is still in use today. Missing data can be classified as:

1. missing completely at random (MCAR)
2. missing at random (MAR)
3. missing not at random (MNAR)

In our case data are MNAR, since missing values are directly and systematically related to use of certain financial/operational tools included in accounting reports and contain information about financial structure.[45]

One possibility how to solve missing data problem is to delete respectively, discard missing values entries for all companies, hence restrict analysis for complete data. Major disadvantages of this approach are reduction of sample size[46], but more importantly, since data are MNAR, the analysis will result in biased estimates as only "good" entries are used. However, this technique was used partially also in this thesis, when around 70 companies with more than ten missing financial ratios and non-bankrupted (!) were deleted from the data sample. These companies generally did not fill in substantial part of financial statements. In such circumstances, even imputing missing values would not have solved problem. Moreover, since companies were only non-bankrupted, solely less than two per cent of companies were omitted. One financial indicator variable was excluded from further analysis because of large number of missing data as well. Interest coverage ratio I was not defined in 43 % of cases. Imputing missing values would not be very helpful in this situation, mainly when Interest coverage ratio II is available. Companies probably report interests in other financial expenses altogether with fees for bank services, therefore values for interest expenses in denominator were not defined.

[45] i.e. equity/debt position or short/long term financing position
[46] By circa 300 companies entries in our case

Single imputation methods, meaning techniques when missing values are replaced by supposedly suitable figures (mean imputation or regression imputation) are not satisfying as well. According to Baraldi and Enders: “Both, mean imputation and regression imputation lead to bias because they fail to account for the variability that is present in the hypothetical data values.” [Baraldi and Enders 2010: 13]

2.7.1 Modern missing data techniques

The most widely and strongly recommended latest “state-of-the-art” techniques for treatment of missing data are multiple imputation and maximum likelihood estimation. Both of these approaches are constructed and better suitable for MCAR and MAR data, on the other hand, according to Baraldi and Enders: ”They too will yield biased parameter estimates when the data are MNAR, although the magnitude of this bias tends to be far less than the bias that results from traditional techniques.” [Baraldi and Enders 2010: 15] One might add that bias could be even larger when excluding observations with missing values.

Amelia II, software package for R, created especially to multiply impute missing data for cross-section or time series data, was employed to complete data set. As the main authors of program states: “Amelia II draws imputations of the missing values using a novel bootstrapping approach, the EMB (expectation-maximization with bootstrapping) algorithm. The algorithm uses the familiar EM (expectation-maximization) algorithm on multiple bootstrapped samples of the original incomplete data to draw values of the complete-data parameters. The algorithm then draws imputed values from each set of bootstrapped parameters, replacing the missing values with these draws.” [Honaker, King, Blackwell 2011: 2]

Main entries from accounting statements, financial statements and other company specific variables were used in multiple imputation. Five imputed datasets were created, using default indicating the cross section variable, in order to catch possible different effects in default groups. Additionally empri function[47] was used to minimize the covariances of the data, although keeping the means and variances the same. Empri function effectively includes artificial variables to the data set with the same means and variances as the existing data but with zero covariances. Five data sets were produced by Amelia II software, from which one final data set of averages of imputed values was calculated with overall number of 862 imputed missing values.

[47] Value of the prior was set to 1 % of number of rows in data set according to recommendations of authors available at http://cran.r-project.org/web/packages/Amelia/vignettes/amelia.pdf

3. Model

3.1 Model estimation

Data sample was divided to estimation and control sample using random selection. Estimation sample consisted of 2 809 companies out of 3 732 (75.3 %) with basically the same distribution of bankrupted and non-bankrupted companies as in original sample. Companies bankrupted up to 400 days after last accounting statement counted for 6.06 % in overall sample and 6.48 % in estimation sample.

Objective of the model building is to minimize number of variables employed as to achieve higher numerical stability. Generally, the higher the number of variables, the higher estimated standard errors.

3.1.1 Estimation techniques

There are several ways toward model estimation, and none of them is definitively denoted in literature to be the best one. Three methods of model estimation were employed. Altogether more than 200 models were estimated and compared according to noted indicators of quality.[48]

An approach proposed by Hosmer and Lemeshow [Hosmer, Lemeshoe 2000] embodies several steps of building-up a model. The first step of selection process of variables rests in univariable analysis of each indicator. Likelihood ratio of chi-square test corresponds, in univariable analysis, to likelihood ratio of the coefficients. Various basic indicators of quality of included variable were noted[49]. Subsequently variables were ranged according to their quality by likelihood ratio. Mickey and Greenland [Mickey and Greenland 1989] recommend using up to 0.25 level of p-value for inclusion of variable into the model, since low, commonly used, levels often overlook important indicators.

After univariable analysis, which provided some basic outline of indicators, stepwise forward, as well as backward selection was executed. In forward selection, indicators were added to model including just intercept, according to their performance in univariable analysis. However, this procedure had two steps. Model consisting only from financial indicators was developed at first. In the second round, macroeconomic indicators were added in stepwise manner, according to their performance in univariable analysis.

[48] z- Wald test; P>|z|; McFadden R^2; log likelihood; D (-2LL); Likelihood ratio test and its pt p value; G statistics, Schwarz criterion and AIC

[49] In univariable analysis also std. error was recorded

By utilization of two step forward selection, significance of adding macroeconomic indicators to purely financial indicators model, could be consecutively verified. When whole list of variables was examined, one more round of sorted variables tryout followed.

Basic criteria for inclusion/exclusion of variable into/from the model were following:

1. Wald statistics and p-value for significance
2. Likelihood ratio tests
3. Pseudo R^2

For supposedly very good model, additional test statistics and criteria were estimated, including: Schwarz criterion, AIC, cases correctly predicted, Pearson's R, area under ROC curve, sensitivity as well as specificity.

Backward selection process was not appropriate in executed analysis, since due to almost exact collinearity of some variables, computer program[50] automatically excluded problematic variables, making backward stepwise selection practically inexecutable. Another problem with backward selection was computation time. Program is defaultly set for up to 16 000 iterations to definitively achieve convergence, considerably prolonging the process.

3.2 Initial best fit model

Decision criteria for the best five models were quite ambivalent. Therefore, there can be no definitive and absolute statement about "the best" model. When fitting various notably solid models, which do not differ significantly from each other, researcher´s preferences according to objective of study are decisive. After estimating fitting models, for which theoretical criteria do not offer conclusive result, practical criteria can be used. Lower number of variables is preferred in order to generalize model. Area under ROC curve, accompanied by Person's R and sensitivity and specificity are exactly those indicators, which were cardinal for the model selection. The model proved to work also according to theoretical fit indicators, since difference between higher pseudo R2 of estimated models and chosen one was just 0,006. Differences among other indicators were genuinely small as well.

[50] STATA/SE 12.00

Table no. 9 - Model decision criteria

Variables included in the model	Number of variables	ROC	Pearson's r	Correct Predictions (out of 2809)	Correct Predictions (%)	Sensitivity	Specificity
LevI RetER CCC QR ICRII SLT rlon i	**8**	**0,8988**	**0,17382565**	**2632**	**93,6988%**	**4,95%**	**99,85%**
LevI RetER CCC QR ICRII SLT rlon i i-1	9	0,8961	0,16015358	2631	93,6632%	4,40%	99,85%
LevI RetER CCC QR ICRII SLT rlon CapR i rdep	10	0,8959	0,16620194	2631	93,6632%	4,95%	99,81%
LevI RetER CCC QR ICRII SLT rlon i gdp1	9	0,8954	0,16620194	2631	93,6632%	4,95%	99,81%
LevI RetER CCC QR ICRII SLT rlon CapR i Rscor	10	0,8919	0,16874539	2632	93,6988%	4,40%	99,89%

Source: own calculation

From table no.9 it can be deducted that preferred model fulfills chosen decision criteria, since it has lowest number of variables included into the model, highest area under ROC and Pearson's R coefficient. In case of number of correct predictions, sensitivity as well as specificity, there are models, which perform competitively. However, chosen model is ~~better~~ superior also in these criteria, since sensitivity can be viewed as more important, when it means percentage of correct predictions of insolvent companies.

3.3 Model diagnosis

3.3.1 Multicollinearity

According to Variance Inflation Factor test, no multicollinearity between the variables was detected. Resulting from correlation matrix of variables, also no significant, serious threat to the model was encountered. Tendency towards large standard errors, which would deteriorate further statistics, is not to be suspected.

Minimum possible value = 1.0
Values > 10.0 may indicate a collinearity problem

LevI	1.001	ICRII	1.041
RetER	1.163	SLT	1.111
CCC	1.010	rlon	1.922
QR	1.000	i	1.920

VIF(j) = 1/(1 - R(j)^2),
where: R(j) is the multiple correlation coefficient between variable j and the other independent variables

3.3.2 Influential observation

Outliers can pose two serious problems for a model. The first problem can be an error in data entry. Outliers can also significantly skew estimated regressors. On the other hand, outliers can be a significant source of information as well. Mainly, in case of prediction of bankruptcy, special cases, meaning companies with abnormal financial ratios, can be especially important. Building model only for "nice" companies could deteriorate possible utilization and detection of special cases. A possible solution for outliers could be their replacement by defined upper and lower boundary values. In my point of view, that would weaken model purpose, as companies in risk of insolvency tend to record deviant financial ratios.

Two approaches to detection of influential observations are possible: distance of observations from the rest of data population and leverage effect on the regression line. Using weighted hat matrix as diagnostics, the diagonal leverage factors were computed. Three observations with potentially the largest effect in the fit are observations number 5 and 6, and 10[51].

After calculating Pearson residuals, observations 5 and 6 were one more time identified as the most influential in accordance with previous case by unstandardized residuals. Subsequently, approximation of Pregibon's influence statistic[52] confirmed observation 5. From graphical analysis of residuals also observation 3 seemed to be problematic. From graphical analysis it can be concluded, that even outliers do not deviate strictly individually, what could signalize some systematic information. After an exclusion of all mentioned observations, no significant difference in model performance was observed. Neither in coefficients, nor in overall indicators. Another possibility how to detect problematic values is an analysis of distribution of employed financial indicators. After plotting their values compared to normal distribution, several observations were visually identified as problematic.[53]

[51] See hat matrix diagnostics in Appendix
[52] Also refered to as Cook's distance
[53] Observations: 2, 3, 4, 5, 6, 10, 12, 43, 278, 2786, 2789, 2793, 2794, 2795

Table no. 10 - Comparison of original model with excluded influential variable model

Model	Number of Observations	ROC	Pearson's r	Correct Predictions	Correct Predictions (%)	Sensitivity	Specificity	McFadden R^2	LL	D (-2LL)	LL ratio test	Schwarz crit.	AIC
Original model	2809	0,8988	0,17383	2632	93,6988%	4,95%	99,85%	0,262	-497,452	994,903	353,16	1066,369	1012,903
Exluded (residual tests)	2805	0.8988	0,17381	2628	93,6898%	4,95%	99,85%	0,262	-497,218	994,435	353,09	1065,888	1012,435
Exluded (normality)	2795	0.8975	0,18398	2622	93,8104%	5,03%	99,89%	0,259	-492,712	985,423	344,72	1056.843	1003.423

Source: own calculation

However, after exclusion even more influential observations, based not on the model statistics, but on the analysis of outliers, model proved to work poorly, when three variables[54] became insignificant. Generally, dysfunction of the model in different data setting is normal, when adjustment according to time and utilized data is needed.

3.4 Data Truncation

Estimation sample was adapted after detection of instability of the model to influential variable, when financial indicators values were truncated from bottom, to the 5th and from top, to the 95th percentile of the values. All major outliers were therefore excluded and in the same time 90 % of population offered comfortable variety of the data. The final model for truncated data sample was outstanding, compared to previous estimates on original data sample.

Table no. 11 - Model estimation on truncated data

ROC	Pearson's r	Correct Predictions	Correct Predictions (%)	Sensitivity	Specificity
0,9475	0,6536947	2707	96,3688%	51,10%	99,51%
McFadden R^2	LL	D (-2LL)	LL ratio test	Schwarz crit,	AIC
0,5249	-320,226	640,452	707,61	703,9763	656,4516

Source: own calculation

The new model on new data performs incredibly well. After general smoothening of extreme values 51.10 sensitivity was achieved, what means that model exactly(!) predicted more than half of defaulted companies compared to 5% in previous cases. Also all other indicators are considerably better than in previous model. Multicollinearity was rejected with VIF test as well as correlation matrix analysis.

Output no. 1 - Model estimation on truncated data

```
Logistic regression                                     Number of obs   =       2809
                                                        LR chi2(7)      =     707.61
                                                        Prob > chi2     =     0.0000
Log likelihood = -320.22581                             Pseudo R2       =     0.5249
```

Def	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
LevI	-.0164266	.005133	-3.20	0.001	-.0264872	-.006366
LevII	4.859957	.3280452	14.81	0.000	4.217	5.502913
DIO	.0037881	.001326	2.86	0.004	.0011892	.0063869
RetER	-.0121221	.0039448	-3.07	0.002	-.0198538	-.0043904
Rscor	.4575584	.1307735	3.50	0.000	.2012471	.7138698
rlon	-11.36963	1.052494	-10.80	0.000	-13.43247	-9.306776
i	-4.343836	.5733381	-7.58	0.000	-5.467558	-3.220114
_cons	42.63912	4.712723	9.05	0.000	33.40236	51.87589

Source: own calculation

Output no. 2 - Initial best fit model performance:

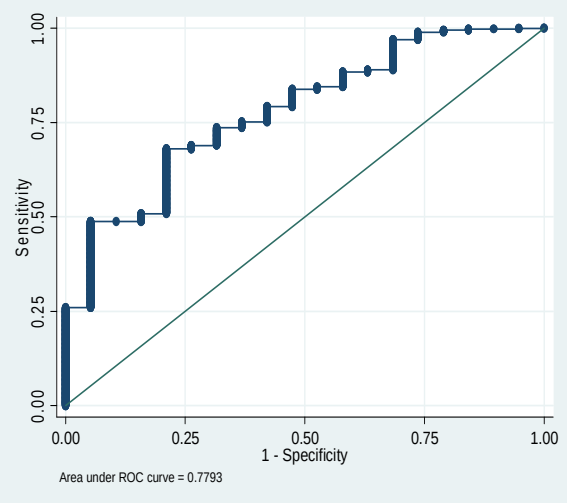
```
Logistic regression                                     Number of obs   =       2809
                                                        LR chi2(7)      =      29.87
                                                        Prob > chi2     =     0.0001
Log likelihood = -98.928681                             Pseudo R2       =     0.1312
```

LevI	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
RetER	.0050469	.0026883	1.88	0.060	-.0002221	.0103159
CCC	2.38e-06	4.27e-06	0.56	0.577	-5.99e-06	.0000108
QR	.0367453	.0075842	4.85	0.000	.0218806	.05161
ICRII	.0000204	.0020978	0.01	0.992	-.0040911	.004132
SLT	.6905645	.2021452	3.42	0.001	.2943671	1.086762
rlon	-.3049405	.5327592	-0.57	0.567	-1.349129	.7392483
i	-.0029971	.1799983	-0.02	0.987	-.3557873	.3497932
_cons	5.354493	1.942182	2.76	0.006	1.547887	9.1611

Source: own calculation

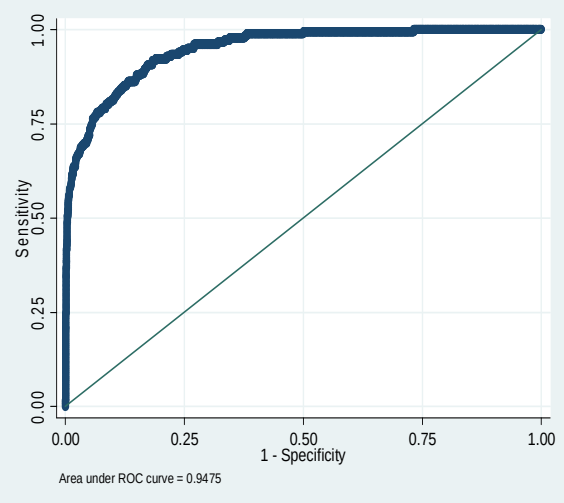
[54] CCC, SLT and QR

Graph no.1 - ROC Curve of initial best fit model



Source: own calculation

Graph no.2 - ROC Curve of new fit model



Source: own calculation

3.5 Complete sample model estimation

For the estimation purposes of fitted model, estimation sample was created. Fit of the models was test on the whole sample. The first best fit model seemed to perform much worse than expected, however after assessment of other models, supposed to be among the best at first, results were quite competitive. As in table no.12 can be seen, all models followed almost the same path, when Sales Turnover Ratio proved to be insignificant for the model. The third model is 6th best from original estimation sample and was not included in comparison of the best fitted models assessed on estimation sample. It was included here, since it does not

involve Sales Turnover Ratio. Quite surprising was also the insignificance of the Leverage I ratio, since it was dismissed in all assessed models. Even though theoretical indicators of fit deteriorated, area under ROC seems to record only small changes and still demonstrate strong applicability. The new best fit models estimated from truncated data sample performed well

Table no. 12 - Performance of models in various data samples[55]

Model Nu.	Variables included in the model	ROC	Pearson's r	Correct Predictions	Correct Predictions (%)	Sensitivity	Specificity	McFadden R^2	D (-2LL)	LL ratio test	Schwarz crit.	AIC
1	**Levl RetER CCC QR ICRII SLT rlon i**	0,8988	0,173826	2632	0,937	4,95%	99,85%	0,262	994,903	353,16	1066,369	1012,903
	Levl* RetER CCC QR ICRII SLT* rlon i	0,885	0,104641	3506	0,939	2,65%	99,83%	0,218	1333,67	371,84	1407,689	1351,667
2	**Levl RetER CCC QR ICRII SLT rlon i i-1**	0,8961	0,160154	2631	0,937	4,40%	99,85%	0,2641	992,05	356,01	1071,456	1012,05
	Levl* RetER CCC QR ICRII SLT* rlon i i-1	0,8867	0,090059	3506	0,939	2,22%	99,83%	0,2214	1327,98	377,53	1410,23	1347,983
3	**Levl RetER CCC QR ICRII CapR e-1 rlon i**	0.8907	0,145589	2630	0,936	3,85%	99,85%	0,2445	1018,43	329,63	1091.75	1038.284
	Levl* RetER CCC QR ICRII CapR e-1 rlon i	0,8783	0,104641	3506	0,939	2,65%	99,83%	0,2235	1324,28	381,24	1406,523	1344,276
4	**Levl LevII DIO RetER Rscor rlon i**	0,9475	0,653695	2707	0,937	51,10%	99,51%	0,5249	994,903	707,61	703,9763	656,4516
	Levl* LevII* DIO* RetER Rscor* rlon i	0,8529	0,101973	3508	0,940	1,77%	99,94%	0,206	1354,1	351,41	1419.902	1370.104

Source: own calculation

3.6 Complete truncated sample model estimation

As in the case of model for unadjusted data sample, the final model for truncated data sample was tested on overall data, but modified to match the same estimating condition. Model performed reasonably well, when just Retained earnings ratio proved to be insignificant and was subsequently excluded from estimation. Following models resulted from estimation:

Model	ROC	Pearson's r	Correct Predictions	Correct Predictions (%)	Sensitivity	Specificity	McFadden R^2	LL	D (-2LL)	LL ratio test	Schwarz crit.	AIC
Levl LevII RetER* Rscor rlon i	0,9403	0,6305	3732	96,409%	48,23%	99,52%	0,5018	-424,875	849,751	855,76	907,324	863,751
Levl LevII Rscor rlon i	0,9406	0,6305	3732	96,409%	48,23%	99,52%	0,5015	-425,091	850,182	855,33	899,53	862,182

Source: own calculation

Even original model performed very good, when omitting retained earnings brought enhancement in theoretical indicators as Schwarz and Akaike's ciretria.

[55] First model of a kind (**bold)** is model fitted on its own estimation sample. Model nu. 1,2,3 were estimated on original estimatedata sample, while model 4 on truncated.

Output no.3 - Complete truncated sample model estimation

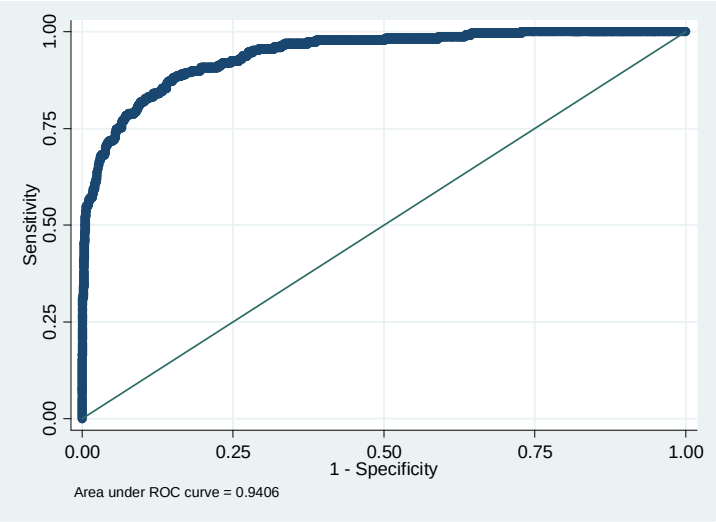
```
Logistic regression                               Number of obs   =       3732
                                                  LR chi2(5)      =     855.33
                                                  Prob > chi2     =     0.0000
Log likelihood = -425.09091                       Pseudo R2       =     0.5015
```

Def	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
LevI	-.0154407	.0049767	-3.10	0.002	-.0251948	-.0056867
LevII	5.055113	.3005643	16.82	0.000	4.466018	5.644208
Rscor	.4911028	.1170006	4.20	0.000	.2617858	.7204198
rlon	-10.46552	.8610703	-12.15	0.000	-12.15319	-8.777857
i	-3.595094	.4137461	-8.69	0.000	-4.406022	-2.784167
_cons	37.94727	3.761684	10.09	0.000	30.5745	45.32003

```
Note: 662 failures and 1 success completely determined.
```

Source: own calculation

Graph no.3 - Complete truncated sample model estimation



Source: own calculation

4. Interpretation of the results

The new fit data model from truncated data estimation was selected as the final model worth interpreting because of two main reasons. From estimation point of view, it does much better job in all measured indicators. The second reason is somehow anomalously estimated coefficients of initial estimated model, since they are all negative. The most likely reason behind negative coefficient is the presence of influential observations and outliers, which significantly bias estimate.

Output no.4 – Final model

```
Logistic regression                             Number of obs   =       2809
                                                LR chi2(7)      =     707.61
                                                Prob > chi2     =     0.0000
Log likelihood = -320.22581                     Pseudo R2       =     0.5249
```

Def	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
LevI	.9837076	.0050494	-3.20	0.001	.9738605	.9936542
LevII	129.0186	42.32393	14.81	0.000	67.82968	245.4058
DIO	1.003795	.001331	2.86	0.004	1.00119	1.006407
RetER	.9879511	.0038973	-3.07	0.002	.980342	.9956192
Rscor	1.580211	.2066497	3.50	0.000	1.222927	2.041878
rlon	.0000115	.0000121	-10.80	0.000	1.47e-06	.0000908
i	.0129866	.0074457	-7.58	0.000	.0042215	.0399505
_cons	3.30e+18	1.55e+19	9.05	0.000	3.21e+14	3.38e+22

```
Note: 503 failures and 1 success completely determined.
```

Source: own calculation

While the effect in case of retained earnings is consistent with a priory assumption, mainly interest rates on loans effect, in terms of magnitude and sign, is surprising. Recall that interest rates on loans is determined on the base of new contracts, therefore effect for companies with ongoing financial loan does not necessarily pose immediate threat. Answer to this ratio can be in financial system and with macroeconomic view. When interest rates for loans go up, they reflect reference rate of central bank, which increases reference interest rates in good times. When interest rates drops, central bank follows expansionary monetary policy usually to stimulate declining economy. Rate on loans is countercyclical, therefore when it grows; the lower amount of insolvencies in economy can be expected. Microeconomic view, seeing rate on loans directly in the accounts of company, can be in this case misleading.

Inflation can have, from company´s specific perspective, ambivalent effect, since it can have effect on costs as well as on revenues. Seeing that inflation has relatively large negative

effect, response of growth of costs is from microeconomic point of view, more plausible. In macroeconomic context, inflation follows the same logic, but in different time horizon as mentioned inflation rate on loans. When taking time horizon[56] of collected data into account, it makes sense that declining inflation reflected deteriorating economy, therefore raising number of insolvency procedures.

A conflict can be seen in Leverage ratios, since the first one has negative effect and second positive. Generally speaking, these two ratios accompany each other, when catching financial structure of the company. Leverage to equity ratio can have negative effect, since ineptness to certain level can bring healthy leverage effect for company. Resolution can also lie in prudence of banks, where higher leverage to equity ratio means higher confidence in company's business. Therefore, when indebtness grows opposite to equity, but it keeps track with growth of assets, odds of healthy company to fill for insolvency are lower than one.

On the other hand, when debt compared to assets grows faster, it means that firm´s business deteriorates and wealth covering for debt decreases, what automatically agitate creditors, therefore odds of bankruptcy is significantly high.

Inventory turnover correspond to supposed effect. When company needs one more day for inventory to sell, or process, it loses money stocked in passive assets and odds of filling for insolvency grows.

Significance of Representation score is notable, since there are many attempts in contemporary bankruptcy probability research to explore area of corporate governance. Effect of this score is however surprising, since negative effect on insolvency likelihood was expected. In this case, explanation can be found in type of companies in research because mostly small companies[57] were collected in dataset. The bigger number responsible persons in management, the more likely conflict inside a company can occur, posing operational problems. If number of responsible representatives grows by one, odds of insolvency in 400 days horizon grows 1.5 times.

[56] Financial statement from 2008 to 2012 and start of insolvency procedure up to 2013.

Output no.5 – Classification table

Logistic model for Def

Classified	True D	True ~D	Total
+	93	13	106
-	89	2614	2703
Total	182	2627	2809

Classified + if predicted Pr(D) >= .5
True D defined as Def != 0

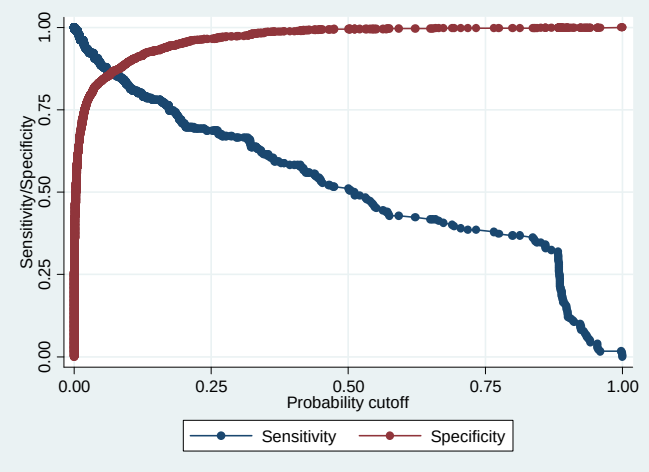
Sensitivity	Pr(+| D)	51.10%
Specificity	Pr(-|~D)	99.51%
Positive predictive value	Pr(D| +)	87.74%
Negative predictive value	Pr(~D| -)	96.71%
False + rate for true ~D	Pr(+|~D)	0.49%
False - rate for true D	Pr(-| D)	48.90%
False + rate for classified +	Pr(~D| +)	12.26%
False - rate for classified -	Pr(D| -)	3.29%
Correctly classified		96.37%

Source: own calculation

Classification table shows very high sensitivity, what is extremely desirable. Recognizing company, which will default, is major role of the model, since protective measures can be set by existing creditor, and in case of potential loan client, bank can successfully reject loan proposal.

[57] Median Asset size is 8,78 millions CZK

Graph no.5 – Sensitivity, specificity plot



Source: own calculation

Quality of the model can be even better observed from Graph no.5, when specificity keeps good pattern staying close to zero showing good estimation of non-bankrupt companies´ classification, while sensitivity showed balanced path across cutoff points.

One should not forget, that logit model differs from univariate analysis, and more important is the performance of the model as such, than individual effects.

Summary

From theoretical point of view, this thesis presents concise but comprehensive overview of the most important papers dedicated to prediction of corporate bankruptcy. Not only final models and overall contribution of used papers is portrayed, but also logic behind models´ construction and data selection was noted. In this case, this thesis differs from ordinary papers, where literature overview is a compressed list of brief statements about previous papers. Quality of analysis was preferred over the quantity, since it could be easier to just recall main content of many papers summarized in conclusion, but it allowed me for deeper understanding not even of methodology, but also development of this discipline through history and context as well.

Overview of the theory behind the models employed is included in methodological part altogether with list of main qualitative indicators and ratios, which are crucial for quality assessment and comparison of the models. Collection of data is described relatively in detail. Quality of works in this area usually rests from bigger part on the quality and volume of available data. Since resources for purchasing big and supreme data file were not available, data had to be collected manually through lengthy procedure. Moreover collected data includes information, which are usually not available in automatic download from big databases, especially specifications of management and responsible persons. From this point of view, data collected are of high quality and in Czech Republic relatively unique[58]. Noticeable is also multiple imputation method used, current "state-of-the-art" technique for missing data treatment, is exceptional, since almost no authors in problematic report missing data problem. Methodological part offers therefore general guideline for future potential followers in this area of research as well.

Practical part concentrates on models estimation for various data setting. It is obvious, that quality of the model is considerably dependent on data quality, what was show by contrasting models on raw and truncated datasets. All important qualitative indicators improved after smoothing outliers and influential observations in the data. Conclusion can be made, that by smoothing data, significantly better model can be estimated with superior discriminating power on the same data points. From the models variables point of view, noticeable is a successful inclusion of macroeconomic variables into the model, which was set as objective at the beginning of this thesis. Even more significant, taking into account latest research, is

inclusion of governance indicators, from which number of persons in management presented in governmental score has significant effect in final model.

Záver

Z teoretického uhľa pohľadu, táto práce prináša výstižný a súhrny prehľad najdôležitejších vedeckých článkov a statí venovaných téme predikcie firemných bankrotov. Nielen finálne výsledky modelov a celkový prínos zahrnutých prác je obsiahnutý v prehľade literatúry, ale aj logika v pozadí budovania modelov a zberu dát bola popísaná. V tomto prípade sa táto práca odlišuje od zvyčajných vedeckých článkov, kde sa prehľad literatúry obmedzuje na zhustený výpočet najdôležitejších konštatovaní predchádzajúcich prací. Kvalita analýzy bola v tomto prípade nadriadená kvantite, keďže by mohlo byť oveľa ľahšie spomenúť hlavné výsledky a prínos práce obsiahnutý zvyčajne v závere samotných prací. Takto bolo možné hlbšie pochopiť nie len použitú metodológiu, ale aj vývoj tejto disciplíny v čase a kontexte.

Prehľad teórie použitých modelov je obsiahnutá v metodologickej časti spolu s hlavnými kvalitatívnymi indikátormi a ukazovateľmi, ktoré sú nevyhnuté na správne posúdenie a porovnanie modelov. Zber dát je pomerne detailne popísaný, keďže kvalita práci v tejto oblasti zvyčajne z väčšej časti spočíva na kvalite and množstve dostupných dát. Keďže zdroje na kúpu kvalitných a obsiahlych dát neboli k dispozícií, dáta museli byť zbierané ručné pomerne zdĺhavou a náročnou procedúrou. Napriek tomu obsahujú práve údaje, ktoré nie sú zvyčajne dostupné pri veľkokapacitnom sťahovaní dát z databáz, najmä údaje o zodpovedných osobách a pozíciách vo firmách. Z tohto uhľa pohľadu sa jedná o veľmi kvalitný datasúbor, a na pomery Českej Republiky pomerne jedinečný. Technika imputácie chýbajúcich hodnôt je taktiež hodná pozornosti, keďže sa jedná o najnovšie a momentálne najviac odporúčanú metódu na riešenie problému nevyplnených hodnôt. Najviac takmer žiadni autori publikujúci v tejto oblasti nespomínajú riešenie chýbajúcich hodnôt. Z tohto dôvodu môže byť táto práca taktiež prínosná pre budúcich bádateľov v tejto oblasti.

Praktická časť tejto práce sa zameriava na odhad modelu v prípadoch s rozdielnymi dátami. Je zjavné, že kvalita modelu je značne závislá na kvalite dát, čo potvrdzuje porovnanie modelu čistých a upravených dát. Všetky dôležité ukazovatele kvality sa zlepšili po vyhladení odľahlých a ovplyvňujúcich pozorovaní v data súbore. Z analýzy plynie záver, že vyhladenie dát umožňuje odhad značne lepšieho modelu s lepšími diskriminujúcimi vlastnosťami

[58] i.e. [Jakubik Teply 2008] have in their dataset 151 bankrupted companies and 606 of good ones, compared to 228 insolvent resp. 831 healthy employed in this work

rovnakých pozorovaní. Z pohľadu premenných modelu, zahrnutie makroekonomických ukazovateľov do modelu, ktoré bolo vytýčené ako jeden z cieľov tejto práce, prinieslo úspech. Ešte viac dôležité, z pohľadu súčasného výskumu, je zahrnutie premenných popisujúcich kvalitu správy podniku, z ktorých množstvo osôb v manažmente podniku, zahrnuté v premennej riadiace skóre, malo značný efekt vo výslednom modely.

References

ALLISON, P. D., *Missing data*. Thousand Oaks, Sage Publications, 2002

ALTMAN, E., *Financial ratios, discriminant analysis and the prediction of corporate bankruptcy*. in Journal of Finance 23, Vol. 4, 1968, pages 589–609

ALTMAN, E., HALDEMAN, R., NARAYANAN, P., *ZETA analysis: A new model to identify bankruptcy risk of corporations* in Journal of Banking and Finance 1, Vol. 1, 1977, pages 29–54.

ALTMAN, E., EINSENBEIS, R.A., *Financial Applications of Discriminant Analysis: A Clarification* in Journal of Financial and Quantitative Analysis, 1977b, pages 185-195.

ALTMAN, E., *Predicting financial distress of companies: Revisiting the z-score and zeta models*, Update of Journal of Finance 1968 and Journal of Banking & Finance 1997, 2000

BALTAGI, B.H., *Econometrics*. Berlin, Springer, 2008

BEAVER W.H., *Financial ratios as predictors of failures. Empirical research in accounting: Selected studies* in Journal of Accounting Research 3, 1966, pages 71–111.

BARALDI, A. N., ENDERS C. K., *An introduction to modern missing data analyses* in Journal of School Psychology 48, 2010, pages 5–37

CIPRA, T. *Finanční ekonometrie*. Praha, Ekopress, 2008

DEAKIN, E.B., *A Discriminant Analysisi of Predictors of Business Failure* in Journal of Accounting Research, Vol. 10, No. 1, 1972, pages 167-179

ENDERS, C. K., *Applied missing data analysis*. New York, Guilford Press, 2010

HONAKER, J., KING, G., BLACKWELL, M., *Amelia II: A Program for Missing Data* in Journal of Statistical Software, Vol. 45, No. 7, 2011, pages 1-47

HOSMER, D. W., LEMESHOW, S., *Applied logistic regression*. New York, Wiley & Sons, 2000

JAKUBIK, P., TEPLY, P., *The Prediction of Corporate Bankruptcy and Czech Economy's Financial Stability through Logit Analysis* ,IES Working Paper 19/2008. IES FSV. Charles University.

JAKUBIK, P. *Exekuce, bankroty a jejich makroekonomické determinant*, IES Working Paper 29/2007. IES FSV. Charles University.

KALLBERG, J.G., UDELL, G.F., *The value of private sector business credit information sharing: The US case* in Journal of Banking and Finance 27, 2003, pages 449–469

LITTLE, R.J.A., RUBIN, D.B., *Statistical analysis with missing data*, Hoboken, Wiley, 2002

LIU, J., *Macroeconomic determinants of corporate failures: evidence from the UK* in Applied Economics, 36, 2004, pages 939-945

MARTIN, D., *Early warning of bank failure: A logit regression approach* in Journal of Banking and Finance, Vol. 1, No. 6, November, 1977, pages 249-276.

MCFADDEN, D. *Conditional Logit Analysis of Qualitative Choice Behavior* in Frontiers in Econometrics, edited by ZAREMBKA, P., New York, Academic Press, 1973

MORAVEC, T., *The Bankruptcy in the Czech Republic – Influence of Macroeconomic variables* in Acta academica karviniensia 3, 2013, pages 136 - 145

MICKEY, J., GREENLAND, S., *A study of the impact of confounder-selection criteria on effect estimation* in American Journal of Epidemiology, 1989, pages 125 - 137

OHLSON, J.A., *Financial ratios and the probabilistic prediction of bankruptcy* in Journal of Accounting Research 18, No.1, 1980, pages 109–131.

PALEPU, K. G., *Business analysis and valuation : text and cases* (IFRS ed). Thomson Learning, London, 2007

RUBIN, D. B., *Inference and missing data* in Biometrika, 63, 1976, pages 581−592.

SANTOMERO, A., VINSO J.D., *Estimating the Probability of Failure for Commercial Banks and the Banking System* in Journal of Banking and Finance, Vol 2, No. 1, 1977, pages 185–205

SCHAFER, J. L., OLSEN, M. K., *Multiple imputation for multivariate missing-data problems: A data analyst's perspective* in Multivariate Behavioral Research, 33, 1998, pages 545−571

VALLINI, C., CIAMPRI, F., GORDINI, N., BENVENUTI, M., *Can credit scoring models effectively predict small enterprise default? Statistical evidence from Italian firms*. Proceedings of the 8th Global Conference on Business & Economics (Florence, Italy), 2008

VALLINI, C., CIAMPRI, F., GORDINI, N., *Using artificial neural networks analysis for small enterprise default prediction modeling: Statistical evidence from Italian firms*. Oxford Business and Economics Conference Program, 2009

VIROLAINEN, K., 2004. *Macro Stress Testing with a Macroeconomic Credit Risk Model for Finland*, Bank of Finland, Discussion Paper, no. 18.

WADHWANI, S.B., *Inflation, Bankruptcy, Default Premia and the Stock Market* in The Economic Journal, Vol. 96, No. 381, 1986, pages 120-138

ZMIJEVSKI, M.E., *Methodological issues related to the estimation of financial mistress predictionmodels* in Journal of Accounting Research (Studies on Current Econometric Issues in Accounting Research) 22, 1984, pages 59–82.

List of appendices

Appendices

Appendix no. 1: Initially best model (output)

```
logit Def LevI RetER CCC QR ICRII  SLT rlon i, iterate(1000)

Iteration 0:   log likelihood = -674.02936
Iteration 1:   log likelihood = -608.66889
Iteration 2:   log likelihood = -602.62752
Iteration 3:   log likelihood = -572.68999
Iteration 4:   log likelihood = -520.76263
Iteration 5:   log likelihood = -498.92729
Iteration 6:   log likelihood = -497.49605
Iteration 7:   log likelihood = -497.45194
Iteration 8:   log likelihood = -497.45168
Iteration 9:   log likelihood = -497.45168

Logistic regression                     Number of obs   =       2809
                                        LR chi2(8)      =     353.16
                                        Prob > chi2     =     0.0000
Log likelihood = -497.45168             Pseudo R2       =     0.2620

------------------------------------------------------------------------------
         Def |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        LevI |  -.0069224   .0034921    -1.98   0.047    -.0137668    -.000078
       RetER |  -.1028106   .0116104    -8.86   0.000    -.1255666   -.0800547
         CCC |  -4.78e-06   2.55e-06    -1.87   0.061    -9.78e-06    2.21e-07
          QR |  -.0183514   .0062431    -2.94   0.003    -.0305877   -.0061151
       ICRII |  -.0028817   .0009012    -3.20   0.001     -.004648   -.0011153
         SLT |   -.024347   .0036273    -6.71   0.000    -.0314564   -.0172376
        rlon |  -10.23731   .8851493   -11.57   0.000    -11.97217   -8.502449
           i |  -4.748769   .5761118    -8.24   0.000    -5.877928   -3.619611
       _cons |   44.33587    4.21499    10.52   0.000     36.07464     52.5971
------------------------------------------------------------------------------
Note: 505 failures and 3 successes completely determined.
 estat ic

-----------------------------------------------------------------------------
      Model |    Obs    ll(null)   ll(model)     df         AIC        BIC
-------------+---------------------------------------------------------------
          . |   2809   -674.0294   -497.4517      9     1012.903    1066.369
-----------------------------------------------------------------------------
               Note:  N=Obs used in calculating BIC; see [R] BIC note
. estat classification

Logistic model for Def

              -------- True --------
```

Classified	D	~D	Total
+	9	4	13
-	173	2623	2796
Total	182	2627	2809

Classified + if predicted Pr(D) >= .5
True D defined as Def != 0

Sensitivity	Pr(+| D)	4.95%
Specificity	Pr(-|~D)	99.85%
Positive predictive value	Pr(D| +)	69.23%
Negative predictive value	Pr(~D| -)	93.81%
False + rate for true ~D	Pr(+|~D)	0.15%
False - rate for true D	Pr(-| D)	95.05%
False + rate for classified +	Pr(~D| +)	30.77%
False - rate for classified -	Pr(D| -)	6.19%
Correctly classified		93.70%

. lroc

Logistic model for Def

number of observations = 2809
area under ROC curve = 0.8988

Gretl

Variance Inflation Factors

Minimum possible value = 1.0
Values > 10.0 may indicate a collinearity problem

LevI	1.001
RetER	1.163
CCC	1.010
QR	1.000
ICRII	1.041
SLT	1.111
rlon	1.922
i	1.920

VIF(j) = 1/(1 - R(j)^2), where R(j) is the multiple correlation coefficient

between variable j and the other independent variables

correl LevI RetER CCC QR ICRII SLT rlon i
(obs=2809)

	LevI	RetER	CCC	QR	ICRII	SLT	rlon	i
LevI	1.0000							
RetER	0.0019	1.0000						
CCC	-0.0067	0.0935	1.0000					
QR	-0.0000	-0.0011	0.0015	1.0000				
ICRII	0.0000	-0.1857	-0.0177	0.0021	1.0000			
SLT	-0.0016	-0.3071	0.0019	-0.0061	-0.0031	1.0000		
rlon	0.0248	0.0231	0.0064	-0.0041	0.0268	-0.0017	1.0000	
i	0.0081	-0.0187	-0.0049	-0.0040	0.0258	-0.0072	0.6910	1.0000

Appendix no. 2: Influencial observations analysis (output)

sum lev if Def==1

Variable	Obs	Mean	Std. Dev.	Min	Max
lev	182	.0081991	.0310128	0	.2512377

. sum lev if Def==0

Variable	Obs	Mean	Std. Dev.	Min	Max
lev	2627	.0028773	.0303149	0	.9854512

Hat matrix diagnostics

. gsort -lev

. list LevI RetER CCC QR ICRII SLT rlon i lev dr if Def==1 in 1/6

	LevI	RetER	CCC	QR	ICRII	SLT	rlon	i	lev	dr
5.	-384.08	-.86	-235.2	.05	-35.46	.31	3.5608333	1.9166667	.2512377	.600201
6.	-307.73	.15	70.21	1.4	-1.52	1.81	3.5608333	1.9166667	.2490554	.8508323

list LevI RetER CCC QR ICRII SLT rlon i lev dr if Def==1 in 1/10

```
    +-------------------------------------------------------------------------------------------+
    |   LevI   RetER       CCC    QR    ICRII    SLT       rlon         i       lev        dr |
    |-------------------------------------------------------------------------------------------|
  5. | -384.08    -.86     -235.2   .05   -35.46    .31   3.5608333   1.9166667   .2512377
.600201 |
  6. | -307.73    .15       70.21   1.4    -1.52   1.81   3.5608333   1.9166667   .2490554
.8508323 |
 10. |  -1.19   -5.21   -389521.8   .12   -45.86    .01   3.5608333   1.9166667   .1648581
.6891451 |
    +-------------------------------------------------------------------------------------------+
```

dropped observations 3,5,6,10

model estim

Pearson's residuals

predict ps, rs
(2 missing values generated)

. gen ds = dr/sqrt(1-lev)

. gen sc = ps^2
(2 missing values generated)

. gsort -sc

. list LevI RetER CCC QR ICRII SLT rlon i ps ds if Def==1 in 1/6

```
    +-------------------------------------------------------------------------------------------+
    | LevI   RetER      CCC    QR     ICRII    SLT     rlon         i        ps         ds |
    |-------------------------------------------------------------------------------------------|
  2. | -7.59    -.35     47.23   .36     -.85    .92   4.2725        1.05   8.707442   2.947654 |
  3. |   .76    -.72     89.34   .66   -18.43   1.13   4.2725        1.05   8.620425   2.940905 |
  4. | -1.46   -1.26   -290.04    .1    -7.37   3.11   4.2725        1.05   8.607964   2.939938 |
  5. | 127.9    -.03   -106.71   .47      2.2   2.06   3.9775   1.4666667   8.575056   2.939717 |
  6. | -1.52    -.51   -164.64   .19   -116.22   4.86   4.2725        1.05   7.820676   2.875181 |
    +-------------------------------------------------------------------------------------------+
```

Cook's Distance

list LevI RetER CCC QR ICRII SLT rlon i dr cook if Def==1 in 1/6

```
    +-------------------------------------------------------------------------------------------+
    |  LevI   RetER     CCC    QR   ICRII   SLT       rlon         i        dr       cook |
    |-------------------------------------------------------------------------------------------|
```

```
  5. | 319.09     .03   -95.7   .69   -5.28   .31   3.5608333   1.9166667   2.633426   1.252854 |
    +-------------------------------------------------------------------------------------------+
```

Appendix no. 3: Model after excluded influencial variables (output)

```
logit Def LevI RetER CCC QR ICRII  SLT rlon i

Iteration 0:   log likelihood = -673.76122
Iteration 1:   log likelihood = -608.48033
Iteration 2:   log likelihood = -602.43509
Iteration 3:   log likelihood = -572.52626
Iteration 4:   log likelihood = -520.43257
Iteration 5:   log likelihood = -498.71788
Iteration 6:   log likelihood = -497.26941
Iteration 7:   log likelihood =  -497.2181
Iteration 8:   log likelihood = -497.21757
Iteration 9:   log likelihood = -497.21757

Logistic regression                     Number of obs   =       2805
                                        LR chi2(8)      =     353.09
                                        Prob > chi2     =     0.0000
Log likelihood = -497.21757             Pseudo R2       =     0.2620

------------------------------------------------------------------------------
         Def |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        LevI |  -.0069226   .0034923    -1.98   0.047    -.0137674   -.0000778
       RetER |  -.1028664   .0116099    -8.86   0.000    -.1256213   -.0801115
         CCC |  -4.78e-06   2.55e-06    -1.87   0.061    -9.78e-06    2.27e-07
          QR |  -.0183586   .0062436    -2.94   0.003    -.0305958   -.0061214
       ICRII |  -.0028809   .0009012    -3.20   0.001    -.0046472   -.0011147
         SLT |  -.0243618   .0036277    -6.72   0.000    -.0314719   -.0172517
        rlon |  -10.24397   .8851034   -11.57   0.000    -11.97875   -8.509204
           i |   -4.75161   .5760348    -8.25   0.000    -5.880617   -3.622602
       _cons |   44.36696   4.214692    10.53   0.000     36.10632     52.6276
------------------------------------------------------------------------------
Note: 502 failures and 3 successes completely determined.

. estat ic

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df         AIC        BIC
-------------+---------------------------------------------------------------
           . |   2805   -673.7612   -497.2176      9    1012.435    1065.888
-----------------------------------------------------------------------------
                Note:  N=Obs used in calculating BIC; see [R] BIC note

. estat classification
```

Logistic model for Def

```
            -------- True --------
Classified |        D           ~D  |      Total
-----------+--------------------------+-----------
    +      |        9            4  |         13
    -      |      173         2619  |       2792
-----------+--------------------------+-----------
   Total   |      182         2623  |       2805

Classified + if predicted Pr(D) >= .5
True D defined as Def != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)    4.95%
Specificity                     Pr( -|~D)   99.85%
Positive predictive value       Pr( D| +)   69.23%
Negative predictive value       Pr(~D| -)   93.80%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    0.15%
False - rate for true D         Pr( -| D)   95.05%
False + rate for classified +   Pr(~D| +)   30.77%
False - rate for classified -   Pr( D| -)    6.20%
--------------------------------------------------
Correctly classified                        93.69%
--------------------------------------------------
```

. lroc

Logistic model for Def

number of observations = 2805
area under ROC curve = 0.8988

.

Appendix no. 4: Model´s estimation output after futher exclustion of variables (output)

(residual tests & visual normality identification), dropped 14 variables

logit Def LevI RetER CCC QR ICRII SLT rlon i

Iteration 0: log likelihood = -665.0698
Iteration 1: log likelihood = -601.05656
Iteration 2: log likelihood = -595.26861
Iteration 3: log likelihood = -565.87938
Iteration 4: log likelihood = -514.91662
Iteration 5: log likelihood = -493.97855

```
Iteration 6:   log likelihood =  -492.7361
Iteration 7:   log likelihood = -492.71163
Iteration 8:   log likelihood = -492.71159

Logistic regression                             Number of obs   =       2795
                                                LR chi2(8)      =     344.72
                                                Prob > chi2     =     0.0000
Log likelihood = -492.71159                     Pseudo R2       =     0.2592

------------------------------------------------------------------------------
         Def |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        LevI |  -.0068422    .003474    -1.97   0.049    -.0136512   -.0000332
       RetER |  -.1019913   .0116105    -8.78   0.000    -.1247474   -.0792351
         CCC |  -4.82e-06   2.54e-06    -1.90   0.058    -9.79e-06    1.55e-07
          QR |  -.0182471   .0062352    -2.93   0.003     -.030468   -.0060263
       ICRII |  -.0028593   .0009027    -3.17   0.002    -.0046286     -.00109
         SLT |   -.024118   .0036103    -6.68   0.000    -.0311942   -.0170419
        rlon |  -10.14675    .885842   -11.45   0.000    -11.88297   -8.410529
           i |  -4.709321    .575983    -8.18   0.000    -5.838227   -3.580415
       _cons |   43.91748   4.217864    10.41   0.000     35.65061    52.18434
------------------------------------------------------------------------------
Note: 501 failures and 3 successes completely determined.

. estat ic

-----------------------------------------------------------------------------
      Model |    Obs    ll(null)   ll(model)     df         AIC        BIC
-------------+---------------------------------------------------------------
          . |   2795   -665.0698   -492.7116      9    1003.423    1056.843
-----------------------------------------------------------------------------
               Note:  N=Obs used in calculating BIC; see [R] BIC note

. estat classification

Logistic model for Def

              -------- True --------
Classified |         D          ~D  |      Total
-----------+--------------------------+-----------
     +     |         9           3  |         12
     -     |       170        2613  |       2783
-----------+--------------------------+-----------
   Total   |       179        2616  |       2795

Classified + if predicted Pr(D) >= .5
True D defined as Def != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)    5.03%
Specificity                     Pr( -|~D)   99.89%
```

```
Positive predictive value       Pr( D| +)   75.00%
Negative predictive value       Pr(~D| -)   93.89%
-------------------------------------------------
False + rate for true ~D        Pr( +|~D)    0.11%
False - rate for true D         Pr( -| D)   94.97%
False + rate for classified +   Pr(~D| +)   25.00%
False - rate for classified -   Pr( D| -)    6.11%
-------------------------------------------------
Correctly classified                        93.81%
-------------------------------------------------
```

```
. lroc

Logistic model for Def

number of observations =     2795
area under ROC curve   =   0.8975

.
```

Appendix no. 5: Final model fitting on overall truncated data (output)

```
Logistic regression                     Number of obs   =       3732
                                        LR chi2(6)      =     855.76
                                        Prob > chi2     =     0.0000
Log likelihood = -424.87541             Pseudo R2       =     0.5018

------------------------------------------------------------------------------
         Def |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        LevI |  -.0154776   .0049825    -3.11   0.002    -.0252432    -.005712
       LevII |   5.083151   .3045253    16.69   0.000     4.486293     5.68001
       RetER |    .001265    .001865     0.68   0.498    -.0023903    .0049202
       Rscor |   .4911089   .1171122     4.19   0.000     .2615732    .7206447
        rlon |   -10.4739   .8611352   -12.16   0.000     -12.1617    -8.78611
           i |  -3.584165   .4125724    -8.69   0.000    -4.392792   -2.775538
       _cons |   37.93464   3.758246    10.09   0.000     30.56861    45.30066
------------------------------------------------------------------------------
Note: 662 failures and 1 success completely determined.

estat ic

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df         AIC        BIC
-------------+---------------------------------------------------------------
           . |   3732   -852.7557   -424.8754      7     863.7508    907.3237
-----------------------------------------------------------------------------
               Note:  N=Obs used in calculating BIC; see [R] BIC note
```

```
. estat classification

Logistic model for Def

           -------- True --------
Classified |        D          ~D  |     Total
-----------+-------------------------+-----------
    +      |      109           17  |       126
    -      |      117         3489  |      3606
-----------+-------------------------+-----------
  Total    |      226         3506  |      3732

Classified + if predicted Pr(D) >= .5
True D defined as Def != 0
-------------------------------------------------
Sensitivity                     Pr( +| D)   48.23%
Specificity                     Pr( -|~D)   99.52%
Positive predictive value       Pr( D| +)   86.51%
Negative predictive value       Pr(~D| -)   96.76%
-------------------------------------------------
False + rate for true ~D        Pr( +|~D)    0.48%
False - rate for true D         Pr( -| D)   51.77%
False + rate for classified +   Pr(~D| +)   13.49%
False - rate for classified -   Pr( D| -)    3.24%
-------------------------------------------------
Correctly classified                        96.41%
-------------------------------------------------

.

 lroc

Logistic model for Def

number of observations =     3732
area under ROC curve   =   0.9403

Logistic regression                             Number of obs   =       3732
                                                LR chi2(5)      =     855.33
                                                Prob > chi2     =     0.0000
Log likelihood = -425.09091                     Pseudo R2       =     0.5015

------------------------------------------------------------------------------
         Def |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        LevI |  -.0154407   .0049767    -3.10   0.002    -.0251948   -.0056867
       LevII |   5.055113   .3005643    16.82   0.000     4.466018    5.644208
       Rscor |   .4911028   .1170006     4.20   0.000     .2617858    .7204198
        rlon |  -10.46552   .8610703   -12.15   0.000    -12.15319   -8.777857
```

```
           i |  -3.595094   .4137461    -8.69   0.000    -4.406022   -2.784167
       _cons |   37.94727   3.761684    10.09   0.000      30.5745    45.32003
------------------------------------------------------------------------------
Note: 662 failures and 1 success completely determined.

. estat ic

-----------------------------------------------------------------------------
      Model |    Obs    ll(null)   ll(model)     df         AIC        BIC
------------+----------------------------------------------------------------
          . |   3732   -852.7557   -425.0909      6     862.1818      899.53
-----------------------------------------------------------------------------
               Note:  N=Obs used in calculating BIC; see [R] BIC note

. estat classification

Logistic model for Def

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |       109            17  |        126
     -     |       117          3489  |       3606
-----------+--------------------------+-----------
   Total   |       226          3506  |       3732

Classified + if predicted Pr(D) >= .5
True D defined as Def != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   48.23%
Specificity                     Pr( -|~D)   99.52%
Positive predictive value       Pr( D| +)   86.51%
Negative predictive value       Pr(~D| -)   96.76%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    0.48%
False - rate for true D         Pr( -| D)   51.77%
False + rate for classified +   Pr(~D| +)   13.49%
False - rate for classified -   Pr( D| -)    3.24%
--------------------------------------------------
Correctly classified                        96.41%
--------------------------------------------------

. lroc

Logistic model for Def

number of observations =     3732
area under ROC curve   =   0.9406

. correl Def LevI LevII Rscor rlon i
```

(obs=3732)

	Def	LevI	LevII	Rscor	rlon	i
Def	1.0000					
LevI	-0.0816	1.0000				
LevII	0.0633	-0.0011	1.0000			
Rscor	0.0145	0.0052	-0.0015	1.0000		
rlon	-0.1969	0.0151	-0.0347	0.0342	1.0000	
i	-0.0770	0.0068	-0.0449	0.0007	0.6856	1.0000