University of Rostock

Doctoral Thesis

Modeling Default Probability in Credit Portfolios: Evidence based on Debt Securities Issuers in Germany

Author:
Lavri Labi

Supervisor:
Prof. Dr. Rafael Weissbach

A thesis submitted in fulfilment of the requirements for the degree of Doctor rerum politicarum (Dr. rer. pol.)

in the

Chair of Statistics and Econometrics
Faculty of Economic and Social Sciences

June 2016

Universität Rostock Traditio et Innovatio

urn:nbn:de:gbv:28-diss2017-0122-9

Gutachter:

1. Gutachter:

Prof. Dr. Rafael Weißbach
Lehrstuhl für Statistik und Ökonometrie

2. Gutachter:

Prof. Dr. Doris Neuberger
Lehrstuhl für Geld und Kredit

Datum der Einreichung: 24. Februar 2016

Datum der Verteidigung: 17. Juni 2016

Eidesstattliche Versicherung

Ich erkläre hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer Prüfungsbehörde zur Erlangung eines akademischen Grades vorgelegt.

Unterschrift: ____________________

Datum: ____________________

"You know that I write slowly. This is chiefly because I am never satisfied until I have said as much as possible in a few words, and writing briefly takes far more time than writing at length."

G. Simmons (Calculus Gems New York 1992).

UNIVERSITY OF ROSTOCK

Abstract

Faculty of Economic and Social Sciences
Department of Economic

Doctor rerum politicarum (Dr. rer. pol.)

Modeling Default Probability in Credit Portfolios: Evidence based on Debt Securities Issuers in Germany

by Lavri Labi

One of the main concern that investors have while buying debt securities is the assessment of securities issuers' creditworthiness. This is usually done by requesting for ratings of securities' issuers, produced by external rating agencies such as Standard & Poor's, Moody's and Fitch. However, since the financial crisis of 2007/2008, there is an increasing mistrust in external ratings and therefore investors now find the need to develop internal rating systems for a more reliable result. However the implementation of reliable rating systems is hindered by non-observed default events in databases and short time series of data available. In this study we propose an approach to handle those two challenges while developing rating systems. We further extend the approach by estimating systematic risk, that is, co-movements of creditworthiness of debt securities' issuers over time. Based on financial information from the PSVaG's debt securities portfolio, we demonstrate that one can adequately assess the creditworthiness of debt securities' issuers using publicly available data. Moreover, we could significantly increase the accuracy of the model by including a systematic risk component.

Acknowledgements

I would like to express my special appreciation and thanks to my supervisor Prof. Dr. Rafael Weissbach, you have been a tremendous mentor for me. I would like to thank you for encouraging my research and for giving me suitable advises. I would also like to thank Dr. Hubert Eckelmann for supporting me in the early stage of this thesis. I would like to thank the management of the PSVaG for providing data and support to conduct the study. A special thanks goes to Cuili Sun, which was always available for any question during the process of data preparation and results interpretation, and to Rebecca Gachago for her support during the whole time of this thesis and particularly for her excellent proofreading.

A special thanks to my family. Words cannot express how grateful I am to my mother and father for all of the sacrifices that you've made on my behalf. Your prayer for me was what sustained me thus far. I would also like to thank all of my friends who supported me in writing, and incented me to strive towards my goal.

Contents

List of Figures

List of Tables

Abbreviations

BaFin	**B**undes**a**nstalt für **Fin**anzdienstleistungsaufsicht
BIS	**B**ank of **I**nternational **S**ettlement
IFRS	**I**nternational **F**inancial **R**eporting **S**tandards
KwG	**K**redit**w**esen**g**esetz
MSE	**M**ean **S**quared **E**rror
MLE	**M**aximum **L**ikelyhood **E**stimator
PSVaG	**P**ension **S**icherungs**v**erein auf **G**egenseitigkeit
PD	**P**robability of **D**efault
RMLE	**R**estricted **M**aximum **L**ikelyhood **E**stimator
SME	**S**mall and **M**edium sized **E**nterprise
OLS	**O**rdinary **L**east **S**quare
PCA	**P**rincipal **C**omponent **A**nalysis
WLS	**W**eighted **L**east **S**quare

Chapter 1

Introduction

To proceed with investment and business expansion activities corporates occasionally have to invest more capital in their businesses than available in their assets. Issuing stocks and bonds provides them with the necessary capital. In comparison to other securities, debt securities offer a higher rate of return however there is a risk of default with corporates not being able to meet their repayment obligations to lenders. Due to this risk of default, lenders perform due diligence checks including an analysis of the creditworthiness of the securities' issuers before acquiring the issued securities. Using financial statements, lenders conduct financial analysis of the securities' issuers, that is, calculating financial ratios and comparing results to the industry sector for the vertical analysis and/or to past ratios for the horizontal analysis.

With the introduction of rating-systems in the early 20th century lenders are provided with an additional tool to assess the creditworthiness of the securities' issuers. Rating-systems provide rating-grades, which reflect the default probability of securities' issuers. The rating-grades are the categorized prediction of multivariate models, which are based on a large number of financial characteristics and take historical information from a high number of securities' issuers into account.

Rating-systems have enormously simplified the credibility analysis by providing an unique grade to assess the creditworthiness of securities' issuers and therefore the rating-systems have gained significant importance over time in the financial community. Lenders have been highly dependent on rating-grades as unique indicators that are taken into account before acquiring debt securities. Consequently, misleading rating-grades of mortgage securities were one of the elements leading to financial crisis in 2008.

As one of the consequence of the financial crisis of 2007/2008, the regulators now encourage corporates that invest in securities not to rely on external ratings alone, but also to develop internal rating systems to support their investment decisions. This is the case for the Pension Sicherungsverein auf Gegenseitigkeit (PSVaG). The PSVaG is a mutual organization. It is authorized to levy contributions under public law, but organized under civil law. The funds required to provide insolvency insurance come from legally mandated contributions paid by employers who provide old age pensions in the form of benefits subject to insurance. The PSVaG invest his asset only in securities of issuers in Germany. Therefore the PSVaG has a relative small portfolio of securities' issuers that is mainly composed of financial institutions and a small number of non-financial institutions such as local governments.

In this study we aim to develop a statistical model to predict the probability of default (PD) of securities' issuers in the PSVaG-Portfolio. For reason of homogeneity, the focus will be set on security issuers from the financial industry. Since historical default has not yet been observed in the PSVaG-Portfolio, external ratings will be used as credit performance indicators, that is the performance in repaying the credit. In the first step, information from financial statements of the securities' issuers will be used as explanatory variables in the model. In the second step systematic risk will be taken into account by adding corresponding indicator in the model. The following research questions will therefore be addressed: a) using rating-grades as proxy for default indicator, b) developing a cross-sectional model based on a small number of observations, c) developing longitudinal model for short time series of sparse data and d) estimating systematic risk.

In order to select suitable information for the model, one needs to have an understanding of the financial statements of corporates and, in this context, the particularities of financial reporting in the financial industry. We will therefore make a general presentation of the different steps of corporate valuation (incl. specificity of financial corporates) in chapter II. The chapter III will present detail of the data preparation. In Chapter IV we will describe our methodological approach for the model development. Chapter V and VI will present resulting model and related comments. Chapter VII will close the study by recapitulating the result and indicating further research possibilities to enhance the research findings.

Chapter 2

Identification of Relevant Information and Related Studies

In this study we aim to develop a model to predict default probability of financial corporates using information from financial reports. We use a method of corporate valuation to identify relevant financial information and adapt it to the specificity of financial corporates. We further review related empirical studies to identify additional relevant financial information and methodological approaches. We then finally put the resulting set of financial variables in relation with the credit performance indicator. The latter is determined by external rating, since our dataset does not contain corporates that have defaulted.

2.1 Corporate Valuation

At least once every 12 months, corporates publish financial statements comprised of a balance sheet, income statement, cash flow and stockholders' equity. The balance sheet makes a snapshot of the asset, liabilities and equity of the corporate at the end of the stated accounting period (usually the end of the year). The income statement determines the earnings of the corporate for the stated accounting period (usually a year) by describing how asset and liabilities were used. The cash flow statement explains inflows and outflows to determine the cash balance held by the corporate at the end of the stated accounting period. The stockholders' equity statement describes the change in the different types of stockholder's equity and determines the total stockholders' equity during the accounting period. See ifr [1] for more about International Financial Reporting Standards.

There are different approaches in valuating corporates. The market valuation uses market prices of shares to valuate publicly traded firms, which is not the case for the financial corporates that we are going to analyze. The discounted cash flow statements focus on discounting expected cash flows to present value in order to value corporates. This is not interesting for our study, since we intend to use a larger set of financial information for our prediction. We will therefore focus on the multiple approach, which generate ratios based on the information in the financial statements and then compare them to past values and/or values of similar corporates in the same industry sector. The ratios are generally generated from following categories:

Growth ratios describe the relative growth of corporate over time. The higher the growth rate the positiver the corporate value.

Profitability ratios indicate the efficiency in using resources to generate profit and shareholders' value. High level of profitability have a positive effect on corporate value.

Leverage ratios indicate the level of debt in the corporate and his relation to equity. High level of Leverage has a negative influence on corporate value.

Liquidity ratios give indications on the ability of corporate to pay-off it short-term debt obligations. High liquidity level has a positive effect on corporate value.

Management Efficiency ratios give indications on the ability of corporate to efficiently use human capital.

2.2 Specificity of Financial Corporate

Unlike non-financial corporates, which generate revenues by selling products and services, financial corporates generates their main income from interest spreads between deposits and loans. Fees generated from the services sold by financial corporates, e.g. charges for account administration, constitute a small part of their income. There is therefore a difference in the structure of financial statements and relevance of financial related information items for financial corporates versus non-financial corporates.

2.2.1 Balance Sheet

Most of the non-current asset of financial corporate are traded on the market, so that their actual value is reported in the balance sheet. In contrary most of the non-current asset of non-financial corporate are not reported with their actual value in the balance sheet, unless their value have been estimated based on an expensive process usually determined by regulatory requirements. Financial corporates have more asset classes than non-financial corporate, such as mortgage servicing rights, federal funds sold and trading assets. This is also the case for liability side with a liability classes, such as deposits, federal fund purchased and commercial papers. An other particularity of Financial corporate is the asset class allowance for loan losses, which is in fact a contra-asset, because one subtract it from the gross loans to get the net loans. The allowance for loan losses is designed to compensate the expected amount of loan defaulted.

2.2.2 Income Statement

The income statement of financial corporates has a different structure than the non-financial corporate one. It is divided into two parts, a) non-interest revenue containing incomes such as investment banking fees, lending- and deposit related fees, asset management, administration and commissions etc, and b) interest income and related expenses.

2.2.3 Cash Flow

Cash Flow Statement of financial and non-financial corporate have a similar structure containing operating activities, investing activities and financing activities. However, measuring reinvestment for financial corporate based on financial indicators such as free cash flow is problematic. In order to calculate free cash flow, one needs to determine net capital expenditure and working capital. We arrive at working capital by calculating the difference between current assets and current liabilities. Because of the particularity of business model of financial corporates, which use debt as their raw material, classification in current asset resp. current liabilities is misleading, so that the resulting working capital will not reflect the reinvestment activities of the financial corporate. In the case of net capital expenditure, the difference is that, non-financial corporates invest in fixed asset i.e. equipment, where financial corporates invest primarily in intangible assets such as brand name. Therefore the investment of financial corporate are "misclassified" as operating expense in accounting statement. Consequently the cash flow statement of financial corporate shows low or unexciting capital expenditure.

2.2.4 Regulation

The financial services sector is a highly regulated industry. Banks are the main focus of this study and they are, in particular required to maintain regulatory capital ratios, that are determined on the basis of their operations and the value of their equity. This is a measure to reduce the risk of depositors and has implications on the investment activities of banks. For more information about bank regulation see [2].

Resulting from the specificity of financial corporates two categories of financial ratios have been add to our previous analysis, so that our final list of financial ratio is as follows:

Growth ratios describe the relative growth a corporate over time. Higher growth rates depict a more positive corporate value.

Profitability ratios indicate the efficiency in using resources to generate profit and shareholders' value. High level of profitability have a positive effect on corporate value.

Leverage ratios indicate the level of debt in the corporate in relation to equity. A high level of Leverage has a negative influence on corporate value.

Liquidity ratios give indications on the ability of corporate to pay-off it short-term debt obligations. A high liquidity level has a positive effect on corporate value.

Management Efficiency ratios focus in this study on the ability of corporate to efficiently use human capital.

Capital Adequacy ratios give indication on the quality of the capital of the corporate.

Regulatory Capital ratios are related to the requirement of Basel Committee on Banking Supervision (see [2]) as explained in section 2.2.4.

Income Structure ratios give an indication on the income distribution between interest and non-interest revenue.

2.3 Credit Performance Indicator

The credit performance indicator is usually a binary vector taking value 1 if the borrower has defaulted and 0 in the case of no default. In the case of banks and financial institutions insolvency is rarely observed. Due to their high relevance in the economy, banks and financial institutions are mostly bailed out by their related states, taken over by other banks/financial institutions or merge with other banks/financial institutions. Refer to [3] for details about banks bailout-programs, acquisitions and fusions in Germany. One possibility is then to define financial criteria whereby banks and financial institutions are observed as defaulted and deduce the corresponding default vector. An other is to use an external rating as a proxy for the credit performance indicator. This approach is called shadow rating (see Engelmann and Rauhmeier [4] p. 39 and Cardoso et al. [5]). We will use the shadow rating approach in this study.

2.4 Related Studies

Because of their high relevance for the economy, a large number of research works have been conducted on predicting banks' bankruptcies. However research publications on quantitative assessment of creditworthiness of banks in Germany are scarce. Porath [6] uses Germany's central bank data to assess creditworthiness of banks in Germany based on both bank specific indicators and macroeconomic factors. He performed the estimation using a panel binary response model and could show that estimating both idiosyncratic and systematic risk simultaneously significantly increase the accuracy of the prediction in contrary to findings of Nuxoll [7] on the American Banks. Porath [6] result is supported by the findings of Hamerle et al. [8]. In their study they analyze the creditworthiness of a large number of German firms. They define a time discrete hazard rate related to the asset value process of the firms. They then propose a linear panel model including time-lagged firm specific risk drivers, macroeconomic factors and contemporary systematic random effect to estimate the underlying asset value. The latter is linked to the default indicator through logit and probit.

The majority of the publication on corporate default prediction focus on implementing cross-sectional model approaches using econometric/statistic and/or machine learning methods. Drobetz and Heller [9] use an ordered logit regression model to predict rating-grades of SMEs and large corporates in Germany based on both quantitative and qualitative factors. They could show the relevance of qualitative factors in assessing creditworthiness of corporates in Germany.

Outside Germany one can mention the study of Andersen [10], which uses a logit regression model to assess creditworthiness of Norway's banks and identifies relevant financial indicators. Canbas et al. [11] analyze the Turkey case using PCA to identify relevant factors based on banks' financial indicators and perform probit, logit and discriminant analysis on them. Espahbodi [12] and Tatom [13] analyze the creditworthiness of US Banks using logit resp. probit model.

In their study Boyacioglu et al. [14] make a comparative analysis of statistical and machine learning methods on their ability to predict bank failure in Turkey. They found out that neuronal network with multi-layer perceptron outperforms the other methods. Kumar and Ravi [15] give a review of statistical and intelligent techniques used for bankruptcy prediction of banks and firms.

Chapter 3

Data Preparation

3.1 Data Description

As described in section 4.3, our model for the cross-sectional approach is based on external ratings as dependent variable and financial ratios as predictors.

3.1.1 External Ratings

External ratings of 102 security issuers in PSV investment-portfolio are available. The ratings are provided by the three main rating agencies S&P, Mood's and Fitch for the time period 1995 - 2015 on a yearly basis.

For each securities' issuer we use one rating-grade per year. Since some securities' issuers have several rating-grades per year either from the same rating agency or from different rating agencies, we select the suitable rating-grade according to the BaFin-requirement as presented in [16]. The BaFin recommend in case of two rating-grades, to choose the "worst" rating-grade in the corresponding year, that is, the rating-grade with the highest probability of default. In case of more than two rating-grades, the BaFin recommend to choose the second best rating-grade. The rating-grade have been ordered according to the classification in figure 3.1.

Moody's		SP		Fitch	
Order	Grade	Order	Grade	Order	Grade
1	Aaa	1	AAA	1	AAA
2	Aa1	2	AA+	2	AA+
3	Aa2	3	AA	3	AA
4	Aa3	4	AA-	4	AA-
5	A1	5	A+	5	A+
6	A2	6	A	6	A
7	A3	7	A-	7	A-
8	Baa1	8	BBB+	8	BBB+
9	Baa2	9	BBB	9	BBB
10	Baa3	10	BBB-	10	BBB-
11	Ba1	11	BB+	11	BB+
12	Ba2	12	BB	12	BB
13	Ba3	13	BB-	13	BB-
14	B1	14	B+	14	B+
15	B2	15	B	15	B
16	B3	16	B-	16	B-
17	Caa1	17	CCC+	21	CCC
18	Caa2	18	CCC		
19	Caa3	19	CCC-		
21	Ca	20	CC		
		21	C		
22	C	24	D	22	DDD
23	/			23	DD
24	/			24	D

FIGURE 3.1: An Overview of the Rating Order

3.1.2 Financial Ratios

Financial data of 48 securities' issuers in PSV investment-portfolio are available. The financial data are provided by Bloomberg for the time period 1995 - 2014 on a yearly basis. The tables contain more than 300 financial indicators for the category Income Statement, Balance Sheet, Cash Flow, Ratio Analysis, Enterprise Value, Price Ratio Analysis, Environmental, Social and Corporate Governance and Valuation.

As presented in section 2.2 we extract relevant resp. compute new financial ratios based on this dataset and obtain the list below.

1. Growth
    - Sales - 1 Yr Growth
    - EBIT - 1 Yr Growth
    - Net Income - 1 Yr Growth
    - Assets - 1 Yr Growth
    - Net Worth - 1 Yr Growth
    - Capital - 1 Yr Growth
2. Profitability
    - Operating Margin
    - Pretax Margin
    - Profit Margin
    - Return on Assets
    - Return on Common Equity
    - Return on Capital
    - Financial Leverage
    - Annualized Return on Common Equity
    - Operating ROE
    - 5 Year Average Return On Equity
    - Return on Equity - 5 Yr Geometric Growth

- Net Fixed Asset Turnover
- Asset Turnover
- Dividend Payout Ratio

3. Leverage

- Assets/Equity
- Tangible Common Equity Ratio
- LT Debt/Common Equity
- LT Debt/Total Capital
- LT Debt/Total Assets
- Total Debt/Common Equity
- Total Debt/Tangible Book Value
- Total Debt/Total Equity
- Total Debt/Total Capital
- Total Debt/Total Assets
- Total Debt/Market Cap
- Net Debt/Shareholders Equity
- T12M CFO/Total Debt
- T12M FCF/Total Debt

4. Liquidity

- Total Debt/Total Capital
- CFO/Total Debt
- Cashflow/Total Liabilities

5. Management Efficiency

- Employees - 1 Yr Growth
- Net Income Per 1000 Employees
- Actual Sales Per Employee
- Assets Per 1000 Employees

6. Capital Adequacy
    - Tangible Common Equity Ratio
    - Tangible Common Equity to Risk-Weighted Assets
    - Tot loan/tot dep
    - Equity Ratio
    - Net Change in Liabilities
    - Increase In Equity
7. Regulatory Capital
    - Tier 1 Common Equity Ratio
    - Core Tier 1 Capital Ratio
    - Tier 1 Risk-Based Capital Ratio
    - Total Risk-Based Capital Ratio
    - Ratio of Risk weighted assets to Total assets (RWA/TA)
    - Changes in indicator (5) measured in percentage points (rwa/ta)
    - Est Basel III Tier 1 CE Ratio Fully Phased In
    - Est Basel III RWA
8. Income Structure
    - interest income / (interest + non-interest income)
9. Others
    - Sales/Cash
    - Sales/Marketable Securities
    - Sales/Net Fixed Assets
    - Sales/LT Investments
    - Sales/Other Assets
    - Sales/Total Assets
    - Per Share
    - Price/T12M Earnings per Share

- Book Value per Share

The following rules have been used in generating additional financial ratios such as "interest income/total income".

1. If at least one input-indicator is missing, the resulting output-ratio is set to missing.
2. If the denominator is zero, the resulting output-ratio is not a number. We will consider it as missing.
3. If the numerator is not zero and the denominator is near zero, the resulting output-ratio is a very large number. We will consider it as missing.

3.2 Plausibility Check

Before we start building our model, we are going to perform a check on the quality of our data and eventually correct the data where necessary. The first step in improving the data quality is the plausibility check, where one checks if the data structure fits within expectations partly based on expert knowledge.

3.2.1 External Ratings

The plausibility of the external rating is checked under two aspect, namely cross-sectional and longitudinal.

3.2.1.1 Cross-Sectional Plausibility Check

For the cross-sectional plausibility check we analyze the distribution of the ordered rating-grades for all the 102 security issuers at each point in time. One can observe three phases in the rating-grades distribution over time. The first phase is from 1995 to 2000, where the span of the rating-grades goes from 1 to 6 with concentration in class 1. In the second phase from 2001 to 2004 the distribution of the rating-grades are shifted to the right and span from 1 to 8 resp. 9. In the third phase from 2005 to 2008 the rating-grades concentrate between 1 and 6 and we observe relatively extreme rating-grades, that is, grades 11 and 12. The fourth phase starts in 2009 and ends in 2015 with rating-grades from 1 to 9 resp. 10. Related graphics can be found in appendix A. In the next section 3.2.1.2 we place a special focus on the issuers with relatively extreme rating-grades.

3.2.1.2 Longitudinal Plausibility Check

With the longitudinal plausibility check we analyze for each relevant issuer the distribution of the rating-grades over time. The relevant issuers are the issuers for which both financial information and rating-grades are available as indicated in section 3.3. We observe that apart of the LFA Bayern, the SAB Sächs.Aufb., the KFW, the FMS Wertmanagement and the Erste Abw.Anst. which shows a constant rating-grade of 1 resp 2 over time, the rating-grade of all other security issuers tend to deteriorate over time with change of one or two grades from one year to an other. Particular remarkable is the HRE from which the rating-grade jump from 6 to 9 between 2007 and 2008. This seems to be plausible since the HRE was one of the most affected financial institute in Germany by the financial crisis of 2007/2008.

3.2.2 Financial Ratios

In this section we perform both cross-sectional and longitudinal plausibility check. We start by analyzing the distribution of each financial ratio at each point in time with the objective of finding outliers or heterogeneous structures and check their plausibility by performing a longitudinal analysis of the related outlier. See appendix B and C for examples of corresponding graphics.

1. Growth

   Sales 1 Yr Growth do not show remarkable structure heterogeneity over time. Extreme value are observed almost in each point of time, but not constantly for the same security issuers. For example LBK Bremen, HRE and IKB have an enormous positive growth of sales 1 Yr in 2008, which get negative the following years.

   EBIT 1 Yr Growth do not show remarkable structure heterogeneity over time. However outliers are observed in 1997, 2005, 2008, 2011, 2012 and 2013 for different security issuers. For example in 2005 we observe an enormous growth of EBIT 1 Yr of DEKABANK, while it span approximately between -100 and 70 for the other years.

Net Income 1 Yr Growth do not show remarkable structure heterogeneity over time. However outliers are observed in 2008, 2010, 2011, 2013. In particular IKB shows strong negative value of -5320.5607 in 2008, Düssel Hyp -1865.5872 in 2011, HSH -576.6667 in 2013 and Sab Sächs.Aufb. a strong positive value of 16623.3094 in 2010.

Asset 1 Yr Growth do not show remarkable structure heterogeneity over time. Outliers are observed in 1996, 1997, 1999, 2005 and 2007. This is the case for example for HRE with a value of 147.6432 in 2007, while it span approximately between -3 and 6 for the other years.

Net Worth 1 Yr Growth do not show remarkable structure heterogeneity over time. Outlier is observed in 2005 with NRW Bank showing a value of 423.3441.

Capital 1 Yr Growth do not show remarkable structure heterogeneity over time. Extreme value are observed in 1996, 1997, 1999, 2005 and 2007, that is for example HRE with a value of 156.7324 in 2007.

2. Profitability

Operating Margin do not show remarkable structure heterogeneity over time. However outliers are observed in 2000, 2007, 2008, 2009 2011, 2012, 2013. This is the case in particular for KFW which shows strong negative operating margin in 2007 (-598.4466) and 2008 (-1437,0558), IBB in 2011 (-673.8318), Landwirtschaftliche Rentenbank in 2009 (-418.8172) and 2011 (-1621.4286) and Dexia Kommunal in 2012 (-540.3571)

Pretax Margin do not show remarkable structure heterogeneity over time. Extreme values are observed in 2000, 2003, 2006, 2007, 2009, 2011 and 2013. Eurohypo shows a particularly high negative value of -102.4622 in 2011.

Profit Margin do not show remarkable structure heterogeneity over time. Outlier are observed in 2011 and 2012. In particular Dexia Kommunal shows an high negative value in 2012 (-540.3571) and in 2011 IBB (-672.8972) and West Immo (-364.5012).

Return on Asset do not show remarkable structure heterogeneity over time. Extreme values are observed in 2007, 2009, 2010 and 2013, such as SAB Sächs.Aufb. with a value of 2.7187 in 2010.

Return on Common Equity do not show remarkable structure heterogeneity over time. Extreme value are observed in almost each year, such as FMS Wertmanag. with 151.7425 in 2013 or Eurohyp with -94.6038 in 2011.

Return on Capital do not show remarkable structure heterogeneity over time.

Financial Leverage do not show remarkable structure heterogeneity over time. Extreme value are observed in 2009, 2012 and 2013. This is the case for HRE in 2009 (378.6832), Dexia S.A. in 2012 (923.2242) and FMS Wertmanag. in 2013 (2167.1232).

Annualized Return on Common Equity do not show remarkable structure heterogeneity over time. Extreme value are observed in 2000, 2008, 2009, 2011 and 2013. This is the case i.a. for FMS Wertmanagement with 151.7425 in 2013.

Operating ROE do not show remarkable structure heterogeneity over time. Extreme value are observed in 2000, 2008, 2009, 2011 and 2013. This is the case for FMS Wertmanagement with 151.7425 in 2013.

5 Year Average Return on Equity do not show remarkable structure heterogeneity over time.

Return on Equity - 5 Year Geometric Growth do not show remarkable structure heterogeneity over time. Extreme value are observed in 1997, 1999, 2006, 2011 and 2013. This is for example the case for Berlin Hyp, West Immo and WL Bank in 2013 with value -100.0000.

Net Fixed Asset Turnover do not show remarkable structure heterogeneity over time. Extreme value are observed in all years. This is the case i.a. for FMS Wertmanagement with 10094.0602 in 2012 and 10704,7308 in 2013.

Asset Turnover do not show remarkable structure heterogeneity over time.

Dividend Payout Ratio do not show remarkable structure heterogeneity over time. Extreme value are observed in several years. This particularly the case i.a. of Dt. Bank with 216.0428 in 2001, 224.4444 in 2002, 290.4943 in 2012 and 114.8649 in 2013.

3. Leverage

Asset/Equity do not show remarkable structure heterogeneity over time. Extreme value are observed in 2000, 2008, 2011, 2012, 2013. This is in particular the case for FMS Wertmanag. with value 5162.2288 in 2012 and 1281.1579 in 2013.

Tangible Common Equity Ratio do not show remarkable structure heterogeneity over time. Extreme value are observed 2004, 2005, 2006, 2007, 2008, 2009 and 2013. This in particular the case for NRW Bank, for which the value stays high since 2005 (15.1021).

LT Debt/Common Equity do not show remarkable structure heterogeneity over time. Extreme value are observed in several years, particularly in 2012 and 2013 with value of 233600.4673 resp. 72087.4526 for FMS Wertmanag.

LT Debt/Total Capital, LT Debt/Total Assets do not show remarkable structure heterogeneity over time.

Total Debt/Common Equity, Total Debt/Tangible Book Value do not show remarkable structure heterogeneity over time, with FMS Wertmanag. as outlier.

Total Debt/Total Equity, Total Debt/Total Capital do not show remarkable structure heterogeneity over time, with FMS Wertmanag. as outlier.

Total Debt/Total Assets, Total Debt/Market Cap do not show remarkable structure heterogeneity over time, with FMS Wertmanag. as outlier.

Net Debt/Shareholders Equity do not show remarkable structure heterogeneity over time. Extreme value are observed in 2000, 2012 and 2013, particularly for FMS Wertmanag. with value 240558.1149 and 70271.3846 in 2012 resp. 2013.

T12M CFO/Total Debt, T12M FCF/Total Debt do not show remarkable structure heterogeneity over time.

4. Liquidity

Total Debt/Total Capital do not show remarkable structure heterogeneity over time. HSBC Trinkaus with an extreme low value from 2010 to 2013.

CFO/Total Debt do not show remarkable structure heterogeneity over time.

Cashflow/Total Liabilities do not show remarkable structure heterogeneity over time.

5. Management Efficiency

Employees - 1 Yr Growth do not show remarkable structure heterogeneity over time. However extreme value are observed in all years.

Net Income Per 1000 Employees heterogeneous structure over time, with several extreme value in each year.

Actual Sales Per Employee do not show remarkable structure heterogeneity over time. However extreme value are observed in all years.

Assets Per 1000 Employees do not show remarkable structure heterogeneity over time. However extreme value are observed in all years.

6. Capital Adequacy

Tangible Common Equity Ratio do not show remarkable structure heterogeneity over time. However extreme value are observed in several years.

Tangible Common Equity to Risk-Weighted Assets do not show remarkable structure heterogeneity over time. However extreme value are observed in several years.

Tot loan/tot dep do not show remarkable structure heterogeneity over time. However extreme value are observed in several years.

Equity Ratio is not available in all years.

Net Change in Liabilities do not show remarkable structure heterogeneity over time. However extreme value are observed in all years.

Increase In Equity do not show remarkable structure heterogeneity over time. However extreme value are observed in all years.

7. Regulatory Capital

Tier 1 Common Equity Ratio is only available from 2011 to 2014 for a small number of securities' issuers.

Core Tier 1 Capital Ratio is available from 2005 to 2014 with an heterogeneous structure over time.

Tier 1 Risk-Based Capital Ratio do not show remarkable structure heterogeneity over time.

Total Risk-Based Capital Ratio do not show remarkable structure heterogeneity over time.

Ratio of Risk weighted assets to Total assets (RWA/TA) is available from 2001 to 2014 and do not show remarkable structure heterogeneity over time.

Changes in indicator (5) measured in percentage points (rwa/ta) is available from 2004 to 2014 with an heterogeneous structure over time.

Est Basel III Tier 1 CE Ratio Fully Phased In is only available from 2011 to 2014 for a small number of securities' issuers.

Est Basel III RWA is only available from 2013 to 2014 for a small number of securities' issuers.

8. Income Structure

interest income / (interest + non-interest income) do not show remarkable structure heterogeneity over time. However extreme value are observed in all years.

3.2.3 Concluding Remarks to Plausibility Check

Several securities' issuers such as IKB, Eurohyp and HRE have shown extreme value at certain points in time for both plausibility checks of ratings and financial information. This is partly due to the impact of the financial crisis in 2007/2008. Particularly HRE had big troubles, so that FMS Wertmanag. have been created to baillout HRE. FMS Wertmanag. is a "Bad Bank". According to the definition in [17], a "Bad Bank" is a bank set up to buy the bad loans of a bank with significant nonperforming assets at market price. FMS Wertmanag. is not a financial corporate as define in KwG(see [18]) and is therefore not required neither to fulfill capital requirement as recommended in Basel II resp. Basel III nor to provide financial reporting according to IFRS. Those facts are reflected in the observations of the plausibility check, where FMS Wertmanag. shows extreme value in almost all financial ratios. We will therefore remove FMS Wertmanag. from our basis dataset.

As an other consequence of the financial crisis of 2007/2008, Basel III have been agreed (see [19]). It has taken effect in Germany in 2014, so that related financial ratio such as "Est Basel III Tier 1 CE Ratio Fully Phased In" and "Est Basel III RWA" are only available for very few years and a small number of securities issuers. We therefore decide not to take both financial information in account in our basis dataset.

3.3 Data Merging and Consolidation

The data merging and consolidation is the last step in generating the basis dataset for our model. It consist in joining the dataset with financial information to the dataset with rating-grades, while taking into account possible differences in key-columns due for example to fusion of financial institutes. From the relevant securities issuers only the LBK Berlin change his name to be Spk Berlin. The merging and consolidation of both dataset results in a dataset of 38 security issuers with rating-grades from 1995 to 2015, because no rating-grades were available for Aareal, Dexia Kommunal, Dexia S.A., HSBC Trinkaus, HSH, Hypovereinsbank, ING BHF, Landwirtschaftliche Rentenbank, Warburger Hypo.

Chapter 4

Methodology

As mention in section 2.3 we propose the shadow-rating approach as described in Engelmann and Rauhmeier [4] p. 39 and Cardoso et al. [5] to build the rating model, since no insolvency have been yet observed in PSVaG-Portfolio.

The shadow-rating approach consist in predicting external PDs[1] using risk factors. In the case of reduced amount of observations, it is a common practice in banks and financial institutions to pool panel data and perform a cross-section analysis, while ignoring time dependencies. Since we are facing the same problem of restricted amount of data in this study, we start the analysis by following the same approach that is used to perform a cross-sectional analysis. We then propose a longitudinal analysis approach to take time dependencies into account.

In this chapter we present the methodological background of our analysis. We start by presenting, in section 4.1 the method selected to impute missing value in the data. In section 4.2 the problem of multicollinearity is explained and methods to handle it are shown. Section 4.3 details our cross-sectional shadow rating approach. Section 4.4 describes our longitudinal shadow rating approach. Section 4.5 presents the chosen methods to assess goodness of fit. We then conclude in section 4.6 by highlighting our scientific contribution.

[1] external PDs corresponding to external rating-classes

4.1 Missing Value Imputation

In order to use the largest possible amount of observations in our data, an imputation of missing values is necessary. Several approaches such as mean- and regression-imputation are available in the literature. The method chosen for this study is the nearest-neighbor approach. It belong to the case based reasoning (CBR) in the area of artificial intelligence and consist of a search of the nearest neighbor/neighborhood of the observation with the missing value in the hyperspace (with axis defined by the variables) and undertake his corresponding value.

The proximity of neighbors is determined by distance metrics. In a space defined by an orthogonal basis, one determines the distance between two points by translating the corresponding vector to the origin and using Pythagoras theorem to estimate his norm. In the particular case of a two dimensional space, the distance d between point $A(x_1, y_1)$ and point $B(x_2, y_2)$ is $d = \left\|\vec{AB}\right\| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$. This metric is known as the Euclidean distance. In case axes x and y are correlated, using the Euclidean distance leads to a biased estimation of distance d. Pythagoras theorem assumed orthogonality of x and y, which is no more the case, because of the correlation.

An approach to handle the problem lies in transforming the axes, while taking their covariance matrix into account. Such a metrics is named Mahalanobis distance and calculated as follow:

$$d(A, B) = \sqrt{(A - B)'\Sigma^{-1}(A - B)}, \tag{4.1}$$

with Σ^{-1} the covariance matrix. Since our variables are partly correlated, we chose the Mahalanobis distance to perform the nearest-neighbor-analysis. For further detail about nearest-neighbor-analysis see Hastie et al. [20] p. 14-18 and 415-431.

4.2 Multicollinearity Problem and Handling

Running a linear regression analysis is submitted to the technical restriction, that extreme multicollinearity makes a parameter estimation impossible. This is due to the fact that, when at least two variables are correlated with factor 1, the inverse $(X'X)^{-1}$ is impossible to compute, since it doesn't have full rank (his rank k is smaller than the number of variables p). See section 4.3 for parameter estimation of linear regression model.

Weak multicollinearity, that is, when at least two variables are highly correlated with a factor less than 1, also leads to problematical parameter estimation in linear regression models. Even if in this case the design matrix $X'X$ has full rank ($k = p$), its determinant $|X'X|$ tend to have a value near 0, which may lead to a very large variance of the estimated regression parameters $\hat{\beta}$. It may hence result in very large confidence intervals for β and a high probability of obtaining implausible estimations of regression parameters. For more details about multicollinearity refer to Toutenburg [21] p. 44-49.

There exist several methods to handle multicollinearity. For this study we have chosen 1) the variables selection and 2) the principal component analysis approaches.

4.2.1 Variables Selection Approach

For this approach we start with performing a pairwise correlation analysis to identify correlated variables. We build groups of correlated variables and choose one variable as a representative for each group. One can then ignore the other variables in the groups, since they have very similar information like the representative variables. The marginal difference of information is neglected.

We perform the pairwise correlation analysis using the Pearson's correlation coefficient. Refer to Agresti [22] for more details about correlation coefficients.

4.2.2 Principal Component Analysis

In opposition to the variables selection approach in section 4.2.1 no high correlated variables are excluded from the predictors set for the principal component analysis. The latter consist of a rotation of the basis (vector of predictors as axes) in order to fulfill the two following conditions:

1. the new axes are orthogonal with each other,
2. the axes are chosen so that the greatest variance of the point cloud comes to lie on the first component, the second greatest variance on the second component, and so on.

We denote by X the $n \times p$ matrix of explanatory variables. Our objective is to obtain the $n \times p$ matrix Y of orthogonal vectors $y_j (j = 1, \cdots, p)$ as a projection of X based on projection matrix P, that is, $Y = PX$. Y being composed of orthogonal vectors means, that her covariance matrix $Y'Y$ is diagonal. We therefore have

$$\begin{aligned} Y'Y &= PX(PX)' \\ Y'Y &= PX'XP'. \end{aligned} \tag{4.2}$$

Since $X'X$ is a symmetric positive (semi-)definite matrix and $Y'Y$ is a diagonal matrix, it is obvious that the solution of equation 4.2 lies in determining P so, that her row vectors are the eigenvector of matrix $X'X$ and the diagonal element of $Y'Y$ are the corresponding eigenvalue. The eigenvalues are ordered from the highest to the lowest as required by condition 2 above. Moreover, since one observe $X'X$ as the variances-covariances matrix, one implicitly admit $E(X) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$, because

$$\begin{aligned} var(x_1) &= \sum_{i=1}^{n} (x_{i1} - \overbrace{E(x_1)}^{=0})^2 \\ var(x_1 x_2) &= \sum_{i=1}^{n} (x_{i1} - \overbrace{E(x_1)}^{=0})(x_{i2} - \overbrace{E(x_2)}^{=0}). \end{aligned}$$

Therefore one need to center (subtract mean) from the variables before performing principal component analysis. Refer to Hastie et al. [20] p. 485-491 for further details about principal component analysis.

4.3 Cross-Sectional Shadow Rating Approach

Cross-Sectional Shadow Rating Approach consist in pooling the data, that is, one consider observations of the same securities' issuer at different points in time as independent. We define the random variable Y_{it}, which takes value 1 if securities issuer i defaults in the credit performance period starting at time t and 0 if not, with $i = 1, \cdots, I$ and $t = 1, \cdots, T$. I is the total number of securities' issuers and T the length of the panel data.

We observe y_{it} as the realization of an independent Bernoulli experiment of random variable Y_{it} (dependencies between entities are neglected), that is

$$Y_{it} \sim Bernoulli(\pi_{it}) \tag{4.3}$$

with $\pi_{it} = P(y_{it} = 1)$ the probability of default. We denote by x_{it} the vector of specific risk factors of securities issuer i at time t. The standard approach in developing credit risk model for banks and financial institutions consists of estimating the parameters of the corresponding logistic regression model,

$$\log\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \beta' x_{it} \tag{4.4}$$

with β the vector of regression coefficients. This approach requires a default indicator vector with sufficient default events. In the case of this study, insolvency has not been observed and additional information to deduce defaults from are not available. We therefore adapt the modeling approach by assuming that the underlying probability of default π_{it}^{PD} of the rating-grade of securities issuer i at time t correspond to the underlying probability of default π_{it}, that is,

$$Y_{it} \sim Bernoulli(\pi_{it}^{PD}). \tag{4.5}$$

Since the probability of success of the Bernoulli distribution is now known, we are interested in estimating the parameters of the corresponding linear regression model,

$$\log\left(\frac{\pi_{it}^{PD}}{1-\pi_{it}^{PD}}\right) = z_{it} = \beta' x_{it} + \epsilon_{it} \text{ with } \epsilon_{it} \sim N(0,\sigma^2). \tag{4.6}$$

$$\text{in matrix form it can be written as} \tag{4.7}$$

$$\log\left(\frac{\pi^{PD}}{\mathbb{1}-\pi^{PD}}\right) = Z = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4.8}$$

$$\boldsymbol{\epsilon} \sim N(0, \sigma^2 I) \tag{4.9}$$

with $\pi^{PD} = \begin{pmatrix} \pi_{11}^{PD} \\ \pi_{21}^{PD} \\ \vdots \\ \pi_{IT}^{PD} \end{pmatrix}$, $Z = \begin{pmatrix} Z_{11} \\ Z_{21} \\ \vdots \\ Z_{IT} \end{pmatrix}$, $X_{n,p} = \begin{pmatrix} X_{11}^1 & X_{11}^2 & \cdots & X_{11}^p \\ X_{21}^1 & X_{21}^2 & \cdots & X_{21}^p \\ \vdots & \vdots & \ddots & \vdots \\ X_{IT}^1 & X_{IT}^2 & \cdots & X_{IT}^p \end{pmatrix}$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_{11} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{IT} \end{pmatrix} \text{ and } \mathbb{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

A widely used method to estimate linear regression parameters consists in choosing the line or hyperplane for which the euclidean distance between it and the points cloud is minimal. In other words, one minimizes the mean squared error. We define the resulting estimation of the linear regression model by

$$\hat{Z} = X\hat{\boldsymbol{\beta}}. \tag{4.10}$$

The error also called residual is the difference between the depending variable Z and its estimation $\hat{Z}$. The mean squared error (MSE) is therefore defined as

$$\text{MSE} = (Z - \hat{Z})'(Z - \hat{Z}) = (Z - X\hat{\boldsymbol{\beta}})'(Z - X\hat{\boldsymbol{\beta}}). \tag{4.11}$$

Assuming full rank for the matrix X and hence that the corresponding design matrix $X'X$ is invertible, we obtain

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'Z \tag{4.12}$$

after minimization of the MSE with regard to $\boldsymbol{\beta}$. This is also the best linear unbiased estimation (BLUE) according to the Gauß-Markov-Theorem. For further details about linear regression analysis refer to Toutenburg [21] p. 21-94.

Neglecting time dependencies in panel data may considerably impact the accuracy of the model. We therefore propose in section 4.4 a model approach, which takes time dependencies into account.

4.4 Longitudinal Shadow Rating Approach

As presented in figure 4.1 one distinguishes three types of dependencies in credit portfolio, that is, the idiosyncratic risk, the concentration risk and the contagion risk. The idiosyncratic risk is particular to each securities' issuer, while the concentration risk affects groups or sub-groups of securities' issuers. Concentration risk happens when co-movements of creditworthiness are observed, due to securities issuers' exposure to common risk factors i.e. industry sector. Contagion risk is related to the bilateral/multilateral interactions between securities' issuers. It happens for instance when securities issuers are linked in their businesses, so that the activities/default of one has a direct impact on the activities/tendency to default of the other and vice versa. Both concentration and contagion risk constitute the systematic risk.

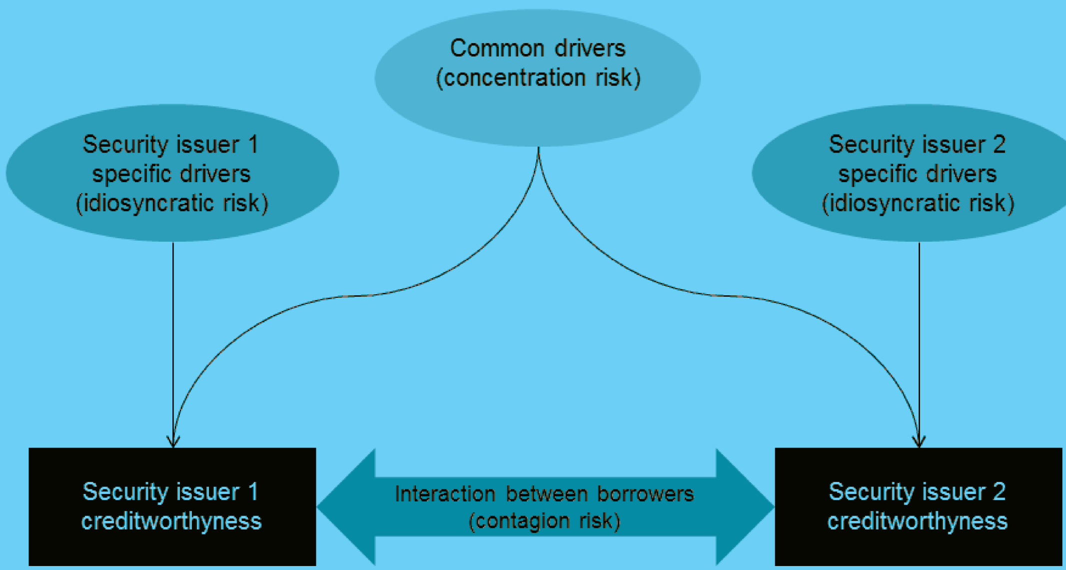


FIGURE 4.1: Overview Dependency Structure in Security Portfolio

In the cross sectional approach in section 4.3 creditworthiness of securities' issuers has been estimated based on idiosyncratic risk, while neglecting related systematic risk. In this section a model approach is proposed to take systematic risk into account by estimation of creditworthiness of securities issuers. Since additional information (which is not available here) is necessary for a suitable modeling of contagion risk, the focus will be set on concentration risk in this section.

But before proceeding with the presentation of our proposed model in sections 4.4.3, we will give in section 4.4.1 and 4.4.2 an overview of approaches in developing rating model accounting for systematic risk. In section 4.4.1 systematic risk is estimated without taking exogenous factors into account, while in section 4.4.2 macroeconomic factors are included.

4.4.1 Short Panel Data with Sufficient Observations at each Point in Time

For the case where sufficient observations per point in time are available, rating models are usually implemented using a two steps approach. The first step consist of estimating the idiosyncratic risk by training a cross-sectional model on the observation of the most actual time point T. In the second step the scores are calibrated while taking systematic risk into account.

The cross-sectional model consist of estimating the score z_i for each entity i for the most actual time point T according to equation (4.13). The parameter of the later are estimated as described in 4.3.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = z_i = \beta' x_i. \tag{4.13}$$

The calibration consist of categorizing the scores z_i matching them to the predefined rating-grades (see figure 4.2). Following conditions must be fulfilled while doing calibration.

1. The default rate in each rating-grade must not be significantly different from the corresponding expected PD.
2. The long term forecast default rate in the portfolio must not be significantly different from the predicted PD for the portfolio.

The predicted PD for the portfolio is computed as the weighted (number of entity in the rating-grade divided by total number of entity) sum of the expected PD for each rating-grade. The long term forecast default rate is estimated by the weighted (weights sum to one) sum of the historical default rates.

Refer to Engelmann and Rauhmeier [4] p. 1-37 and p. 307-345 for common practice to implement rating model.

Rating-Grades	Score-Intervals	PD
A	9,795 – 9,999	1.44 %
B	9,686 – 9,794	2.37 %
C	9,617 – 9,688	3.32 %
D	9,522 – 9,616	4.46 %
E	9,381 – 9,521	5.81 %
F	9,151 – 9,380	7.15 %
G	8,657 – 9,150	11.06 %
H	8,091 – 8,656	16.33 %
I	7,611 – 8,090	21.91 %
K	7,218 – 7,610	27.82 %
L	6,457 – 7,217	28.90 %
M	1 – 6,456	41.91 %

FIGURE 4.2: Schufa's Rating-Grades for SME

4.4.2 Long Times Series with (In)sufficient Observations at each Point in Time

For long panel data with sufficient or insufficient observations per point in time, rating models can be implemented by simultaneously estimating the idiosyncratic and systematic risk in the context of an hierarchic model. Let's define s_t, f_t and ξ_t respectively the systematic risk, the vector of macroeconomic factors and the random error at time t. γ is a coefficients vector. One can then extend the rating model (4.6) in section 4.3 as

$$s_t = \gamma' f_t + \xi_t \text{ with } \xi_t \sim N(0, \omega^2) \tag{4.14}$$

$$z_{it} = \beta' x_{it} + s_t + \epsilon_{it} \text{ with } \epsilon_{it} \sim N(0, \sigma^2). \tag{4.15}$$

We further admit independence of random errors ξ_t and ϵ_{it}. A method to estimate the model parameters consists of taking the means of entities at each point in time and estimates γ, that is,

$$z_{it} = \beta' x_{it} + \gamma' f_t + \xi_t + \epsilon_{it} \text{ with } \epsilon_{it} \sim N(0, \sigma^2). \tag{4.16}$$

$$\bar{z}_t = \overbrace{\beta' \bar{x}_t}^{w_t} + \gamma' f_t + \overbrace{\xi_t + \bar{\epsilon}_t}^{\epsilon_t^*} \text{ with } \bar{\epsilon}_t \sim N(0, \varsigma^2) \tag{4.17}$$

$$\bar{z}_t = w_t + \gamma' f_t + \epsilon_t^* \text{ with } \epsilon_t^* \sim N(0, V). \tag{4.18}$$

One can estimate γ using OLS (ordinary least square). Once γ is estimated, one can use it to estimate β using WLS (weighted least square).

Refer to Porath [6] and Hamerle et al. [8] for details about simultaneous estimation of systematic and idiosyncratic risk in the context of implementing rating model.

4.4.3 Short Panel Data with Scarce Observations at each Point in Time

The dataset for our study is characterized by a short panel data of scarce observations. Therefore the two steps approach proposed in section 4.4.1 can not be used. Since the number of observation to perform OLS for equation (4.18) depends of the length of the time series, the approach presented in section 4.4.2 is not applicable either.

We therefore propose a two steps approach to implement rating model based on short panel data of scarce observations. The first step extend the model in section 4.4.2 by observing the systematic risk s_i in equation (4.14) as a time variable constant, which increase or decrease the level of PD in the portfolio. The parameters of such a model can be estimated using a random effect model (section 4.4.3.1), for which there is no necessity to include any macroeconomic factors to estimate s_i.

The second step of our proposed approach is similar to the second step of the approach for "'short panel data of sufficient observations at each point in time"' in section 4.4.1. It consists of using the historical estimated systematic risk to predict its future value. Unlike the approach in section 4.4.1, which perform the prediction based on a simple weighted sum of historical values, we propose a double exponential smoothing for the prediction (section 4.4.3.2).

4.4.3.1 Linear Random Effects Model for Normal Data

We observe the probability of default of securities' issuers as related to the sum of weighted idiosyncratic and systematic factors. We further postulate the conditional independence of the probability of default and admit that given the systematic factors at time t, the probability distribution of securities' issuers are independent from each others. The model proposed is analog to the random effect model for normal data presented in Fahrmeir and Tutz [23] p. 285-288. The systematic factors are the estimated random effects and the idiosyncratic factors, the estimated fixed effects. Once we obtain the random effects, we rearrange them in the order of time t and apply a double exponential smoothing (section 4.4.3.2) to predict the future value $T+1$. This prediction of the systematic risk can then be used in combination with the idiosyncratic risk (resulting of the application of parameters for fixed effects) to assess creditworthiness of security issuers.

Taking the same annotation as in section 4.3 and adding index t for time we can write

$$Z_t = \overbrace{X_t\boldsymbol{\beta}_t}^{idiosyncratic} + \overbrace{U_t\boldsymbol{\gamma}_t}^{systematic} + \boldsymbol{\epsilon}_t \tag{4.19}$$

$$Z_t|\boldsymbol{\gamma}_t \sim N(X_t\boldsymbol{\beta}_t + U_t\boldsymbol{\gamma}_t, \Sigma_t) \tag{4.20}$$

$$\boldsymbol{\gamma}_t \sim N(0, D) \tag{4.21}$$

$$\epsilon_t \sim N(0, \Sigma_t) \tag{4.22}$$

$$\gamma_1, \cdots, \gamma_n, \gamma_1, \cdots, \gamma_n \text{ are independent} \tag{4.23}$$

with

$$Z_t = \text{vector (length } p\text{) of logit(PD) of all issuers at time } t, t = 1, \cdots, T$$

$$X_t = \text{matrix}(n_t \times p) \text{ of idiosyncratic predictors of all issuers at time } t$$

$$\boldsymbol{\beta}_t = \text{vector (length } p\text{) of coefficients of fixed effects}$$

$$U_t = \text{matrix}(n_t \times q) \text{ of systematic predictors of all issuers at time } t$$

$$\boldsymbol{\gamma}_t = \text{vector (length } q\text{) of coefficients of random effects at time } t$$

$$n_t = \text{number of issuer at time } t$$

In this study we group the entities (securities' issuers) only by time t, so that $q = 1$ and hence U_t is a unit vector of length n_t and γ_t a scalar. Putting securities' issuers of all points in time t together, one can define following matrices and vectors:

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix}, X_{n,p} = \begin{pmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,p} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n,1} & X_{n,2} & \cdots & X_{n,p} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

$$U_{n,(T\cdot q)} = \begin{pmatrix} U_1 & \mathbf{0}_{n_1\times q} & \cdots & \mathbf{0}_{n_1\times q} \\ \mathbf{0}_{n_2\times q} & U_2 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{n_T\times q} & & \cdots & U_T \end{pmatrix}, \mathbf{0}_{n_i\times q} = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix}, \boldsymbol{\gamma} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_T \end{pmatrix},$$

$$G = \begin{pmatrix} D & & \\ & \ddots & \\ & & D \end{pmatrix}, R = \begin{pmatrix} \Sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Sigma_T \end{pmatrix}.$$

We can therefore rewrite the regression equation analog to section 4.3.

$$Z = X\boldsymbol{\beta} + \boldsymbol{\epsilon}^* \tag{4.24}$$

$$\text{with } \boldsymbol{\epsilon}^* = U\boldsymbol{\gamma} + \boldsymbol{\epsilon} \tag{4.25}$$

Now we can determine the distribution of $\boldsymbol{\epsilon}^*$.

$$\boldsymbol{\epsilon}^* = \begin{pmatrix} U & I_{n\times n} \end{pmatrix} \begin{pmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\epsilon} \end{pmatrix}$$

from (4.21), (4.22) and (4.23) it follows

$$\begin{aligned} Var(\boldsymbol{\epsilon}^*) = V &= \begin{pmatrix} U & I \end{pmatrix} \begin{pmatrix} G & \mathbf{0} \\ \mathbf{0} & R \end{pmatrix} \begin{pmatrix} U' \\ I \end{pmatrix} \\ &= UGU' + R \end{aligned} \tag{4.26}$$

The model can then be written as

$$Z = X\boldsymbol{\beta} + \boldsymbol{\epsilon}^* \tag{4.27}$$

$$\boldsymbol{\epsilon}^* \sim N_n(\mathbf{0}, V). \tag{4.28}$$

Estimation of model parameters can be performed using different methods depending on variances R and G being known or unknown.

4.4.3.1.1 Parameters Estimation with Variances-Covariances Components known

For R and G known the maximum likelihood estimator (MLE) coincides with the solution of the weighted least square. One can therefore proceed analog to section 4.3 and find a suitable estimation for $\boldsymbol{\beta}$ by choosing the line or hyperplane for which the distance between it and the points cloud is minimal. But the difference here is, that the selected distance metrics is the Mahalanobis distance, because of the auto-correlation of the disturbance term $\boldsymbol{\epsilon}^*$ (unlike $\sigma^2 I$ in section 4.3, V is non-diagonal). That is

$$\mathrm{MSE} = (Z - \tilde{Z})' V^{-1} (Z - \tilde{Z}) = (Z - X\tilde{\boldsymbol{\beta}})' V^{-1} (Z - X\tilde{\boldsymbol{\beta}}). \tag{4.29}$$

After minimization of the MSE with regard to $\boldsymbol{\beta}$, we obtain:

$$\tilde{\boldsymbol{\beta}} = (X' V^{-1} X)^{-1} X' V^{-1} Z. \tag{4.30}$$

(4.20) and (4.20) imply $Z \sim N(X\boldsymbol{\beta}, V)$ and $\boldsymbol{\gamma} \sim N(\mathbf{0}, G)$. Therefore

$$\begin{pmatrix} Z \\ \boldsymbol{\gamma} \end{pmatrix} \sim N\left(\begin{pmatrix} X\boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} V & UG \\ GU' & G \end{pmatrix} \right). \tag{4.31}$$

Based on (4.31) one can use the related rule for multivariate normal distributed variables to determine the best linear unbiased predictor (BLUP) for γ, which is the conditional expectation of γ given Z,

$$E(\boldsymbol{\gamma}|Z) = GU' V^{-1} (Z - X\boldsymbol{\beta}). \tag{4.32}$$

The empirical BLUP (EBLUP) can then be determined by substituting $\boldsymbol{\beta}$ by $\tilde{\boldsymbol{\beta}}$ in equation (4.36), so that $E(\tilde{\boldsymbol{\gamma}}|Z) = GU' V^{-1} (Z - X\tilde{\boldsymbol{\beta}})$.

4.4.3.1.2 Parameters Estimation for Unknown Variances-Covariances Components

In the case of unknown variances-covariances components one defines the parameter θ, which collects the unknown variances-covariances component G and R, that is $\boldsymbol{\theta} = (G, R)$, and determines the log-likelihood function $l(\boldsymbol{\beta}, \boldsymbol{\theta})$ of the marginal model (4.45), that is,

$$l(\boldsymbol{\beta}, \boldsymbol{\theta}) = -\frac{1}{2}(\log|V(\boldsymbol{\theta})| + (Z - X\boldsymbol{\beta})'V(\boldsymbol{\theta})^{-1}(Z - X\boldsymbol{\beta})). \tag{4.33}$$

One can then determine an estimation of $\boldsymbol{\beta}$ conditioned on $\boldsymbol{\theta}$ by maximizing the log-likelihood function $l(\boldsymbol{\beta}, \boldsymbol{\theta})$ for a fixed value of $\boldsymbol{\theta}$ with regard to $\boldsymbol{\beta}$. One obtains analog to subsection 4.4.3.1.1

$$\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}) = (X'V^{-1}(\boldsymbol{\theta})X)^{-1}X'V^{-1}(\boldsymbol{\theta})Z. \tag{4.34}$$

One replaces β by its estimation in the log-likelihood function $l(\boldsymbol{\beta}, \boldsymbol{\theta})$ to obtain the profile log-likelihood $l_p(\boldsymbol{\theta})$, which is only dependent on $\boldsymbol{\theta}$.

$$l_p(\boldsymbol{\theta}) = -\frac{1}{2}(\log|V(\boldsymbol{\theta})| + (Z - X\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}))'V(\boldsymbol{\theta})^{-1}(Z - X\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta}))) \tag{4.35}$$

The maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ can now be obtained by maximizing $l_p(\boldsymbol{\theta})$ with regard to $\boldsymbol{\theta}$. As one can see above (4.36) the estimation $\tilde{\boldsymbol{\beta}}(\boldsymbol{\theta})$ is just replaced in the log-likelihood function $l(\boldsymbol{\beta}, \boldsymbol{\theta})$ without taking into account the degree of freedom lost through the estimation of $\boldsymbol{\beta}$. This produces a bias in the estimation of $\boldsymbol{\theta}$. Therefore the restricted maximum likelihood estimator (RMLE) is commonly used for parameter estimation in this case, since it less biased than the MLE.

The RMLE consist in determining the marginal log-likelihood $l_R(\theta)$ by integrating the likelihood function $L(\boldsymbol{\beta}, \boldsymbol{\theta})$ over $\boldsymbol{\beta}$, that is, $l_R(\theta) = \log(\int L(\boldsymbol{\beta}, \boldsymbol{\theta})d\boldsymbol{\beta})$. It results that

$$l_R(\boldsymbol{\theta}) = l_p(\boldsymbol{\theta}) - \frac{1}{2}\log|X'V(\boldsymbol{\theta})^{-1}X|. \tag{4.36}$$

The restricted maximum likelihood estimator (RMLE) of $\boldsymbol{\theta}$ can now be obtained by maximizing $l_R(\boldsymbol{\theta})$ with regard to $\boldsymbol{\theta}$. Maximization of $l_R(\theta)$ can be perform using i.e. EM-algorithm (Fahrmeir and Tutz [23] p. 442-443).

Refer to Fahrmeir and Tutz [23] p. 289-292 for further details about parameter estimation for mixed model when variances-covariances components are known/unknown.

4.4.3.2 Double Exponential Smoothing

Double exponential smoothing consist of predicting future values of univariate time series based on its past values. One observes the time series of systematic risk $\{s_t\}_{t=1}^T$ as the sum of a smoothed time series $\{\mu_t\}_{t=1}^T$ and random error ζ_t. The smoothed time series $\{\mu_t\}_{t=1}^T$ is function of its past value, the past value related to $\{s_t\}_{t=1}^T$, the trend and the seasonal component. For this study seasonal component is not taken into account, since the data have been yearly collected and the shortness of the time series do not allow the detection of a cyclical effect. We solely include an additive trend component b_t. We can therefore write the model in state space form as follow:

$$s_t = \mu_t + \zeta_t \tag{4.37}$$

$$\mu_t = l_{t-1} + b_{t-1} \tag{4.38}$$

$$l_t = l_{t-1} + b_{t-1} + \alpha\zeta_t \tag{4.39}$$

$$b_t = b_{t-1} + \alpha\beta\zeta_t. \tag{4.40}$$

As one can see in equation (4.38) the smoothed value μ_t at time t is predicted by the sum of the past level l_{t-1} and trend b_{t-1}. One can analogously predict the future smoothed value F_{T+1} at time $T+1$ by summing level and trend at time T, that is,

$$F_{T+1} = l_T + b_T. \tag{4.41}$$

The actual level l_t is predicted as the sum of the previous level l_{t-1} and previous trend b_{t-1} adjusted by the size of the random error ζ_t. The coefficient α arbitrate the influence of the random error ζ_t on the prediction of the level l_t. The actual trend b_t is predicted by the previously estimated trend b_{t-1} and the size of the random error ζ_t. A combination of the coefficients α and β arbitrate the influence of the random error ζ_t on the prediction of the trend b_t.

The task is then to determine optimal values of coefficients α and β. This can be done by defining a loss function and minimizing it with respect to α and β. A possible loss function is MSE $= \sum_{i=1}^T \zeta_t^2$. Optimization can then be solved i.e. numerically. Refer to Hyndman et al. [24] for more details about automatic forecasting using exponential smoothing methods.

4.5 Assessing Goodness of Fit

In this section we are presenting methods to assess goodness of fit under two approaches. The first approach consists of checking how the model adapts to the training sample. Estimating goodness of fit solely on a training sample involves the danger of overfitting, which may result in a low accuracy of the model in reality. We therefore include another approach, which estimates the goodness of fit on data, which has not been involved in model parametrization.

4.5.1 In-Sample

A very popular indicator of in-sample goodness of fit is the R^2, because it gives a percentage of the proportion of variation of the depending variable explained by the model.

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}(Z_i - \hat{Z}_i)^2}{\sum_{i=1}^{n}(Z_i - \bar{Z})^2}. \tag{4.42}$$

Because of its tendency to increase with growing number of covariates, a correction has been brought by creating an adjusted R^2.

$$\mathrm{R}^2_{adj} = 1 - \frac{\frac{1}{n-p-1}\sum_{i=1}^{n}(\hat{Z}_i - \bar{Z})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z})^2}. \tag{4.43}$$

An increase of covariates tends to decrease the bias of the model, but at the same time increases it variance. An increase of the accuracy of the model is achieved when the biased decreases more than the variance increase. A group of methods have been develop to perform this tradeoff while assessing the goodness of fit. The most popular one are the Akaike Information Criteria and the Bayesian Information Criteria (BIC).

$$\mathrm{AIC} = n\log(\hat{\sigma}^2_Z) + 2k \tag{4.44}$$

$$\mathrm{BIC} = n\log(\hat{\sigma}^2_Z) + k\log(n) \tag{4.45}$$

In Hastie et al. [20] p. 193-213 further explanation is given on assessing goodness of fit on a training sample.

4.5.2 Out-Of-Sample

The most suitable method to validate models on large datasets consists of randomly dividing the data into training and test sample. The usual proportion is 80% for training and 20% for test samples. One then trains the model on the training sample and estimate it out-of-sample goodness of fit on the test sample. Since our dataset is relatively small, this approach cannot be applied.

We therefore choose to apply a 10-fold cross-validation. The 10-fold cross-validation consists of randomly dividing the sample in 10 equal sub-samples. Thereafter one trains the model on the training samples consisting of 9 sub-samples and test it on the remaining sub-sample. This is done 10 times, so that each sub-sample is once the test sample. The out-of-sample mean square error is then the average of the squared error on all test samples. Refer to Hastie et al. [20] p. 214-221 for further details about out-of-sample assessing of goodness of fit.

4.6 Conclusion

Retail credit portfolios are characterized by large numbers of observations. Methodical approach to develop rating systems for such portfolios consists of a two level model. In the first level probability of default for each credit applicant is determined by performing cross-sectional analysis with methods such as logistic, logit or probit regression. In the second step the longitudinal aspect is taken into account by categorizing the probability of default into rating-grades, while ensuring that the probability of default in the portfolio is matched with the historical estimated probability of default prediction. The latter is usually based on a simple weighted sum of historical default rates. (see Engelmann and Rauhmeier [4] p. 25-36)

Contrary to retail credit portfolios, non-retail credit portfolios are characterized by small numbers of observations and often short time series. The common practice to estimate probability of default in such portfolios consists of performing a cross-sectional analysis on the whole time series, while ignoring time dependencies in the data. Systematic risk is taken into account by adding macroeconomic variables on the same level as credit applicants specific variables. This leads to a loss in goodness of fit. (see Engelmann and Rauhmeier [4] p. 13-24 and 37-74)

Our scientific contribution lies in proposing a methodology to suitably take time dependencies into account for short panel data of scarce observations, while determining systematic and idiosyncratic risk simultaneously in the context of the development of rating systems for non-retail credit portfolios.

Chapter 5

Rating Model based on Cross-Sectional Analysis

After preparing the data as explained in chapter 3 and presenting the set of methods that we have selected for this work in chapter 4, we present the first part of our analysis results in this chapter. It is based on a common practices in banks and financial institution, which consist in ignoring the time dependency of the data and applies a linear regression analysis as described in 4.3. This approach has the advantage to bypass the limitation of the small amount of observation at each point of time taken individually. We therefore rearrange the data and describe it in section 5.1.

We start with a simple model approach based on predictors with a minimum of 90% filling rate, which we gradually alter by adding further variables. That is, variables with less filling rates or lagged values.

For each of the model alternatives, we start by analyzing and handling missing values in the first subsection. In the following subsection we analyze pairwise correlations of predictors and find suitable handling of highly correlated predictors. In the third section we present and comment the estimated model parameters.

The analysis have been carried out in R Core Team [25]. In addition to package base, packages MASS (Venables and Ripley [26]), gdata (Warnes et al. [27]) and DAAG (Maindonald and Braun [28]) have been used.

5.1 Data Description

Based on the fact that the ratings of year t are based on financial information of year $t-1$, we join the rating table of year t with the table of financial information of year $t-1$. It results in a table with 414 observations of 36 security issuers from 1997 to 2013. We further assumed, that observations with less than 4 non-missing variables out of a total of 64 variables, are useless. We removed them from the data and obtain a table with 277 observations of 35 security issuers from 1997 to 2013.

For the analysis of the relevance of variables with lagged values we have followed the same principle as above and created tables by joining ratings of year t with financial information of year $t-2$, $t-3$, $t-4$ and $t-5$. See table 5.1 for an overview of the resulting data.

Type	Obs	Issuers	From	To
Lag-1	277	35	1997	2013
Lag-2	255	35	1998	2013
Lag-3	231	33	1999	2013
Lag-4	207	32	2002	2013
Lag-5	181	31	2003	2013

Table 5.1: Data description of tables with lagged values

5.2 Regression Analysis based on Predictors with actual values

5.2.1 Missing Value Analysis and Handling

Our model approach is based on regression analysis, that is, estimating linear dependencies between predictors (financial ratios) and the depending variable, which is in our case the logit of the underlying probability of defaults of the rating-grades. As mentioned in chapter 4, estimating the regression parameters requires an inversion of the design matrix, for which each line contains a set of predictors values for each securities' issuer. Any missing value in the matrix would lead to make related calculus impossible. We therefore decide to keep only predictors with sufficient amount of non-missing values.

After observing an overview of the percentage of missing value of financial ratios (see table 5.2 and 5.3), we decide to choose a threshold of 10%. Financial ratios with less than 90% filling rate are excluded from the set of predictors.

Nr	Variables	Missing percent
1	Sales...1.Yr.Growth_lag1	7.58
4	Assets...1.Yr.Growth_lag1	6.50
5	Net.Worth...1.Yr.Growth_lag1	7.94
6	Capital...1.Yr.Growth_lag1	7.58
9	Asset.Turnover_lag1	7.22
10	Operating.Margin_lag1	3.61
11	Pretax.Margin_lag1	1.08
12	Profit.Margin_lag1	2.17
13	Return.on.Assets_lag1	6.50
14	Return.on.Common.Equity_lag1	7.94
16	Financial.Leverage_lag1	7.22
17	Annualized.Return.on.Common.Equity_lag1	7.94
18	Operating.ROE_lag1	8.30
21	Number.of.Employees_lag1	5.05
23	Net.Income.Per.1000.Employees_lag1	5.05
24	Actual.Sales.Per.Employee_lag1	5.05
25	Assets.Per.1000.Employees_lag1	5.05
26	Total.Debt.Total.Capital_lag1	1.08
29	Assets.Equity_lag1	0.36
30	Tangible.Common.Equity.Ratio_lag1	9.03
31	LT.Debt.Common.Equity_lag1	1.44
32	LT.Debt.Total.Capital_lag1	1.08
33	LT.Debt.Total.Assets_lag1	1.08
34	Total.Debt.Common.Equity_lag1	0.72
35	Total.Debt.Tangible.Book.Value_lag1	9.03
36	Total.Debt.Total.Equity_lag1	1.44
37	Total.Debt.Total.Assets_lag1	0.00
39	Net.Debt.Shareholders.Equity_lag1	1.44
45	Sales.Cash_lag1	1.08
46	Sales.Marketable.Securities_lag1	1.44
47	Sales.Net.Fixed.Assets_lag1	1.08
48	Sales.LT.Investments_lag1	3.25
49	Sales.Other.Assets_lag1	1.08
52	Tangible.Common.Equity_lag1	3.61
54	Tot.loan.tot.dep_lag1	14.08

TABLE 5.2: Variables with less than 10% missing value

Nr	Variables	Missing percent
2	EBIT...1.Yr.Growth_lag1	23.47
3	Net.Income...1.Yr.Growth_lag1	23.83
7	Dividend.Payout.Ratio_lag1	48.38
8	Net.Fixed.Asset.Turnover_lag1	11.19
15	Return.on.Capital_lag1	13.00
19	X5.Year.Average.Return.On.Equity_lag1	100.00
20	Return.on.Equity...5.Yr.Geometric.Growth_lag1	62.45
22	Employees...1.Yr.Growth_lag1	13.00
27	CFO.Total.Debt_lag1	26.35
28	Cashflow.Total.Liabilities_lag1	26.35
38	Total.Debt.Market.Cap_lag1	62.82
40	T12M.CFO.Total.Debt_lag1	26.35
41	T12M.FCF.Total.Debt_lag1	33.57
42	Net.Change.in.Liabilities...of.Total_lag1	11.91
43	Increase.In.Equity...of.Total_lag1	11.91
44	Asset.Utilization_lag1	100.00
50	Price.T12M.Earnings.per.Share_lag1	64.62
51	Book.Value.per.Share_lag1	51.26
53	Tangible.Common.Equity.to.Risk.Weighted.Assets_lag1	65.34
55	Equity.Ratio_lag1	100.00
56	Tier.1.Common.Equity.Ratio_lag1	99.64
57	Core.Tier.1.Capital.Ratio_lag1	88.09
58	Tier.1.Risk.Based.Capital.Ratio_lag1	36.82
59	Total.Risk.Based.Capital.Ratio_lag1	42.96
60	Risk.Weighted.Assets_lag1	65.34
61	Ratio.of.Risk.weighted.assets.to.Total.assets..RWA.TA._lag1	65.34
62	Changes.in.indicator..5..measured.in.percentage.points..rwa.ta._lag1	74.01
63	int.inc....int...non.int.inc._lag1	10.47
64	5.Year.Average.Return.On.Equity_lag1	38.27

TABLE 5.3: Variables with more than 10% missing value

5.2.2 Pairwise Correlation Analysis and Handling

Before running the linear regression analysis, another technical restriction that should be dealt with is multicollinearity. That is when several predictors are highly or totally correlated. One distinguishes between extreme multicollinearity for totally correlated predictors and weak multicollinearity for highly correlated predictors (see section 4.2).

In table 5.4 we present an overview of highly correlated predictors, that is, predictors for which the Pearson's correlation coefficient is higher than 0.70.

Variable 1	Variable 2	Corr
Capital...1.Yr.Growth_lag1	Assets...1.Yr.Growth_lag1	0,88
Profit.Margin_lag1	Operating.Margin_lag1	0,82
Pretax.Margin_lag1	Return.on.Assets_lag1	0,91
Pretax.Margin_lag1	Return.on.Common.Equity_lag1	0,79
Pretax.Margin_lag1	Annualized.Return.on.Common.Equity_lag1	0,79
Pretax.Margin_lag1	Operating.ROE_lag1	0,79
Pretax.Margin_lag1	Net.Income.Per.1000.Employees_lag1	0,72
Return.on.Assets_lag1	Return.on.Common.Equity_lag1	0,88
Return.on.Assets_lag1	Annualized.Return.on.Common.Equity_lag1	0,88
Return.on.Assets_lag1	Operating.ROE_lag1	0,85
Return.on.Common.Equity_lag1	Annualized.Return.on.Common.Equity_lag1	1,00
Return.on.Common.Equity_lag1	Operating.ROE_lag1	0,97
Annualized.Return.on.Common.Equity_lag1	Operating.ROE_lag1	0,97
Actual.Sales.Per.Employee_lag1	Assets.Per.1000.Employees_lag1	0,94
Number.of.Employees_lag1	Tangible.Common.Equity_lag1	0,80
Total.Debt.Total.Capital_lag1	Total.Debt.Total.Assets_lag1	0,80
Assets.Equity_lag1	LT.Debt.Common.Equity_lag1	0,89
Assets.Equity_lag1	Total.Debt.Common.Equity_lag1	0,92
Assets.Equity_lag1	Total.Debt.Tangible.Book.Value_lag1	0,94
Assets.Equity_lag1	Total.Debt.Total.Equity_lag1	0,93
Assets.Equity_lag1	Net.Debt.Shareholders.Equity_lag1	0,89
LT.Debt.Common.Equity_lag1	Total.Debt.Common.Equity_lag1	0,95
LT.Debt.Common.Equity_lag1	Total.Debt.Tangible.Book.Value_lag1	0,91
LT.Debt.Common.Equity_lag1	Total.Debt.Total.Equity_lag1	0,95
LT.Debt.Common.Equity_lag1	Net.Debt.Shareholders.Equity_lag1	0,94
LT.Debt.Total.Capital_lag1	LT.Debt.Total.Assets_lag1	0,79
LT.Debt.Total.Assets_lag1	Total.Debt.Total.Assets_lag1	0,79
Total.Debt.Common.Equity_lag1	Total.Debt.Tangible.Book.Value_lag1	0,97
Total.Debt.Common.Equity_lag1	Total.Debt.Total.Equity_lag1	0,99
Total.Debt.Common.Equity_lag1	Net.Debt.Shareholders.Equity_lag1	0,97
Total.Debt.Tangible.Book.Value_lag1	Total.Debt.Total.Equity_lag1	0,96
Total.Debt.Tangible.Book.Value_lag1	Net.Debt.Shareholders.Equity_lag1	0,95
Total.Debt.Total.Equity_lag1	Net.Debt.Shareholders.Equity_lag1	0,98

Table 5.4: Variables with pairewise correlation more than 70%

5.2.3 Parameter Estimation

As can be seen in table 5.4, a large number of relevant variables are pairwise highly correlated. Two approaches have been implemented to handle correlated predictors while estimating model parameters. We first remove some variables from the predictor set, so that the remaining variables are no more highly correlated. In the second approach we apply principal component analysis to generate uncorrelated components.

All three approaches are done for different percentage of missing value, namely less than 10%, 15% and 25%. This allows to gradually take more variables into account and assess their effect on the model accuracy. For each of the percentage we implement three step model scheme. In the first step we perform a stepwise regression to select relevant variables for the model. In the second step we re-adapt the dataset by only deleting observation from the original dataset, which contain missing in the relevant variables as selected in the stepwise regression of step 1. We perform again a stepwise regression on the enriched dataset. In the third step we perform a residual analysis based on the regression results of step 2 and delete outliers, that is, observations with extreme residuals outside the range $[-2.50, 2.50]$. Thereafter we execute stepwise regression on the shrank dataset.

For the principal component analysis the step 2 will not be executed, since the resulting principal components are rotations based on all variables. So its step 2 correspond to the step 3 of the former.

5.2.3.1 Regression with Correlated Predictors Excluded

After analyzing the pairwise correlation in 5.4, we choose to exclude "Capital...1.Yr.Growth", "Operating.Margin", "Pretax.Margin", "Return.on.Common.Equity", "Annualized.Return.on.Common.Equity", "Operating.ROE", "Actual.Sales.Per.Employee", "LT.Debt.Common.Equity", "Total.Debt.Common.Equity", "Total.Debt.Tangible.Book.Value", "Total.Debt.Total.Equity", "Net.Debt.Shareholders.Equity", "Total.Debt.Total.Assets", "Tangible.Common.Equity" from the predictor set.

We start the analysis by removing all variables with more than 10% missing value from the dataset and delete all observations containing at least one missing value. We obtain a dataset containing 214 observations.

Step 1 results (see table 5.5) in a significant regression model (F-statistic: 4.79, p-value: 2.02e-05) with low goodness (R-squared: 0.158, adjusted R-squared: 0.125). The very skewed distribution of the residuals reflect the dependency between residuals and estimate logit of PD (see figure 5.5). This is an indication of miss-specification of the model.

Variables	Coef	P-value	Signif. level
(Intercept)	$6.77e+00$	0.0998	.
Financial.Leverage_lag1	$-2.24e-02$	0.0547	.
Number.of.Employees_lag1	$9.87e-06$	0.0990	.
Net.Income.Per.1000.Employees_lag1	$1.40e-03$	0.0211	*
Assets.Per.1000.Employees_lag1	$6.45e-06$	0.0993	.
Total.Debt.Total.Capital_lag1	$-1.28e-01$	0.0034	**
Tangible.Common.Equity.Ratio_lag1	$-3.60e-01$	0.0014	**
LT.Debt.Total.Assets_lag1	$3.18e-02$	0.0022	**
Sales.Net.Fixed.Assets_lag1	$-1.88e-03$	0.0024	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.66 on 205 degrees of freedom

Multiple R-squared: 0.158, Adjusted R-squared: 0.125

F-statistic: 4.79 on 8 and 205 DF, p-value: 2.02e-05

10-fold-CV: 2.82, Number of Observations: 214

Table 5.5: 1st step parameter estimation for variables with less than 10% missing values

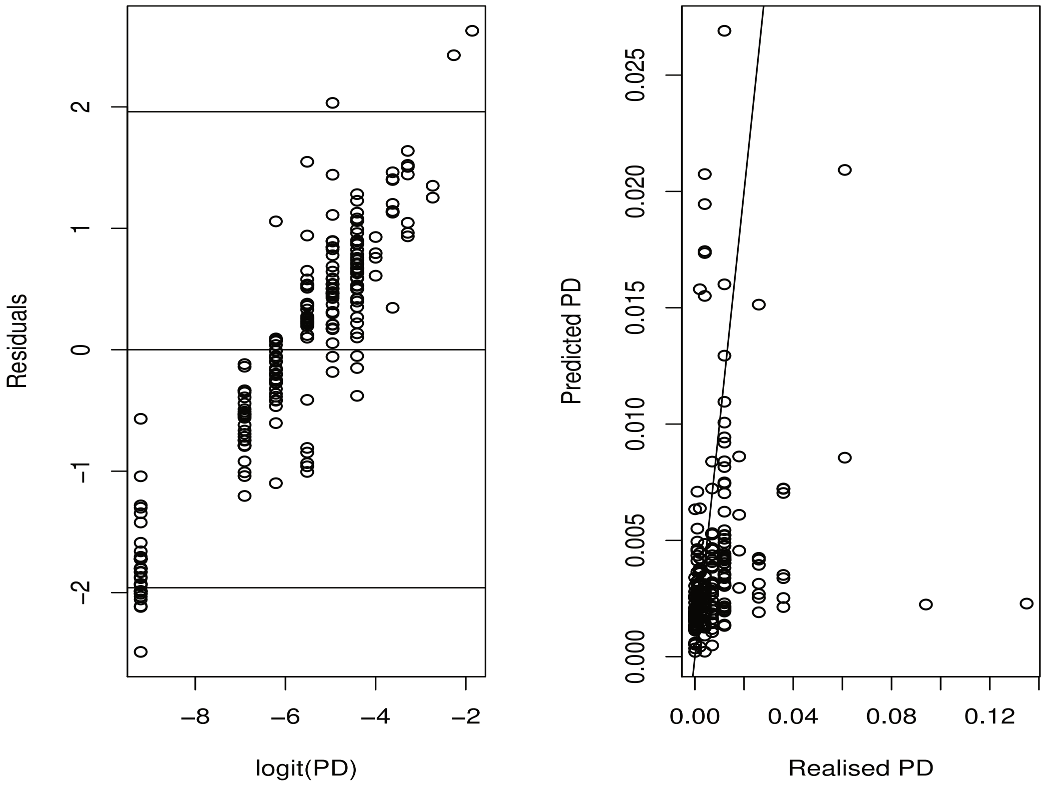


FIGURE 5.1: Residuals and predictions plots for 1st step parameter estimation for variables with less than 10% missing values

After enriching the dataset in step 2, the goodness of the regression model dropped to 0.08 (see table 5.6). This is coupled with the enhancement of the skewness of the residuals distribution in figure 5.2.

Variables	Coef	P-value	Signif. level
(Intercept)	6.367457	0.07926	.
Total.Debt.Total.Capital_lag1	−0.131464	0.00091	***
Tangible.Common.Equity.Ratio_lag1	−0.233208	0.00403	**
LT.Debt.Total.Assets_lag1	0.028933	0.00387	**
Sales.Net.Fixed.Assets_lag1	−0.001552	0.00729	**

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.7 on 220 degrees of freedom

Multiple R-squared: 0.083, Adjusted R-squared: 0.0663

F-statistic: 4.98 on 4 and 220 DF, p-value: 0.000738

10-fold-CV: 2.98, Number of Observations: 225

Table 5.6: 2nd step parameter estimation for variables with less than 10% missing values

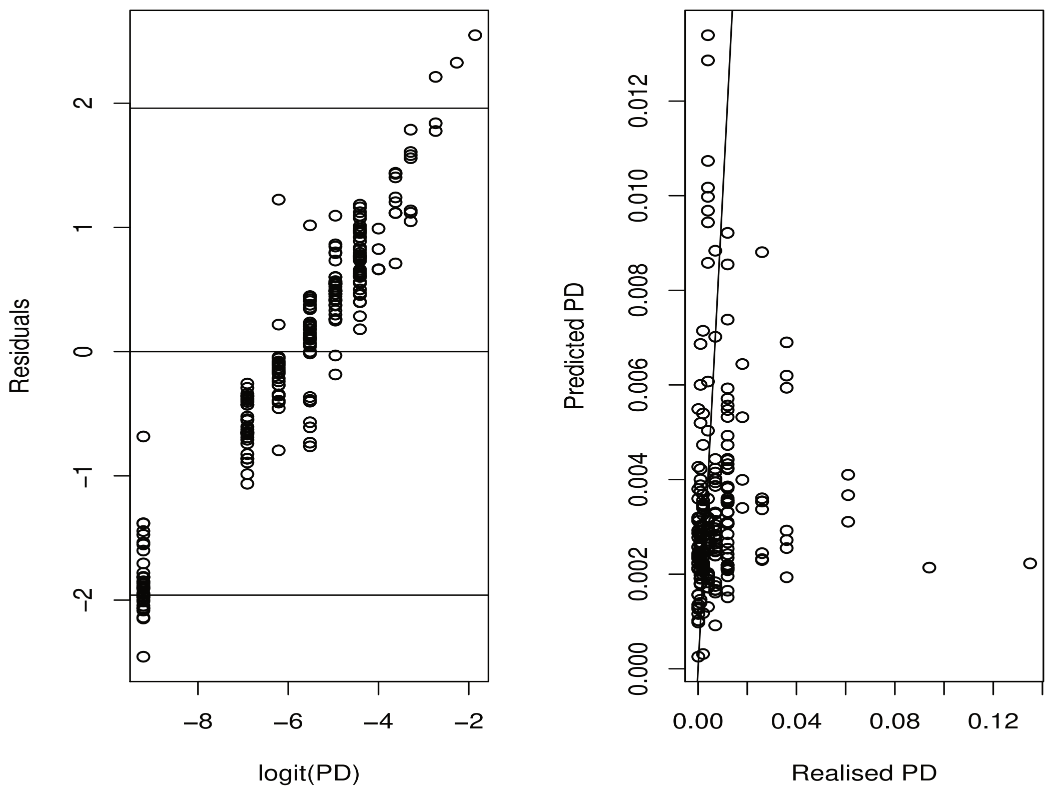


Figure 5.2: Residuals and predictions plots for 2nd step parameter estimation for variables with less than 10% missing values

After removing extreme value in step 3, the goodness of the regression model increase to 0.26 (see table 5.7), but the residuals distribution in figure 5.3 remain very skewed.

Variables	Coef	P-value	Signif. level
(Intercept)	$8.26e-01$	0.7194	
Net.Income.Per.1000.Employees_lag1	$-9.31e-04$	0.0577	.
Assets.Per.1000.Employees_lag1	$-5.34e-06$	0.0078	**
Total.Debt.Total.Capital_lag1	$-6.54e-02$	0.0095	**
Tangible.Common.Equity.Ratio_lag1	$-2.26e-01$	$2.0e-05$	***
LT.Debt.Total.Assets_lag1	$2.85e-02$	$3.4e-05$	***
Sales.Net.Fixed.Assets_lag1	$-1.68e-03$	$4.0e-05$	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.04 on 181 degrees of freedom

Multiple R-squared: 0.267, Adjusted R-squared: 0.242

F-statistic: 11 on 6 and 181 DF, p-value: 2.09e-10

10-fold-CV: 1.19, Number of Observations: 188

Table 5.7: 3rd step parameter estimation for variables with less than 10% missing values

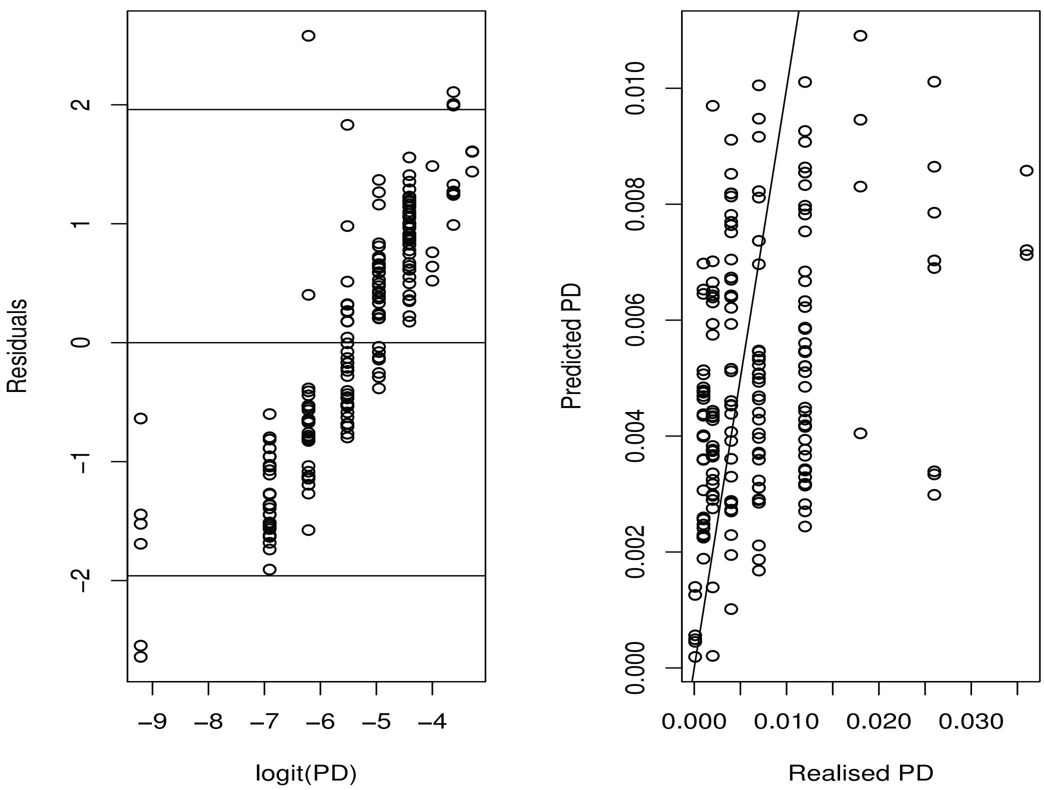


FIGURE 5.3: Residuals and predictions plots for 3rd step parameter estimation for variables with less than 10% missing values

Because of the low goodness of fit and very skewed residual plots, we decide to take more variables into account in the model. We include variables with percentage of missing values between 10% and 15% in the model. Step 1 result in a regression model with almost double goodness compared to the step 1 of the regression model based on variables with less than 10% missing values (see table 5.8). Similarly the shape of the residuals distribution in figure 5.4 shows less skewness.

Variables	Coef	P-value	Signif. level
(Intercept)	$1.90e+00$	0.70849	
Asset.Turnover_lag1	$-1.24e+01$	0.10133	
Return.on.Assets_lag1	$-5.10e-01$	0.16216	
Number.of.Employees_lag1	$2.77e-05$	$9.2e-05$	***
Assets.Per.1000.Employees_lag1	$1.85e-05$	0.00016	***
Total.Debt.Total.Capital_lag1	$-7.79e-02$	0.12214	
Tangible.Common.Equity.Ratio_lag1	$3.60e-01$	0.00495	**
LT.Debt.Total.Capital_lag1	$-2.20e-02$	0.10613	
LT.Debt.Total.Assets_lag1	$5.11e-02$	0.00570	**
Net.Change.in.Liabilities...of.Total_lag1	$2.12e-04$	0.01404	*
Sales.Cash_lag1	$2.96e-04$	0.02870	*
Sales.Net.Fixed.Assets_lag1	$-2.17e-03$	0.00177	**
Sales.LT.Investments_lag1	$-1.53e-04$	0.00601	**
Sales.Other.Assets_lag1	$2.40e-02$	0.13685	
int.inc....int...non.int.inc._lag1	$-2.54e+00$	0.04187	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.46 on 160 degrees of freedom

Multiple R-squared: 0.291, Adjusted R-squared: 0.229

F-statistic: 4.69 on 14 and 160 DF, p-value: 3.89e-07

10-fold-CV: 2.57, Number of Observations: 175

Table 5.8: 1st step parameter estimation for variables with less than 15% missing values

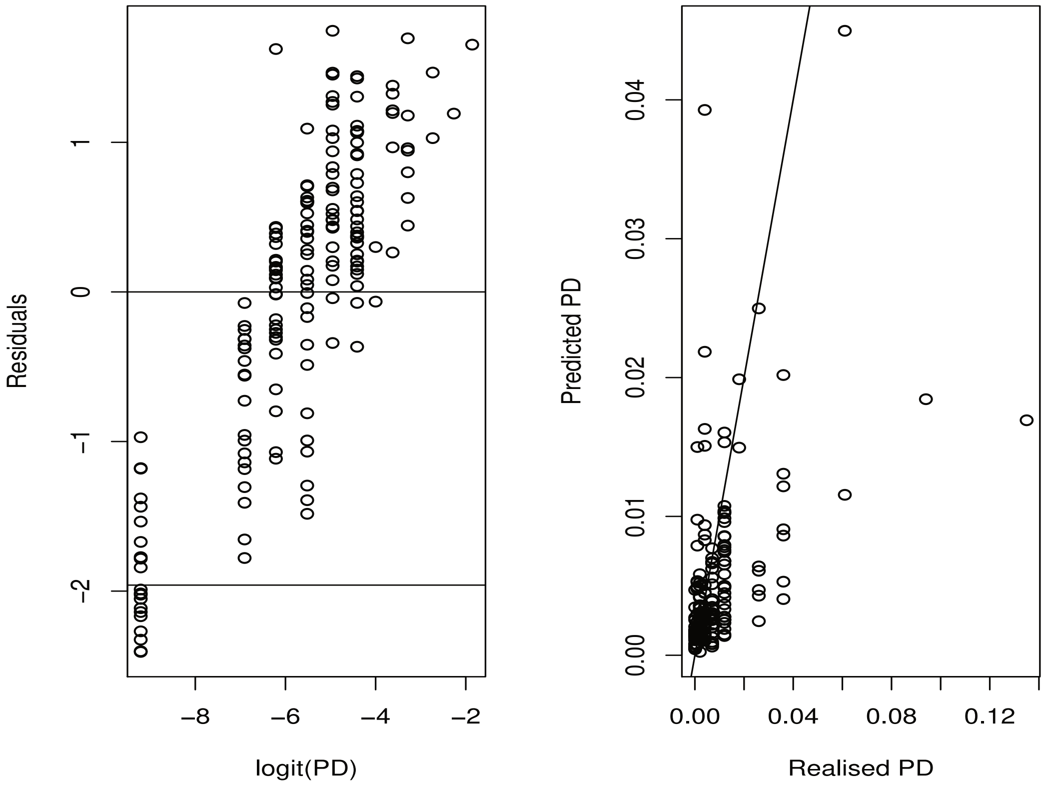


FIGURE 5.4: Residuals and predictions plots for 1st step parameter estimation for variables with less than 15% missing values

Step 2 gives a similar picture like step 1 with a drop of the goodness of fit from 0.291 to 0.219 (see table 5.9 and figure 5.5).

Variables	Coef	P-value	Signif. level
(Intercept_)	$-7.21e+00$	$< 2e-16$	***
Asset.Turnover_lag1	$-1.44e+01$	0.02296	*
Number.of.Employees_lag1	$3.04e-05$	$1e-06$	***
Assets.Per.1000.Employees_lag1	$1.43e-05$	0.00013	***
Tangible.Common.Equity.Ratio_lag1	$3.11e-01$	0.00032	***
LT.Debt.Total.Assets_lag1	$1.56e-02$	0.07767	.
Net.Change.in.Liabilities...of.Total_lag1	$1.87e-04$	0.02852	*
Sales.Cash_lag1	$2.13e-04$	0.07462	.
Sales.Net.Fixed.Assets_lag1	$-1.84e-03$	0.00531	**
Sales.LT.Investments_lag1	$-1.13e-04$	0.00588	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.5 on 187 degrees of freedom

Multiple R-squared: 0.219, Adjusted R-squared: 0.181

F-statistic: 5.81 on 9 and 187 DF, p-value: 4e-07

10-fold-CV: 2.43, Number of Observations: 197

TABLE 5.9: 2nd step parameter estimation for variables with less than 15% missing values

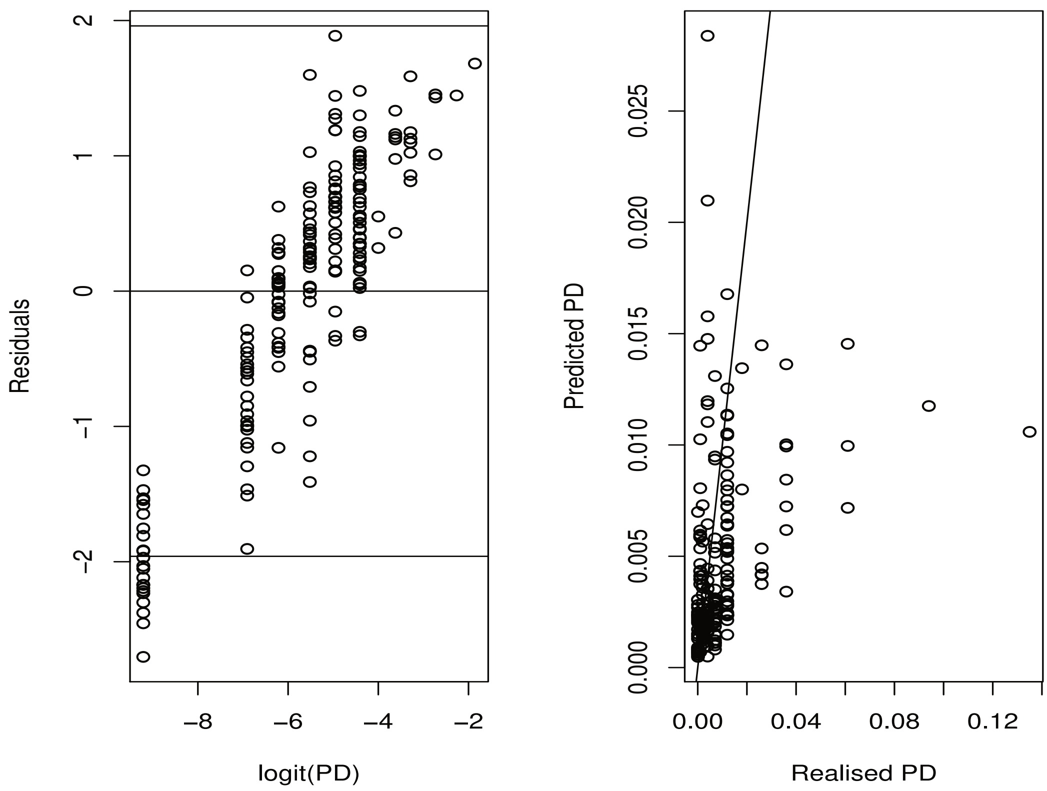


FIGURE 5.5: Residuals and predictions plots for 2nd step parameter estimation for variables with less than 15% missing values

Removing outliers from the model of step 2, one obtain a model with a higher goodness of fit of 0.364 (compared to 0.219). The mean squared error of the 10-fold cross validation MSE also considerably dropped from 2.43 to 1.26. However the residuals plot still show a very skewed distribution (see table 5.10 and figure 5.6). We therefore decide to take more variables in the model and extend the threshold of the missing rate to 25%.

Variables	Coef	P-value	Signif. level
(Intercept_)	$-6.26e+00$	$< 2e-16$	***
Asset.Turnover_lag1	$-2.22e+01$	$8.2e-06$	***
Return.on.Assets_lag1	$-5.03e-01$	0.04322	*
Number.of.Employees_lag1	$2.99e-05$	$2.3e-10$	***
Assets.Per.1000.Employees_lag1	$7.23e-06$	0.00856	**
Tangible.Common.Equity.Ratio_lag1	$2.58e-01$	0.00017	***
LT.Debt.Total.Capital_lag1	$-1.26e-02$	0.11779	
LT.Debt.Total.Assets_lag1	$4.61e-02$	$2.8e-06$	***
Net.Change.in.Liabilities...of.Total_lag1	$1.91e-04$	0.00222	**
Sales.Net.Fixed.Assets_lag1	$-1.85e-03$	0.00029	***
Sales.LT.Investments_lag1	$-5.62e-05$	0.02019	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.08 on 167 degrees of freedom

Multiple R-squared: 0.364, Adjusted R-squared: 0.326

F-statistic: 9.57 on 10 and 167 DF, p-value: 1.64e-12

10-fold-CV: 1.26, Number of Observations: 178

TABLE 5.10: 3rd step parameter estimation for variables with less than 15% missing values

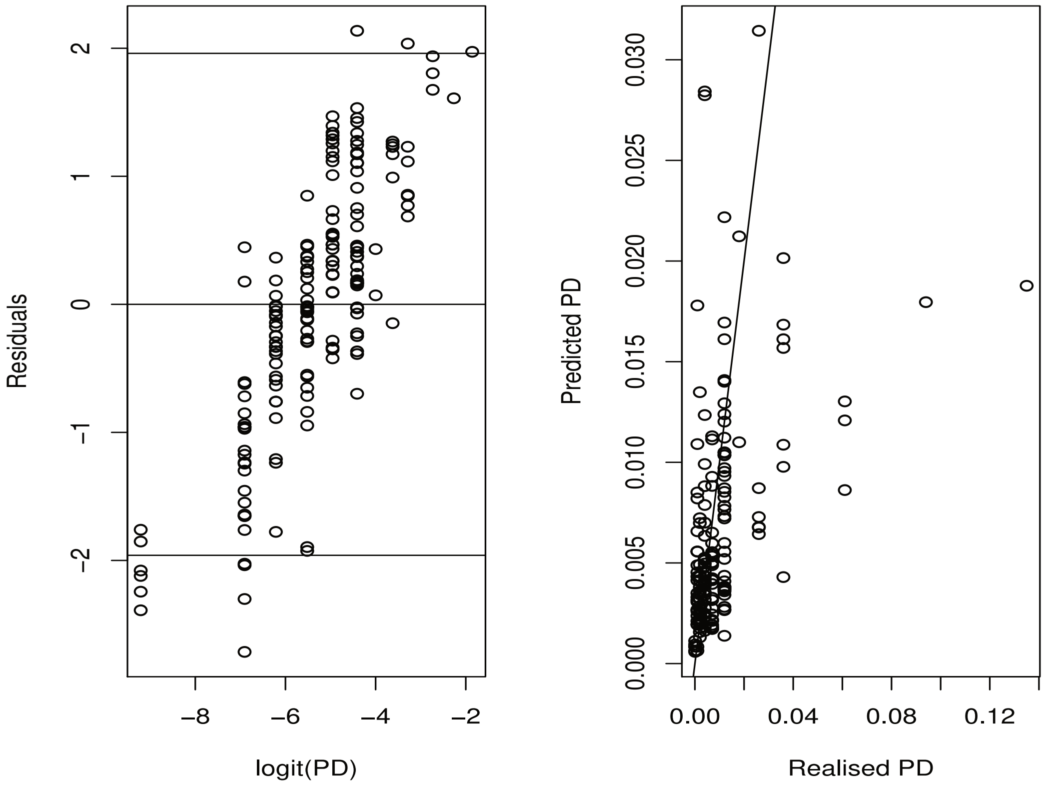


FIGURE 5.6: Residuals and predictions plots for 3rd step parameter estimation for variables with less than 15% missing values

Including more variables leads to an increase of the goodness of fit in step 1 (0.380 vs. 0.291). But the residuals distribution (figure 5.7) remain skewed.

Variables	Coef	P-value	Signif. level
(Intercept)	$2.80e+00$	0.51346	
Net.Income...1.Yr.Growth_lag1	$1.41e-03$	0.12451	
Assets...1.Yr.Growth_lag1	$2.76e-02$	0.00121	**
Asset.Turnover_lag1	$-1.88e+01$	0.01602	*
Profit.Margin_lag1	$-3.54e-02$	0.00080	***
Number.of.Employees_lag1	$2.34e-05$	0.00033	***
Employees...1.Yr.Growth_lag1	$-1.35e-02$	0.05502	.
Assets.Per.1000.Employees_lag1	$1.50e-05$	0.00050	***
Total.Debt.Total.Capital_lag1	$-8.18e-02$	0.10643	
Assets.Equity_lag1	$-5.74e-02$	$3.4e-05$	***
LT.Debt.Total.Assets_lag1	$3.99e-02$	0.00102	**
Sales.Cash_lag1	$2.76e-04$	0.04507	*
Sales.Net.Fixed.Assets_lag1	$-1.34e-03$	0.09175	.
Sales.Other.Assets_lag1	$4.41e-02$	0.02840	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.37 on 130 degrees of freedom

Multiple R-squared: 0.38, Adjusted R-squared: 0.318

F-statistic: 6.12 on 13 and 130 DF, p-value: 7.84e-09

10-fold-CV: 2.37, Number of Observations: 144

Table 5.11: 1st step parameter estimation for variables with less than 25% missing values

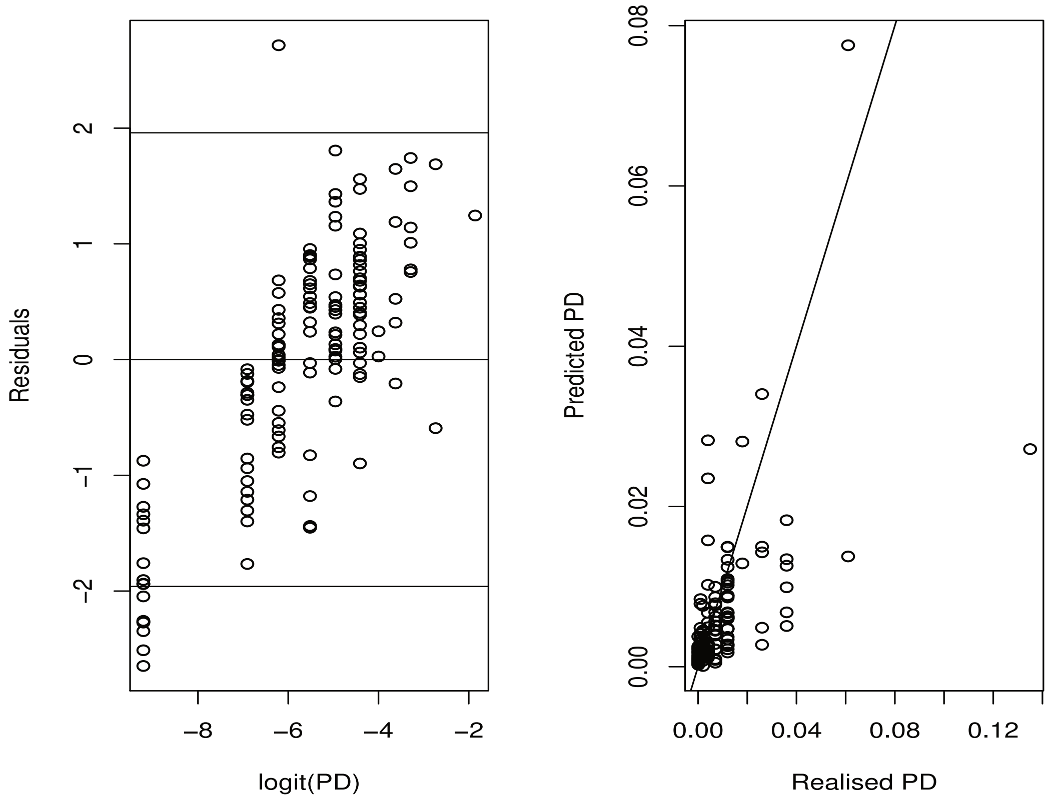


FIGURE 5.7: Residuals and predictions plots for 1st step parameter estimation for variables with less than 25% missing values

In step 2 the goodness of fit considerably dropped from 0.380 in step 1 to 0.141 in step 2.

Variables	Coef	P-value	Signif. level
(Intercept)	$-8.29e-01$	0.81875	
Number.of.Employees_lag1	$1.83e-05$	0.00300	**
Assets.Per.1000.Employees_lag1	$7.62e-06$	0.06277	.
Total.Debt.Total.Capital_lag1	$-6.39e-02$	0.13186	
Assets.Equity_lag1	$-2.04e-02$	0.08011	.
LT.Debt.Total.Assets_lag1	$3.93e-02$	0.00047	***
Sales.Net.Fixed.Assets_lag1	$-2.37e-03$	0.00104	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.67 on 187 degrees of freedom

Multiple R-squared: 0.141, Adjusted R-squared: 0.113

F-statistic: 5.1 on 6 and 187 DF, p-value: 7.11e-05

10-fold-CV: 3.02, Number of Observations: 194

Table 5.12: 2nd step parameter estimation for variables with less than 25% missing values

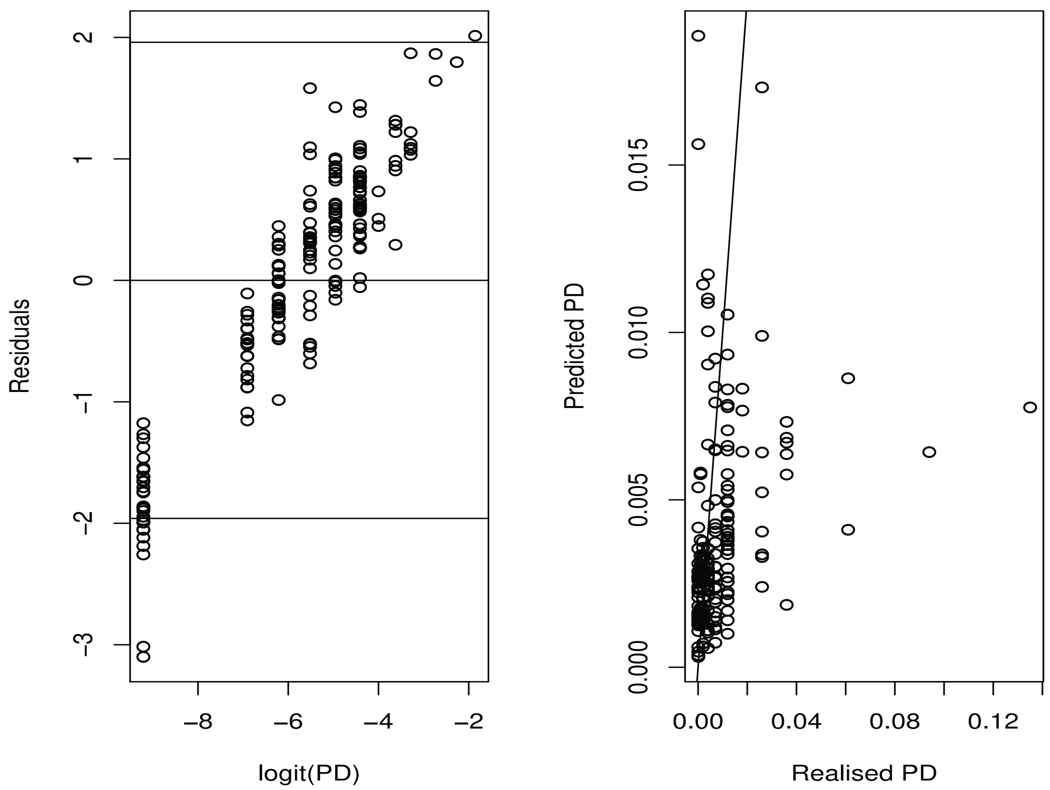


FIGURE 5.8: Residuals and predictions plots for 2nd step parameter estimation for variables with less than 25% missing values

Removing the outlier from the model' data in step 2 leads to a model with a higher goodness of fit of 0.355 in step 3. Compared to step 1 the model in step 3 has a lower goodness of fit (0.380 vs. 0.355), but on the other side it shows a considerably lower mean squared error on the 10-folds cross-validation MSE (1.12 vs. 2.37). However the residuals distribution (figure 5.9) remains skewed. We therefore decide to use an alternative approach to include more variables in the model.

Variables	Coef	P-value	Signif. level
(Intercept)	$-9.02e-01$	0.64822	
Net.Income...1.Yr.Growth_lag1	$1.69e-03$	0.00537	**
Profit.Margin_lag1	$-3.29e-02$	$2.5e-06$	***
Number.of.Employees_lag1	$1.27e-05$	0.00047	***
Total.Debt.Total.Capital_lag1	$-5.83e-02$	0.00988	**
LT.Debt.Total.Assets_lag1	$3.44e-02$	$9.2e-07$	***
Sales.Net.Fixed.Assets_lag1	$-2.55e-03$	$5.9e-09$	***
Sales.Other.Assets_lag1	$2.04e-02$	0.14246	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.989 on 156 degrees of freedom

Multiple R-squared: 0.355, Adjusted R-squared: 0.327

F-statistic: 12.3 on 7 and 156 DF, p-value: 1.8e-12

10-fold-CV: 1.12, Number of Observations: 164

Table 5.13: 3rd step parameter estimation for variables with less than 25% missing values

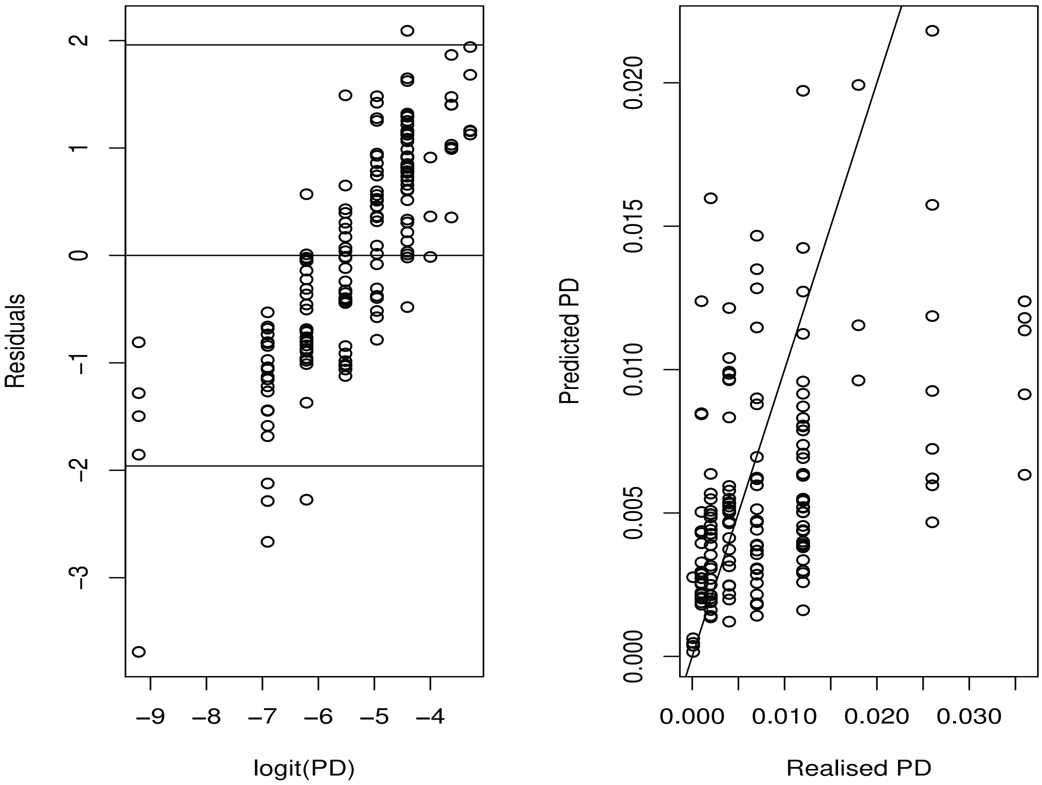


FIGURE 5.9: Residuals and predictions plots for 3rd step parameter estimation for variables with less than 25% missing values

5.2.3.2 Regression with Principal Components

In this section we proceed analog to section 5.2.3.1 with three steps. Starting by selecting variables with at most 10% missing rates, we progressively add at each following steps more variables by increasing the threshold of missing rates. For each step we start by determining principal components. Thereafter we perform a stepwise linear regression based on all principal components. We use all principal component, because our goal by performing the principal component analysis is not dimension reduction, but rather obtaining orthogonal predictors. The combination of forward and backward selection with AIC as information criteria selects the suitable principal components for the final model.

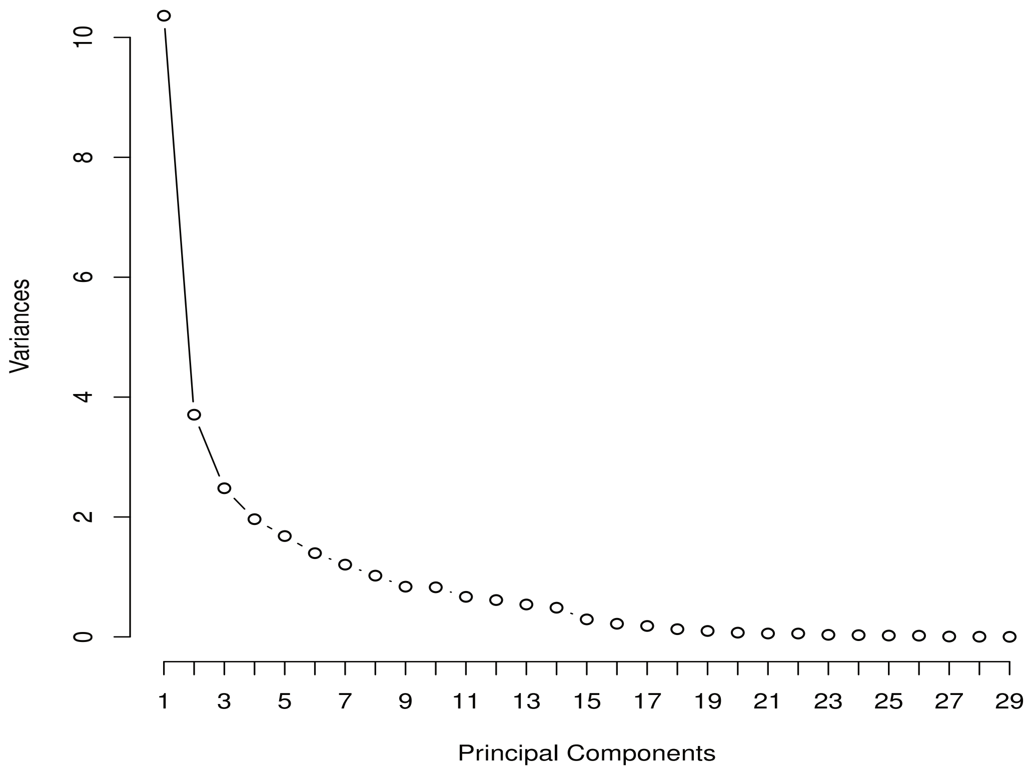


Figure 5.10: Scree plot of principal component for variables with less than 10% missing values

Figure 5.10 shows a scree plot of the principal component obtained. The scree plot represents in a graphic the proportion of variance explain by each principal component. One can see a break in figure 5.10 at the principal component 14 showing that 96% of the variance of the input-data is explained by the 14 first principal component. Principal component 1 and 2 explain 49% of the variance. Figure 5.11 shows a biplot of the two principal component 1 and 2. Biplots consist of representing both observations and variables in the same plot depending of the two first components with highest explained variances. Observations are labeled in the plot and variables are represented as vectors (red colored). The biplot shows that according to the two first components, one can summarize the effects of the variables in 7 main directions.

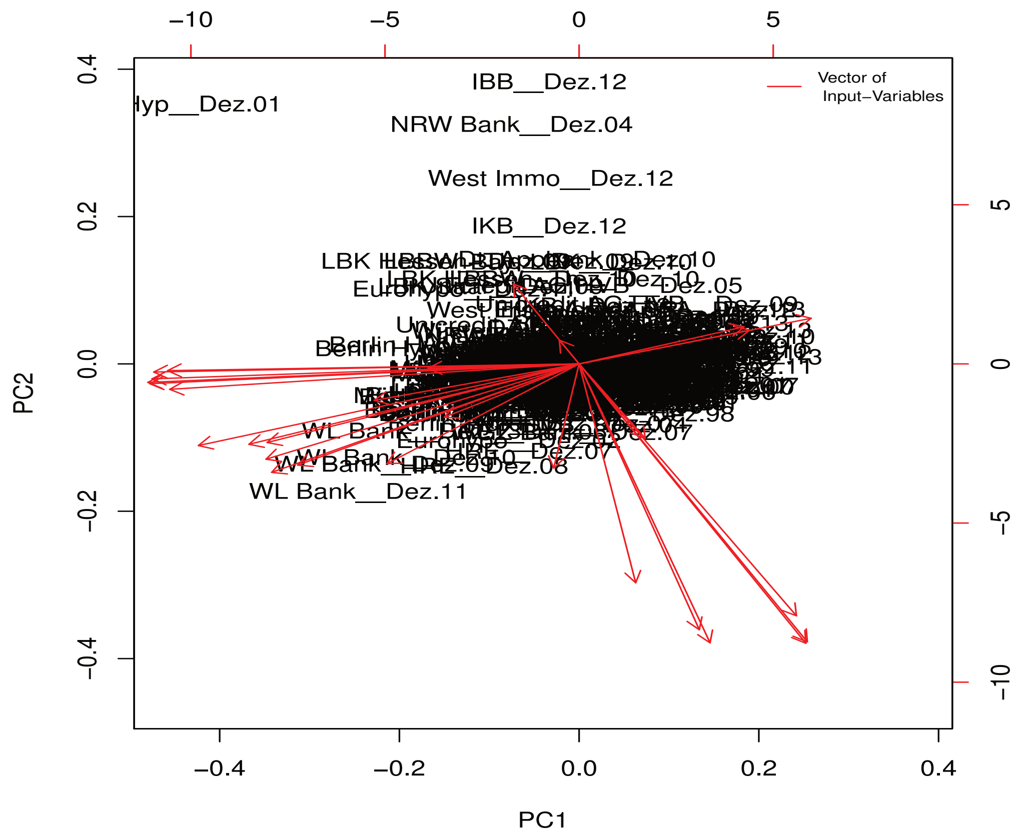


FIGURE 5.11: Biplot of principal component for variables with less than 10% missing values

Compared to step 1 for variables with less than 10% missing rates in section 5.2.3.1 the goodness of fit has considerably increased (0.374 vs. 0.158). The 10-fold cross validation MSE also dropped from 2.82 to 2.34. Table 5.14 shows details of the regression results and figure 5.12 the residuals distribution. The latter shows less skewness as it is the case for the model without principal component in section 5.2.3.1.

Variables	Coef	P-value	Signif. level
(Intercept)	-5.83086	$< 2e-16$	***
PC1	0.04955	0.113578	
PC3	-0.19891	0.002078	**
PC7	0.21147	0.021759	*
PC8	0.17474	0.080074	.
PC9	-0.22301	0.043406	*
PC11	-0.19499	0.113986	
PC13	-0.20681	0.131468	
PC16	-0.79344	0.000283	***
PC17	0.97944	$4.73e-05$	***
PC20	0.58287	0.125514	
PC22	-1.52600	0.000427	***
PC23	1.99512	0.000306	***
PC24	-1.50028	0.012073	*
PC26	-1.52912	0.032027	*
PC27	-4.28871	0.001479	**
PC29	-8.05803	0.021553	*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.458 on 195 degrees of freedom

Multiple R-squared: 0.374, Adjusted R-squared: 0.322

F-statistic: 7.265 on 16 and 195 DF, p-value: 4.023e-13

10-fold-CV: 2.34, Number of Observations: 212

Table 5.14: 1st step parameter estimation for principal components based on variables with less than 10% missing values

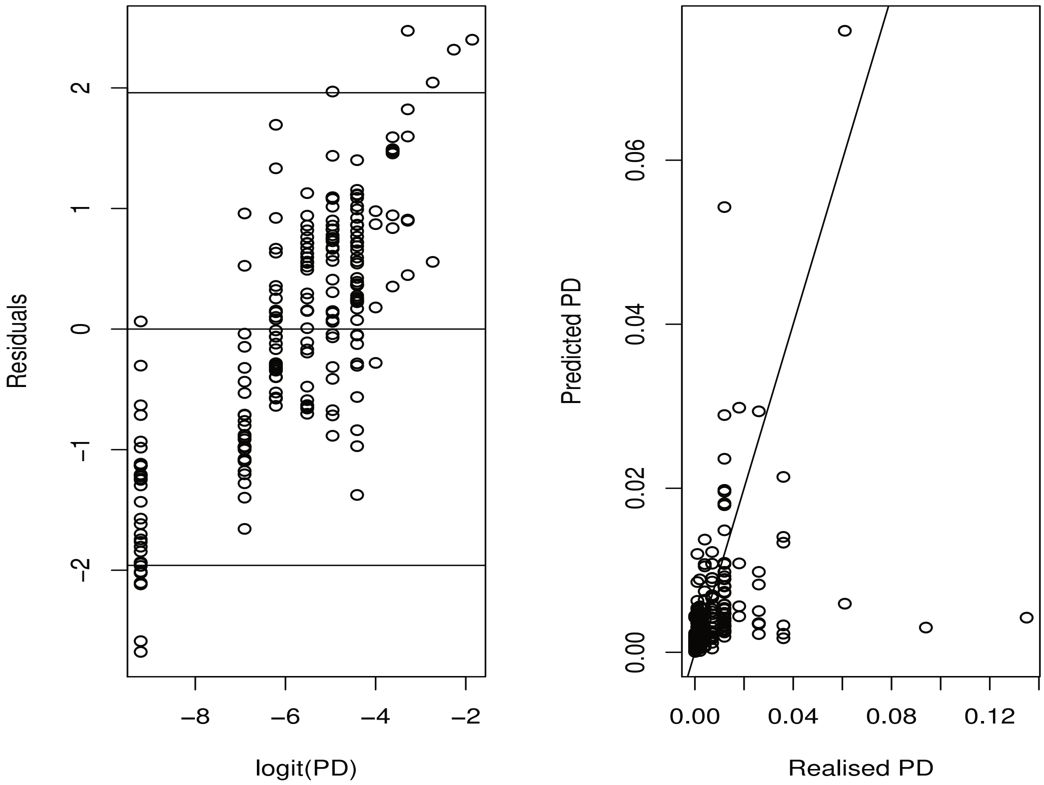


FIGURE 5.12: Residuals and predictions plots for 1st step parameter estimation for principal components based on variables with less than 10% missing values

After removing outliers from the input-data of the model in step 1, we obtain a model in step 2 with a higher goodness of fit of 0.524 (compared to step 1 0.382) and a drop of the 10-fold cross validation MSE from 2.26 to 1.45 (see table 5.15). The residual distribution (figure 5.13) also shows a better shape.

Variables	Coef	P-value	Signif. level
(Intercept)	-5.72513	$< 2e-16$	***
PC3	-0.29094	$1.21e-07$	***
PC7	0.27125	0.000197	***
PC8	0.15979	0.041801	*
PC9	-0.19155	0.027613	*
PC10	-0.14980	0.087229	.
PC12	-0.14139	0.167607	
PC13	-0.34825	0.001894	**
PC15	0.23860	0.111405	
PC16	-0.62651	0.000409	***
PC17	0.87464	$7.25e-06$	***
PC20	1.06489	0.000640	***
PC22	-0.98371	0.005885	**
PC23	2.30624	$2.16e-07$	***
PC24	-1.59702	0.000822	***
PC25	0.86287	0.113832	
PC26	-1.60362	0.004468	**
PC27	-3.13517	0.004430	**
PC29	-4.15594	0.151129	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.13 on 175 degrees of freedom

Multiple R-squared: 0.524, Adjusted R-squared: 0.475

F-statistic: 10.72 on 18 and 175 DF, p-value: < 2.2e-16

10-fold-CV: 1.45, Number of Observations: 194

Table 5.15: 2nd step parameter estimation for principal components based on variables with less than 10% missing values

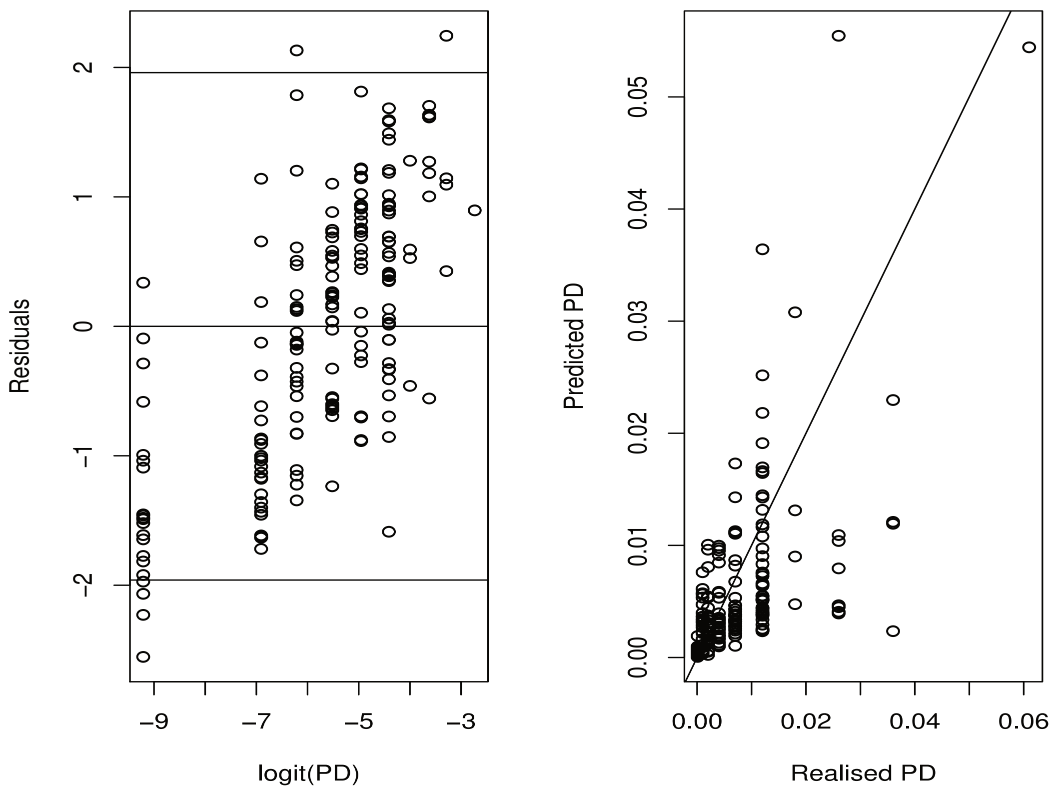


FIGURE 5.13: Residuals and predictions plots for 2nd step parameter estimation for principal components based on variables with less than 10% missing values

Since we still observe a slight pattern of dependency between predictions and residuals, we decide to take more variables into the model by increasing the threshold of missing value rate to 15%. The scree plot (figure 5.14) shows that 96% of the variance of the input-data is explained by the 16 first principal components. That is 2 more than it was the case for the variables with less 10% missing value rate. This is confirmed by the biplot (figure 5.15), which shows more directions as it was the case before.

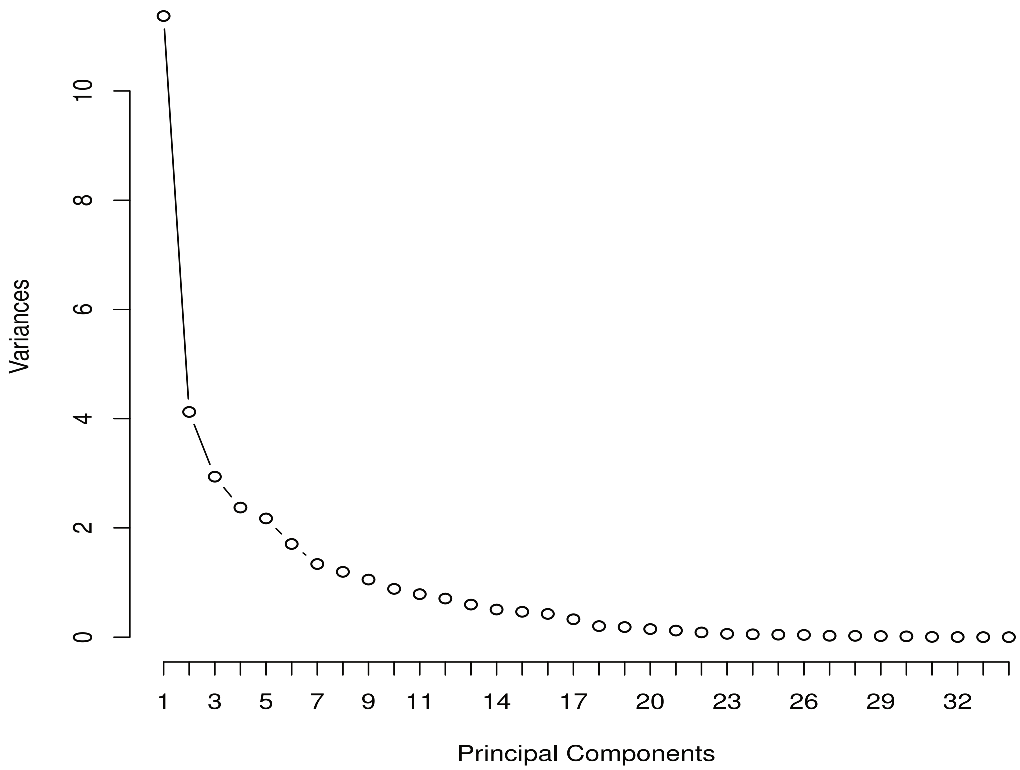


FIGURE 5.14: Scree plot of principal component for variables with less than 15% missing values

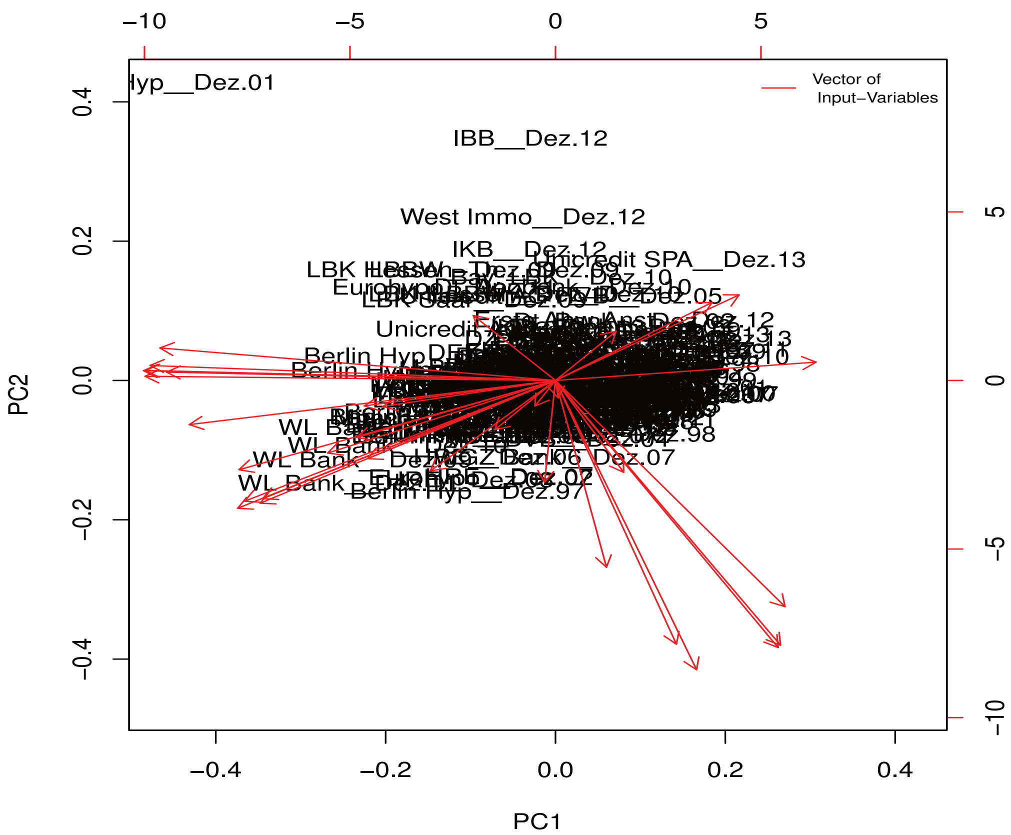


FIGURE 5.15: Biplot of principal component for variables with less than 15% missing values

Regression results are summarized in table 5.16 and residuals distribution in figure 5.16. One notices an increase of the goodness of fit to 0.420 (compared to 0.374) and a drop of the 10-fold cross validation MSE to 2.06.

Variables	Coef	P-value	Signif. level
(Intercept)	$-5.714e+00$	$< 2e-16$	***
PC1	$7.988e-02$	0.012272	*
PC2	$-8.585e-02$	0.103422	
PC3	$2.247e-01$	0.000731	***
PC4	$1.110e-01$	0.147329	
PC5	$3.251e-01$	0.001001	**
PC7	$-2.403e-01$	0.031499	*
PC9	$3.213e-01$	0.002395	**
PC13	$2.532e-01$	0.059287	.
PC16	$6.312e-01$	$9.52e-05$	***
PC19	$-1.206e+00$	$1.30e-06$	***
PC23	$9.876e-01$	0.018874	*
PC24	$-1.815e+00$	$8.33e-05$	***
PC26	$7.121e-01$	0.162670	
PC27	$2.216e+00$	0.000751	***
PC34	$8.078e+14$	$3.39e-05$	***
PC32	$3.922e+00$	0.085170	.
PC25	$7.677e-01$	0.113387	
PC10	$-1.585e-01$	0.147910	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.338 on 154 degrees of freedom

Multiple R-squared: 0.420, Adjusted R-squared: 0.352

F-statistic: 6.185 on 18 and 154 DF, p-value: 3.452e-11

10-fold-CV: 2.06, Number of Observations: 173

TABLE 5.16: 1st step parameter estimation for principal components based on variables with less than 15% missing values

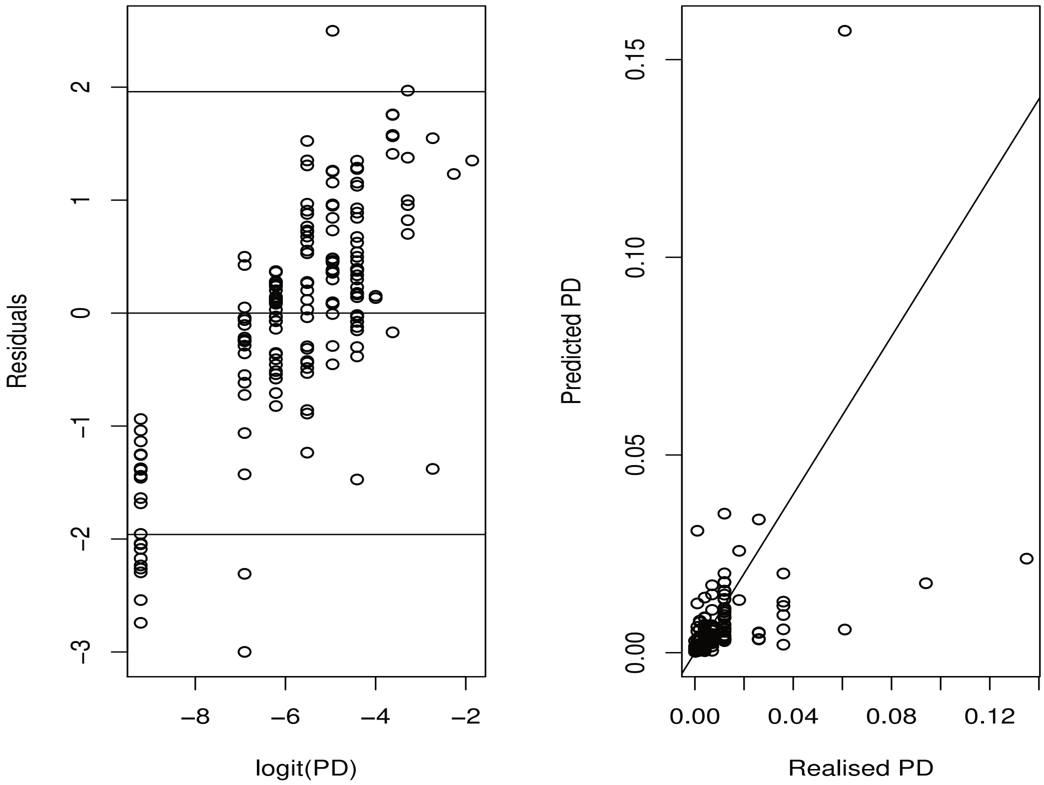


FIGURE 5.16: Residuals and predictions plots for 1st step parameter estimation for principal components based on variables with less than 15% missing values

After removing outliers from the data of the model in step 1, we obtain a model in step 2 (table 5.17) with a considerably higher goodness of fit of 0.624 (compared to 0.420 in step 1) and an enormous drop of the 10-fold cross validation MSE to 1.18 (compared to 2.06 in step 1). There is also a clear amelioration of the shape of the residual distribution (figure 5.17).

Variables	Coef	P-value	Signif. level
(Intercept)	$-5.508e+00$	$< 2e-16$	***
PC1	$6.257e-02$	0.006422	**
PC2	$-8.053e-02$	0.037108	*
PC3	$2.053e-01$	$2.44e-05$	***
PC5	$2.147e-01$	0.001670	**
PC7	$-2.752e-01$	0.000593	***
PC9	$4.256e-01$	$1.05e-07$	***
PC13	$1.879e-01$	0.054866	.
PC16	$6.843e-01$	$1.65e-08$	***
PC19	$-1.097e+00$	$3.02e-09$	***
PC23	$1.272e+00$	$3.90e-05$	***
PC24	$-2.097e+00$	$3.10e-09$	***
PC26	$5.047e-01$	0.180785	
PC27	$2.376e+00$	$2.21e-06$	***
PC31	$2.550e+00$	0.024136	*
PC34	$5.841e+14$	$8.48e-06$	***
PC10	$-1.778e-01$	0.026471	*
PC32	$4.731e+00$	0.004916	**
PC25	$6.467e-01$	0.066928	.
PC29	$-9.290e-01$	0.094979	.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9545 on 139 degrees of freedom

Multiple R-squared: 0.624, Adjusted R-squared: 0.572

F-statistic: 12.12 on 19 and 139 DF, p-value: < 2.2e-16

10-fold-CV: 1.18, Number of Observations: 159

Table 5.17: 2nd step parameter estimation for principal components based on variables with less than 15% missing values

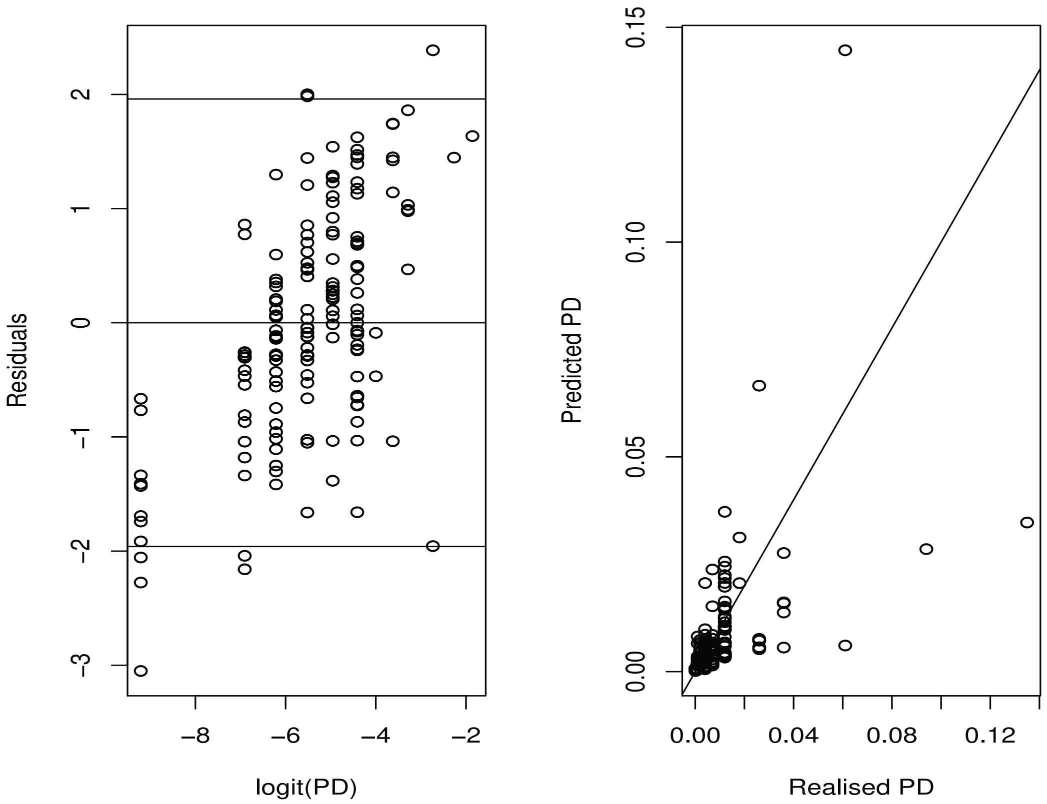


FIGURE 5.17: Residuals and predictions plots for 2nd step parameter estimation for principal components based on variables with less than 15% missing values

Taking more variables into the input-data of the model (see table 5.18 and figure 5.20), by increasing the threshold of missing value rate, leads to an increase of the goodness of fit (0.588) and a drop in the 10-fold cross validation MSE (1.65). Figure 5.18 and 5.19 shows the corresponding scree plot and biplot.

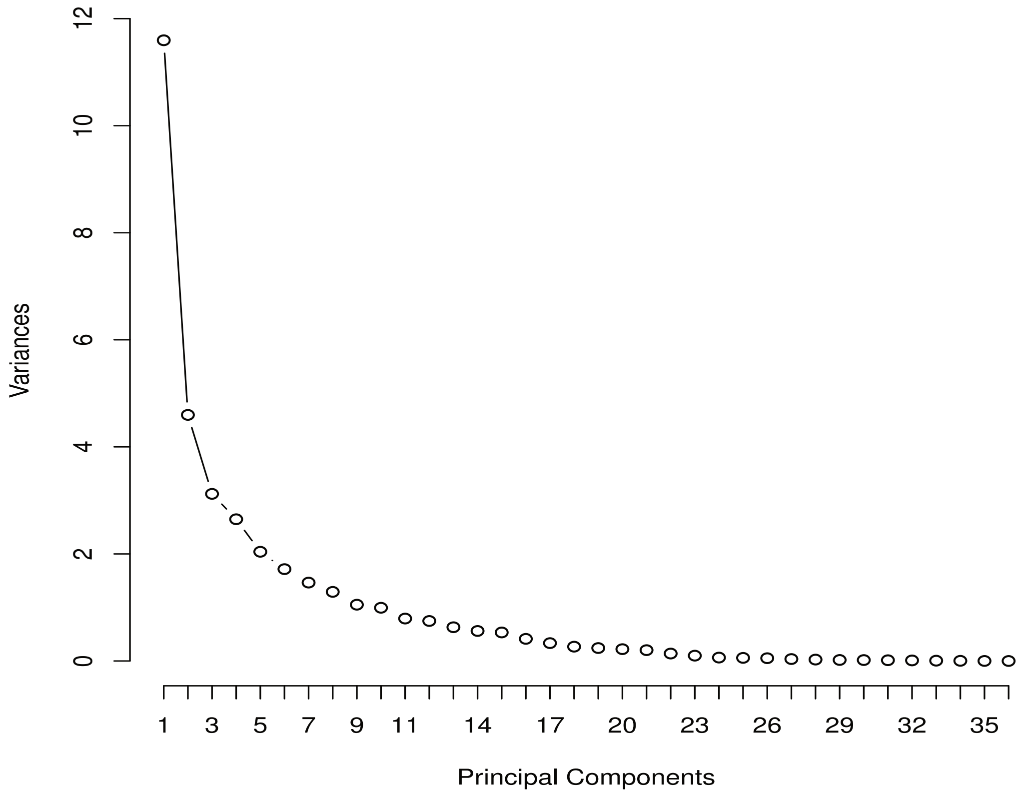


FIGURE 5.18: Scree plot of principal component for variables with less than 25% missing values

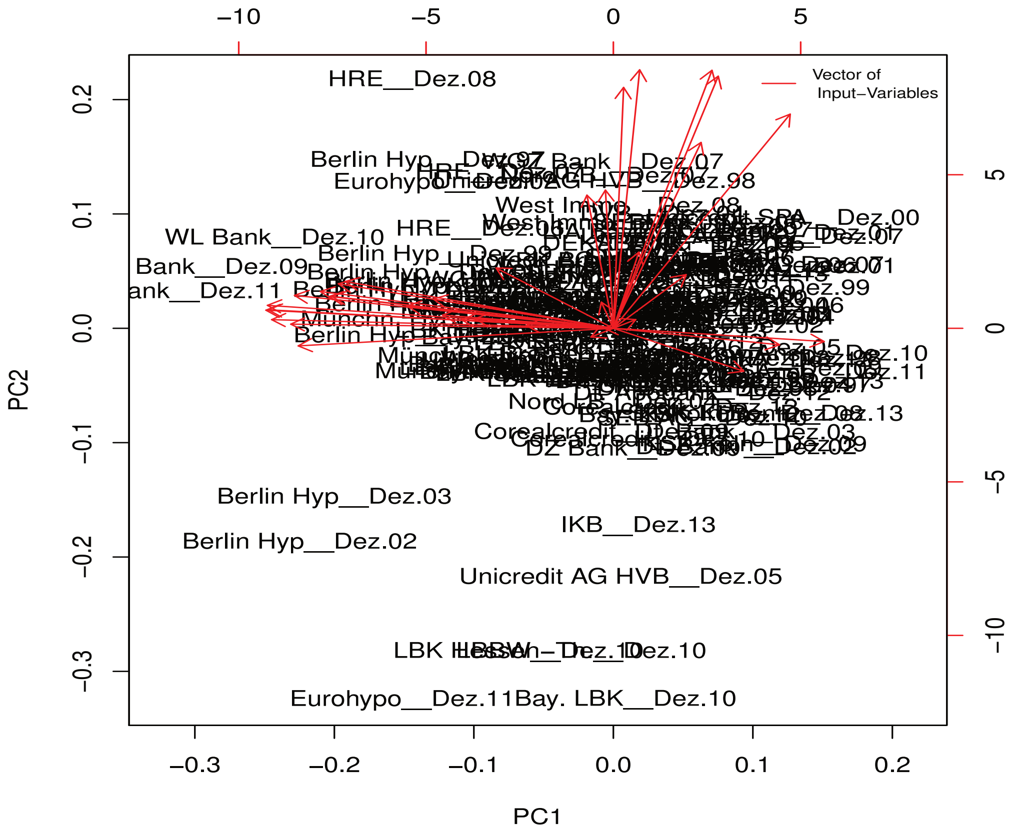


FIGURE 5.19: Biplot of principal component for variables with less than 25% missing values

Variables	Coef	P-value	Signif. level
(Intercept)	$-5.706e+00$	$< 2e-16$	***
PC3	$1.870e-01$	0.010790	*
PC5	$1.588e-01$	0.099890	.
PC9	$-4.158e-01$	$6.48e-05$	***
PC11	$1.958e-01$	0.073372	.
PC12	$-2.407e-01$	0.034311	*
PC16	$7.749e-01$	$1.05e-06$	***
PC17	$-4.622e-01$	0.006776	**
PC20	$4.899e-01$	0.019619	*
PC21	$-1.037e+00$	$3.96e-06$	***
PC22	$-1.005e+00$	0.000180	***
PC23	$-8.410e-01$	0.007050	**
PC24	$-7.779e-01$	0.045369	*
PC26	$-1.169e+00$	0.006671	**
PC27	$1.153e+00$	0.025758	*
PC28	$-2.510e+00$	$5.88e-05$	***
PC30	$-2.688e+00$	0.000429	***
PC34	$4.856e+00$	0.009489	**
PC35	$6.657e+00$	0.174206	
PC36	$3.178e+14$	0.050047	.
PC7	$1.379e-01$	0.087327	.
PC8	$1.325e-01$	0.130620	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.154 on 122 degrees of freedom

Multiple R-squared: 0.588, Adjusted R-squared: 0.518

F-statistic: 8.307 on 21 and 122 DF, p-value: 3.839e-15

10-fold-CV: 1.65, Number of Observations: 144

Table 5.18: 1st step parameter estimation for principal components based on variables with less than 25% missing values

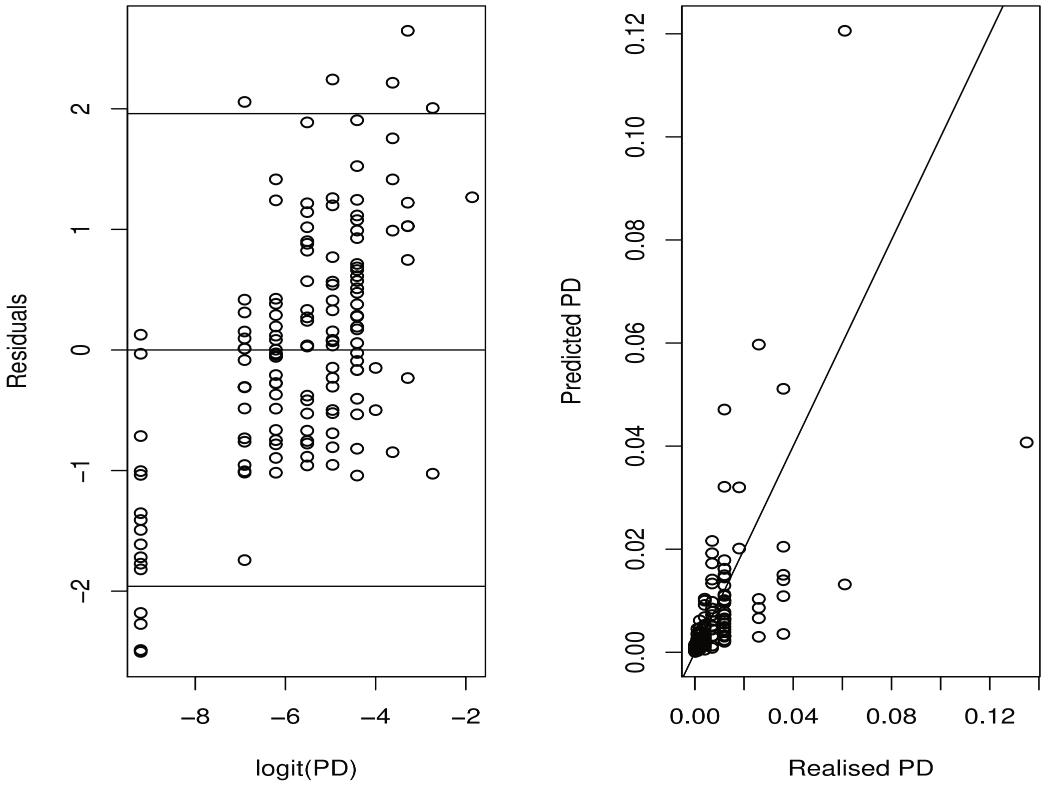


FIGURE 5.20: Residuals and predictions plots for 1st step parameter estimation for principal components based on variables with less than 25% missing values

Removing outliers increase the goodness of fit (0.653 vs 0.588) and reduce the 10-fold cross validation MSE (1.21 vs 1.65).

Variables	Coef	P-value	Signif. level
(Intercept)	$-5.635e+00$	$< 2e-16$	***
PC3	$2.402e-01$	0.000204	***
PC5	$1.494e-01$	0.072867	.
PC9	$-4.492e-01$	$1.13e-06$	***
PC11	$2.384e-01$	0.012821	*
PC12	$-1.936e-01$	0.049916	*
PC16	$7.568e-01$	$6.88e-08$	***
PC17	$-4.028e-01$	0.007229	**
PC20	$4.602e-01$	0.011589	*
PC21	$-8.854e-01$	$6.58e-06$	***
PC22	$-1.098e+00$	$4.09e-06$	***
PC23	$-8.478e-01$	0.002028	**
PC24	$-8.781e-01$	0.009871	**
PC26	$-1.042e+00$	0.005638	**
PC27	$1.144e+00$	0.011585	*
PC28	$-2.509e+00$	$4.81e-06$	***
PC30	$-2.695e+00$	$5.69e-05$	***
PC33	$-2.107e+00$	0.067880	.
PC34	$4.064e+00$	0.012294	*
PC36	$3.164e+14$	0.022744	*
PC7	$9.803e-02$	0.161382	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.998 on 118 degrees of freedom

Multiple R-squared: 0.653, Adjusted R-squared: 0.594

F-statistic: 11.11 on 20 and 118 DF, p-value: < 2.2e-16

10-fold-CV: 1.21, Number of Observations: 139

Table 5.19: 2nd step parameter estimation for principal components based on variables with less than 25% missing values

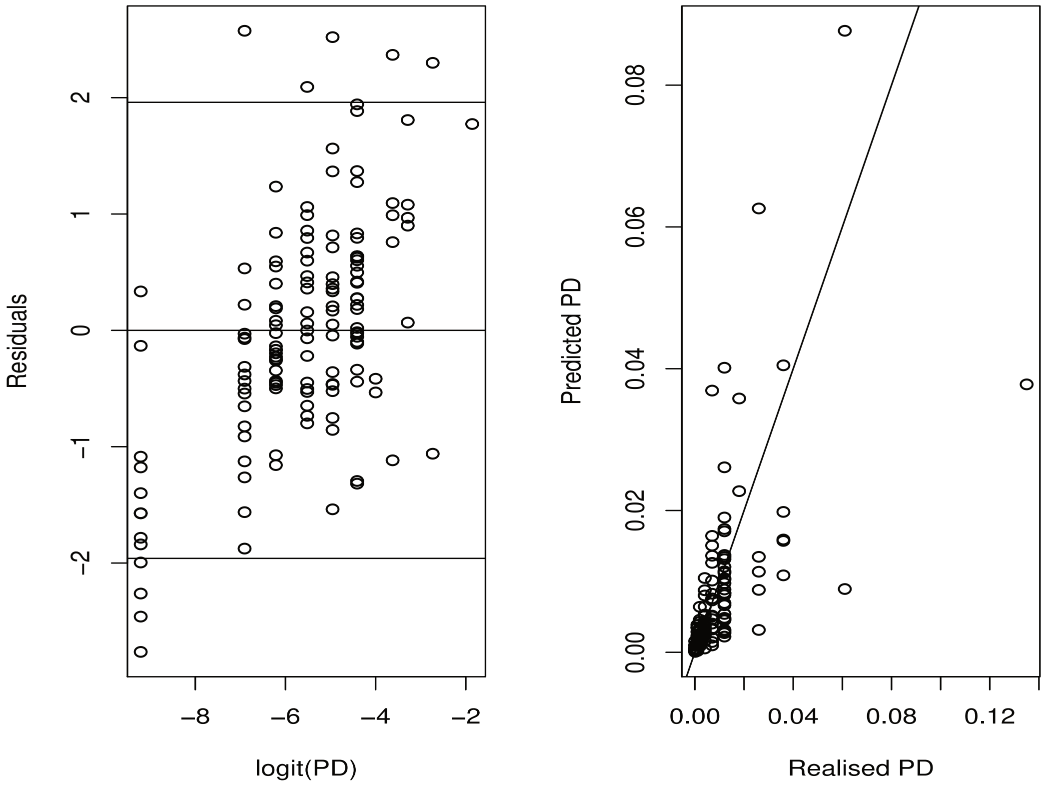


FIGURE 5.21: Residuals and predictions plots for 2nd step parameter estimation for variables with less than 25% missing values

5.3 Conclusion

Based on the observation that financial statement are commonly provided in the last quarter of the year, we assumed, that external rating agencies publish the rating after using financial statement of the immediate past period. We therefore started the analysis by joining the actual (t) credit performance indicators (external ratings) with the predictors (financial ratios) of the year before ($t+1$). In order to verify this assumption and identify possible relevant lagged variables, we have performed several stepwise regressions on lagged data from 1 to 5 years (cleaned from correlated variables). We could observed that with increasing years the goodness of fit of the regression was decreasing, so that the data for lag 1 shows the highest goodness of fit as we have expected. Furthermore we compared the final models for all lagged data to identify lagged variables, which have not been taken into account in the model for lag 1. We could identify some variables, but we finally decided not to include them in the model for lag 1, because it led to a considerable reduction of the number of observations in the input-data.

Thereafter we performed stepwise regressions on lag 1 data for several subset of predictors (depending on missing value rates). It led to regression results with low goodness of fit. Related residuals analysis gave indication on miss-specification of the model and the necessity to include more relevant variables. This is the reason why in the second approach all variables have been included to determine principal components, which we use as predictors in stepwise regressions. It results in significant regression results with much higher goodness of fit. The improvement is also reflected in a better shape of the residuals distribution, which is less skewed than in the case of the first approach.

We found out that the model based on the principal component for variables with less than 25% missing values (see 5.19) is the most suitable model for the cross-sectional approach. It has an adequate combination of high goodness of fit (0.65), low 10-fold cross-validation MSE (1.21) and better shape of residuals distribution (see 5.21). Furthermore variables related to asset show the highest positive impact on creditworthiness and variables related to equity the highest negative effect.

Chapter 6

Rating Model based on Longitudinal Analysis

In this chapter we present the analysis' results for our proposed rating model based on longitudinal approach. We perform the analysis for the model with the highest in-sample goodness of fit for the cross-sectional approach based on principal components (see table 5.19). The corresponding input-data contains 139 observations based on 17 time points from 1997 to 2013. Table 6.1 gives an overview of the number of observations per point in time. We start the analysis by determining fixed and random effects. Thereafter we make a one point prediction of the latter using exponential smoothing.

The analysis have been carried out in R Core Team [25]. In addition to package base, packages nlme (Pinheiro et al. [29]), lmerTest (Kuznetsova et al. [30]) and forecast (Hyndman and Khandakar [31]) have been used.

Date	1997	1998	1999	2000	2001	2002	2003	2004	2005
Nr of Obs	2	3	5	6	5	6	5	7	9
Date	2006	2007	2008	2009	2010	2011	2012	2013	
Nr of Obs	11	13	11	13	10	11	10	12	

Table 6.1: Overview of time series

6.1 Estimation of Fixed and Random Effects

After performing the linear mixed model with random intercept, we obtain the estimates in table 6.3. As one can see the estimates in linear mixed model (table 6.3) are very similar (coefficients almost equal) with the estimate of the corresponding linear model (5.19). The difference lies in the random effect, which moves the general level of the probability of default at each point in time. Table 6.2 gives an overview of the random effects and figure 6.1 the residuals plot. The latter shows a less skewed distribution. The majority of the predicted PD tend to lay on or around the bisectrix.

The estimates of the fixed effects can be used to predict probability of default of debt security issuers in 2014. This is not the case for the random effects since they are particular for each point in time. We therefore need to estimate it for 2014. We perform it in section 6.2 using double exponential smoothing as explained in section 4.4.3.2.

Date	1997	1998	1999	2000	2001	2002	2003	2004	2005
Rd. eff.	0.0353	0.0360	0.0885	0.0474	-0.0580	-0.0236	-0.0042	-0.1278	-0.0499
Date	2006	2007	2008	2009	2010	2011	2012	2013	
Rd. eff.	-0.0038	0.0550	-0.1004	0.0047	0.0183	-0.1480	0.1089	0.1216	

Table 6.2: Overview of random effects

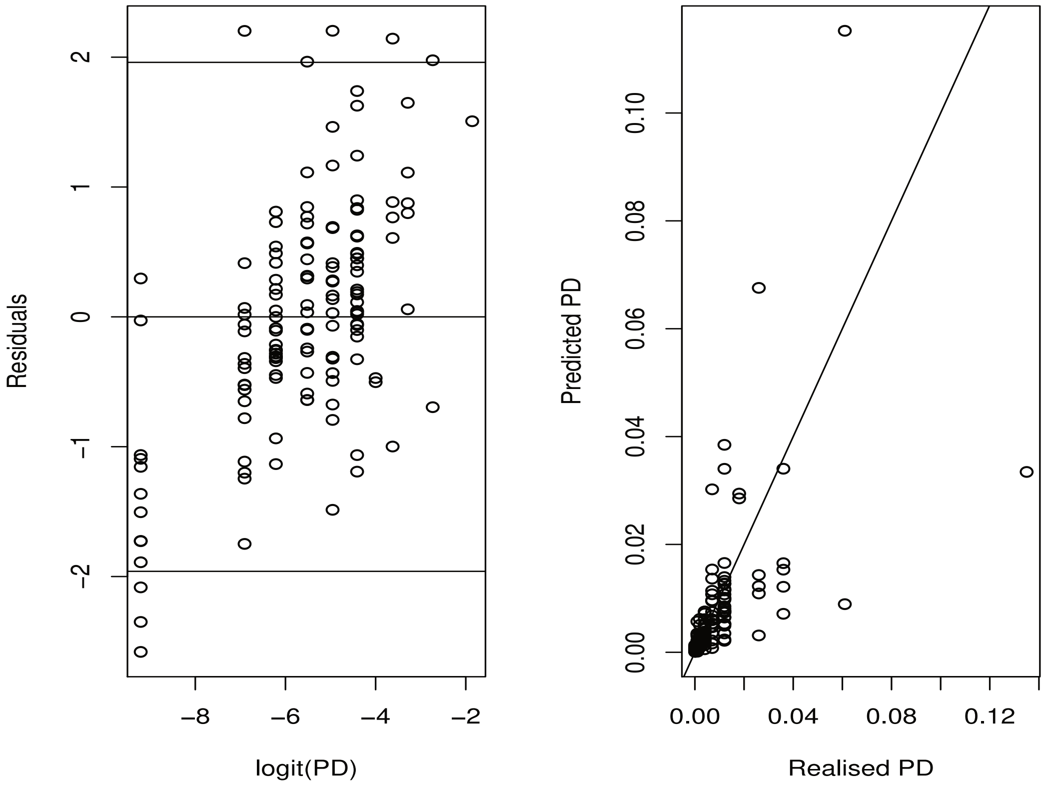


Figure 6.1: Residuals and predictions plots for the linear mixed model based on principal components for variables with less than 25% missing values

Variables	Coef	P-value	Signif. level
(Intercept)	$-5.633e+00$	$5.33e-15$	***
PC3	$2.372e-01$	0.000265	***
PC5	$1.405e-01$	0.090809	.
PC9	$-4.301e-01$	$2.77e-06$	***
PC11	$2.407e-01$	0.011431	*
PC12	$-1.973e-01$	0.046128	*
PC16	$7.579e-01$	$6.29e-08$	***
PC17	$-4.040e-01$	0.007011	**
PC20	$4.553e-01$	0.012817	*
PC21	$-9.077e-01$	$4.34e-06$	***
PC22	$-1.114e+00$	$3.44e-06$	***
PC23	$-8.455e-01$	0.001951	**
PC24	$-9.122e-01$	0.007484	**
PC26	$-1.045e+00$	0.005662	**
PC27	$1.135e+00$	0.011838	*
PC28	$-2.530e+00$	$3.59e-06$	***
PC30	$-2.714e+00$	$4.99e-05$	***
PC33	$-2.212e+00$	0.053573	.
PC34	$4.149e+00$	0.010345	*
PC36	$2.914e+14$	0.035071	*
PC7	$9.769e-02$	0.165260	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

REML criterion at convergence: 332.4, R-squared: 0.6721

Date: 17, Number of Observations: 139

TABLE 6.3: Estimation of fixed effects for mixed model based on principal components for variables with less than 25% missing values

6.2 One Point Prediction of Random Effect

We performed a double exponential smoothing to determine related parameter α and β. We then use the model to predict the random effect in 2014. The figure 6.2 shows the smoothed values of the random effect (blue line) and related prediction for 2014 with confidence interval (blue vertical band).

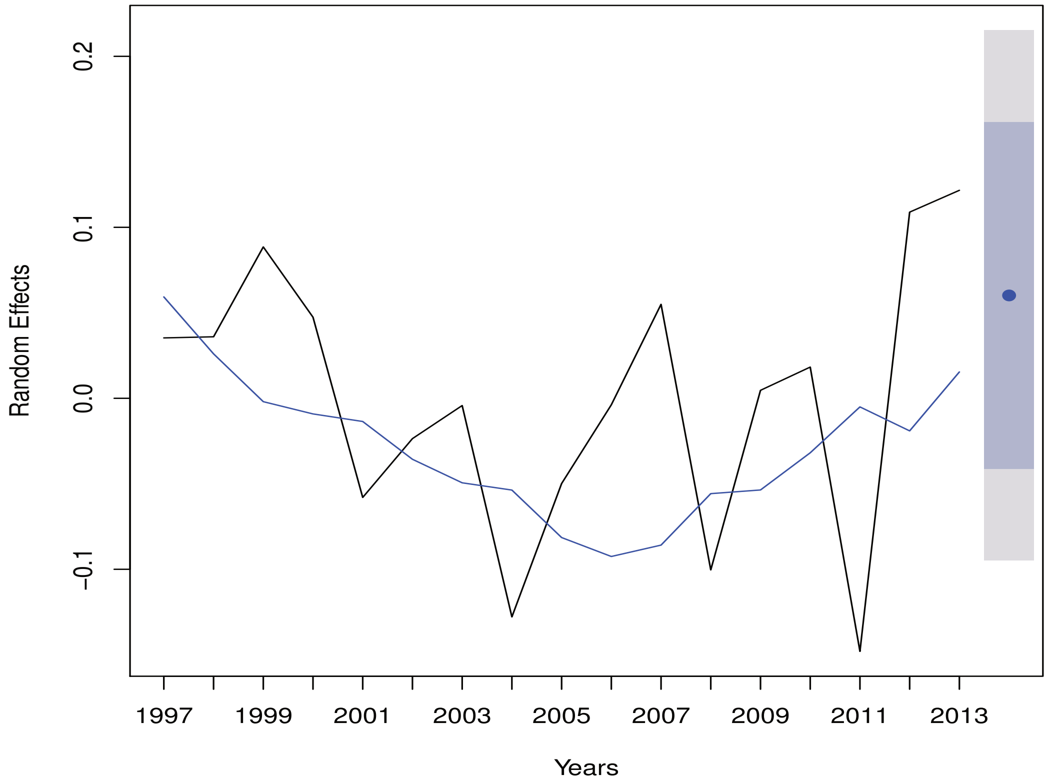


FIGURE 6.2: Forecast of the random effects based on double exponential smoothing

6.3 Cross-sectional vs. Longitudinal Approach

The model based on the longitudinal approach shows an improvement of the goodness of fit compared to the model based on the cross-sectional approach. The R-squared increases from 0.65 to 0.67 and the log-likelihood from −180.62 to −166.21. Others indicators of improvement in goodness of fit are given by the AIC and BIC. Both decrease by around 100, that is, from 415.25 to 378.42 for AIC and from 479.81 to 445.91 for BIC. A better picture of the reduction of miss-classification error is given by table 6.4. It shows that for three observations the differences between the primary ratings and the predicted ratings have reduced. Two with a difference of 2 notches, are now correctly classified and one with a difference of 3 notches shows with the longitudinal model a difference of only 1 notch.

Diff. of rating-orders - cross-sectional	0	1	2	3	4	5
Nr. of Observations	51	53	22	10	2	1
Diff. of rating-orders - longitudinal	0	1	2	3	4	5
Nr. of Observations	53	54	20	9	2	1

TABLE 6.4: Overview miss-classification error

This observation is confirmed by an amelioration of the residuals distribution (figure 6.3) and hence better predictions (figure 6.4).

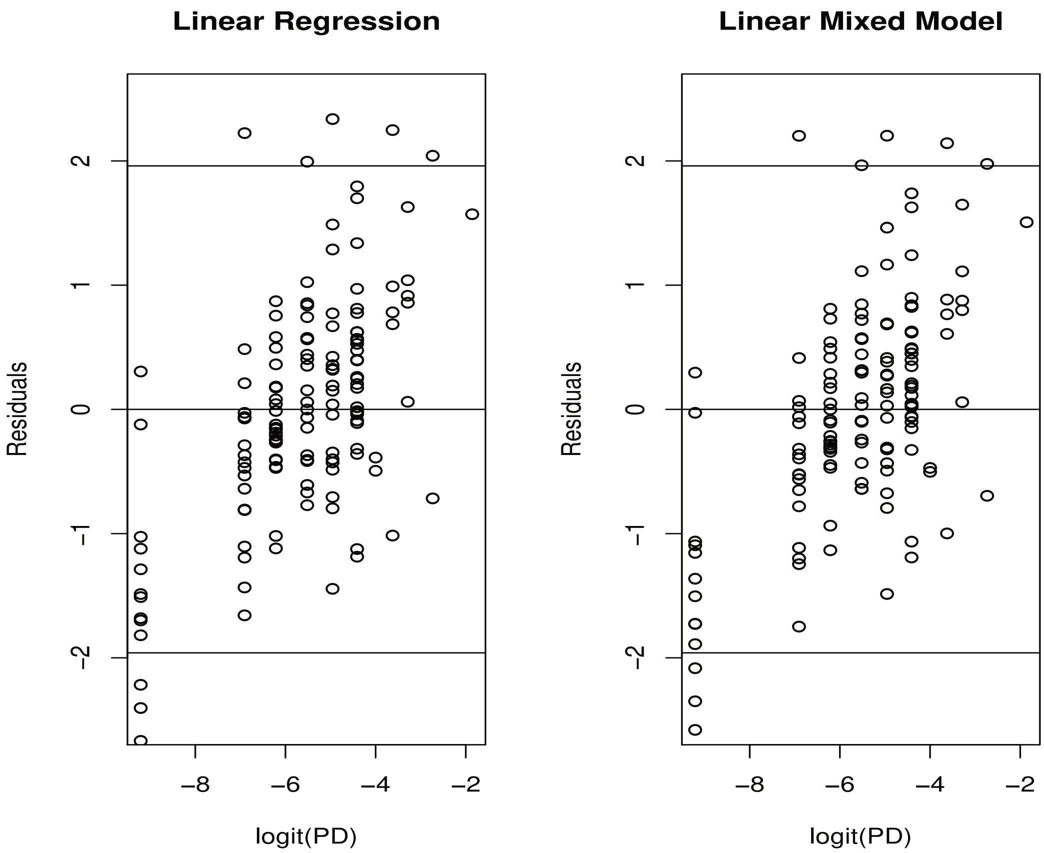


FIGURE 6.3: Forecast of the random effects based on double exponential smoothing

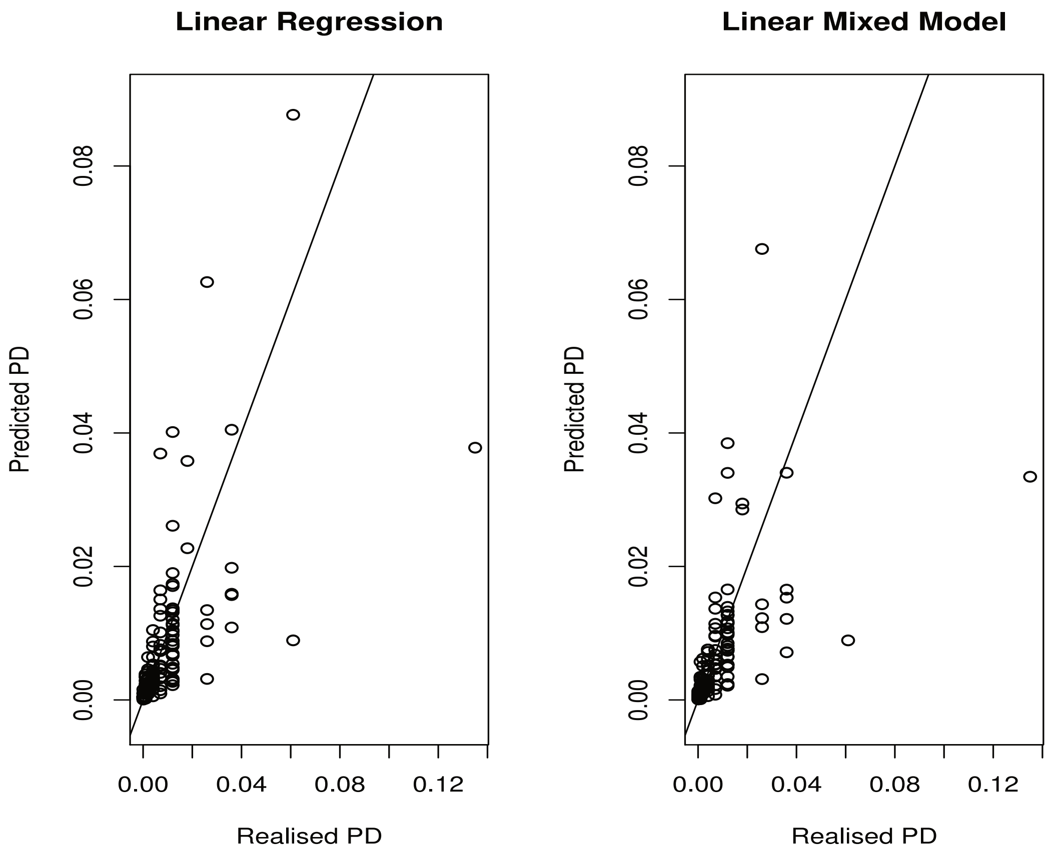


FIGURE 6.4: Forecast of the random effects based on double exponential smoothing

6.4 Conclusion

Based on the best model of the cross-sectional analysis, we performed a longitudinal analysis. Though the relative short time series of 17 points in time, we could determine a longitudinal model with significant systematic and idiosyncratic parameter. Moreover we could show, that taking systematic risk into account while modeling idiosyncratic risk can improve the goodness of fit of the model.

However because of the short size of the time series, it smoothing and prediction using double exponential smoothing shows some inaccuracy. Nonetheless double exponential smoothing deliver the best predictions under all other time series analysis methods, we could try on this short time series.

Chapter 7

Conclusion

The goal of this study was to develop a statistical model to assess creditworthiness of debt securities' issuers in the PSVaG's credit portfolio. In developping the model, we address the research questions of a) using rating-grades as a proxy for default indicator, b) developing a cross-sectional model based on a small number of observations, c) developing a longitudinal model for short time series of sparse data and d) estimating systematic risk.

Our basis dataset contained information about banks and financial institutions in Germany, from which PSVaG had purchased debt securities. It was composed of external ratings of securities issuers as credit performance indicators and a large number of financial ratios as predictors. We had relative short time series of 17 points in time with few observations per point in time. An other data quality issue was the slight heterogeneity of the financial ratios over time. This is due to the fact that with the beginning of the 21st century, German banks and financial institutions enhanced their investment in manifold securitized bonds of mortgage loan in the USA using conduits. Since conduits are not subsidiary of the related investor, their risks did not appear in the corresponding balance sheet, even if they had a repercussion on the return on equity. Following the insolvency of the Texan energy company Eron at late 2001 because of misuse of conduits, the European Union Parliament adopted new regulation requiring banks to publish activities of the conduits in their balance sheets from 2005 onwards. For Landesbanken (state banks) it started in 2008 (see Mueller [3] p. 35 - 40). Hence financial ratios before and after 2005 for banks and 2008 for Landesbanken may show slight structural differences.

In spite of the data quality restriction, we could determine a cross-sectional model with a satisfactory in-sample goodness of fit of 65% and an acceptable out-of-sample one. Variables related to asset show the highest positive impact on the creditworthiness, while variables related to equity show the highest negative effect. We then extend the model by proposing a longitudinal modeling approach of short time series of scarce data. The model was composed of an idiosyncratic (fixed effect) and a systematic (random effect) components for which we determine the best linear unbiased estimator (BLUE) resp. the empirical best unbiased predictor (eBLUP). Properties of our model approach have been proven in P. McCullagh [32]. The longitudinal model enabled us to increase the in-sample goodness of fit up to 67% and obtain a rating-prediction for which 76% of the observations have at most a miss-classification error of 1 notch. Compared to other similar studies such as Porath [6], which achieved a R^2 of 65% using logit, bank internal data and macroeconomic indicators, we can affirm that we could achieve an acceptable prediction result based on our proposed model approach.

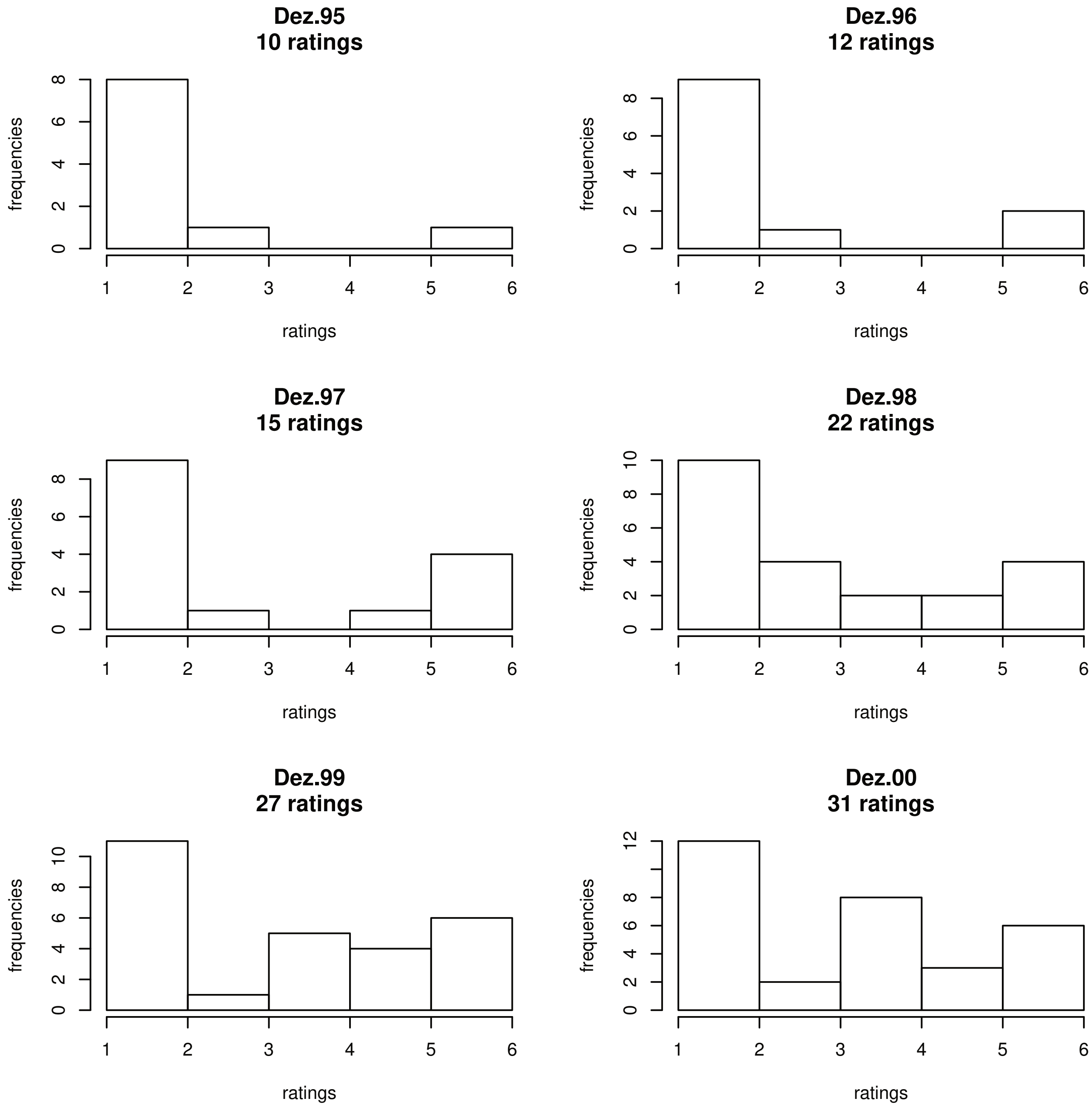
A possible hypothesis for the slight difference in prediction between our proposed model and the historical rating-grades of the debt securities' issuers is, that our model focuses on the quantitative estimation of the creditworthiness, while external ratings agencies use additional mostly subjective information to determine the final probability of default.

For this study we have neglected contagion effects of banks belonging to the same financial consortium such as state banks (Landesbanken) and saving banks (Sparkassen). We further admitted a particular correlation structure with idiosyncratic and systematic components to develop our model. A possible extension of this study could consist of investigating other correlation structures. Alternatively one could also investigate the use of more sophisticated methods such as GARCH to predict systematic risk. Furthermore the inclusion of macroeconomic variables to perform a simultaneous estimation of idiosyncratic and systematic risk could be investigated.

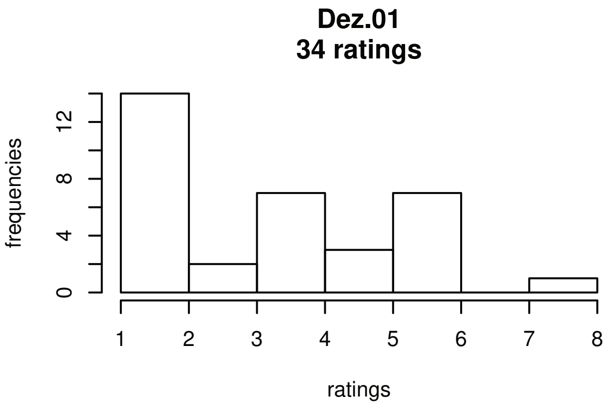
Appendix A

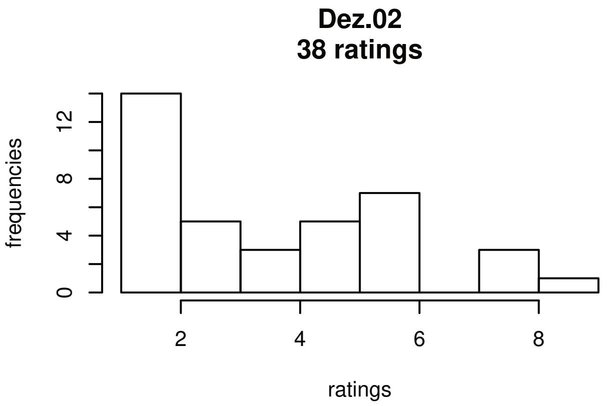
Appendix - Rating - Cross-Sectional Plausibility Check

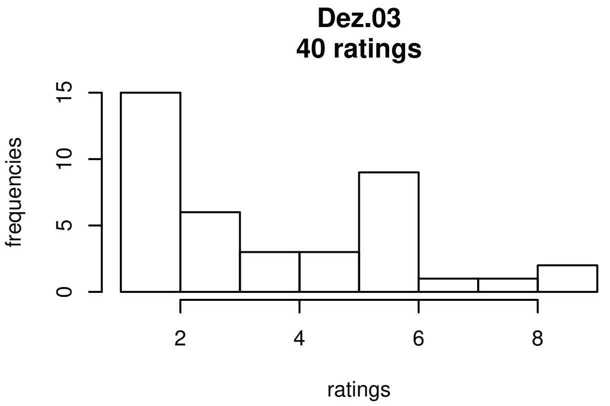
Rating - Cross-Sectional Plausibility Check

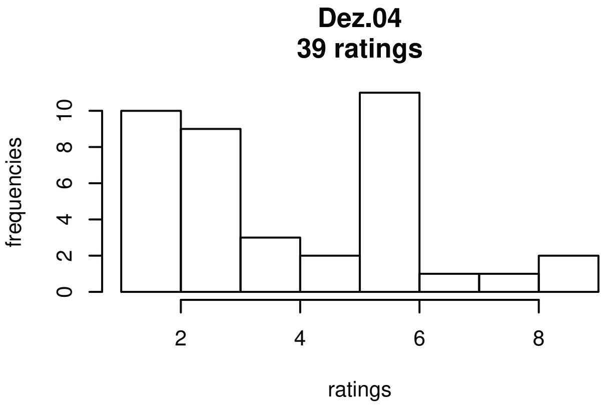


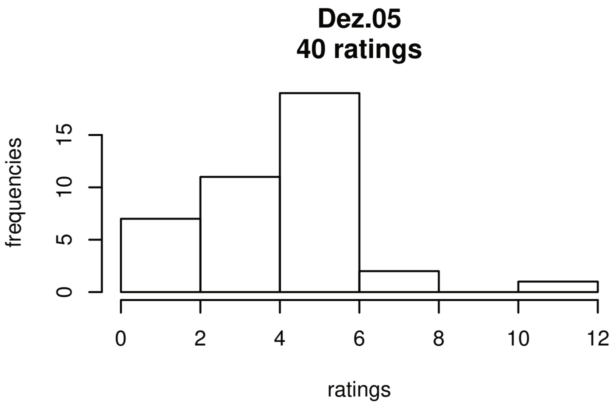
Rating - Cross-Sectional Plausibility Check

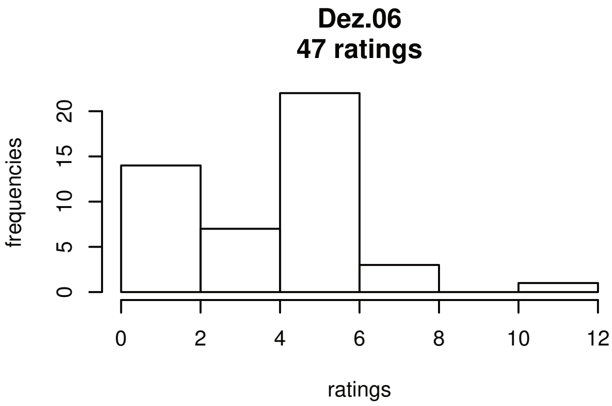




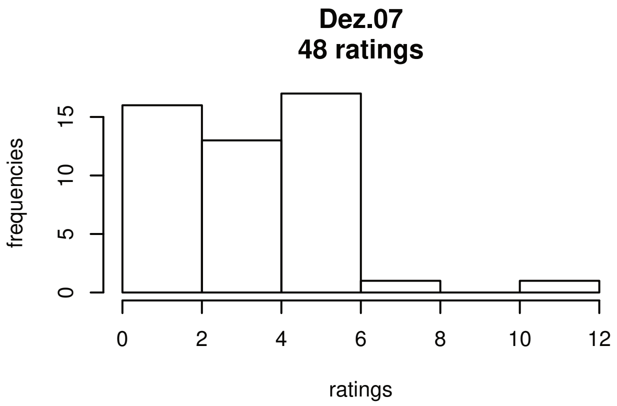


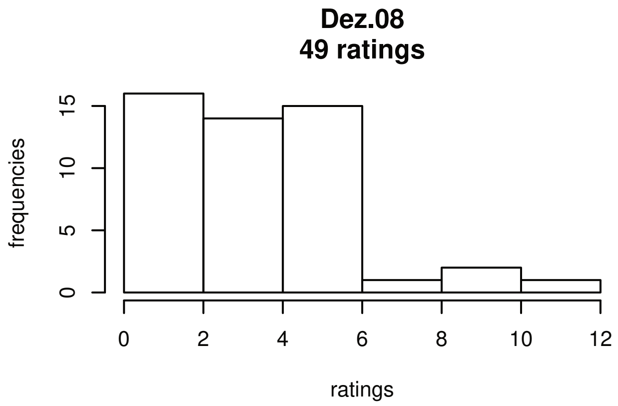


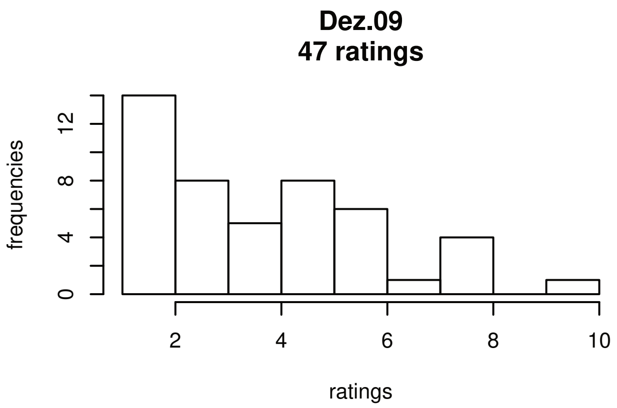


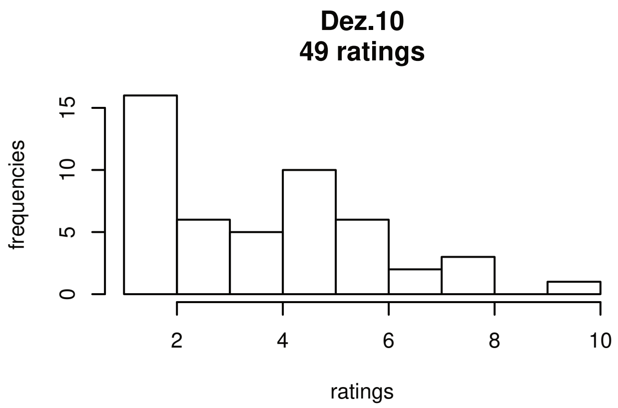


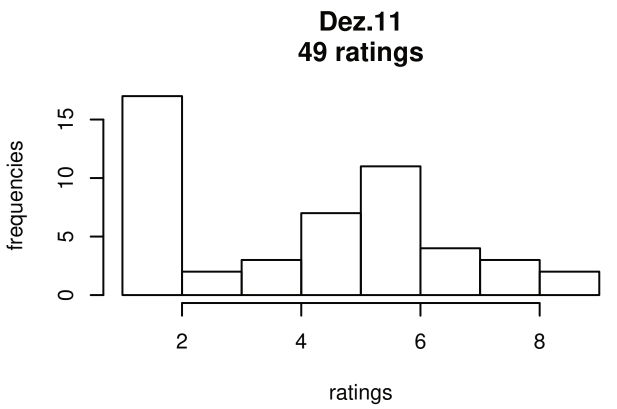
Rating - Cross-Sectional Plausibility Check

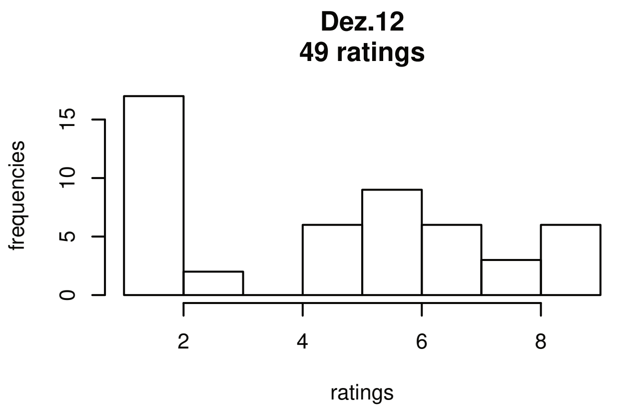




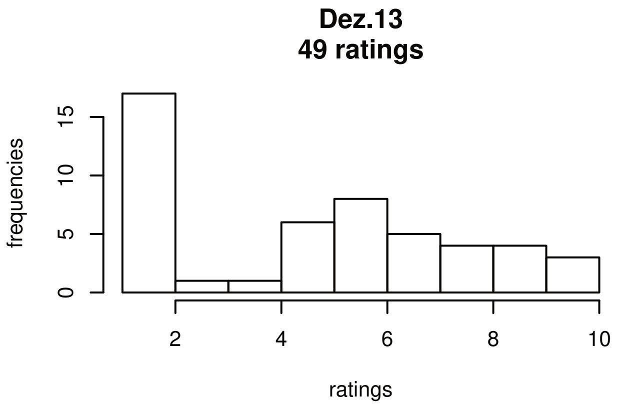


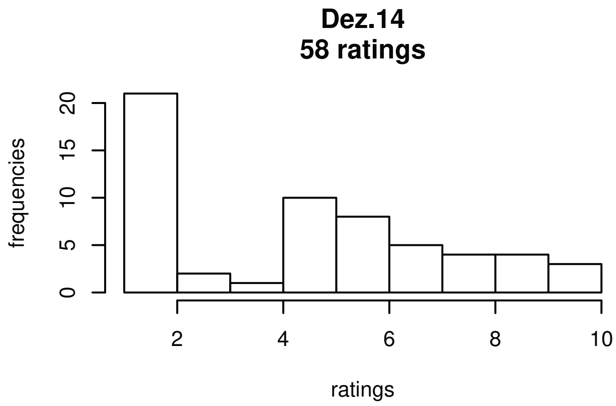


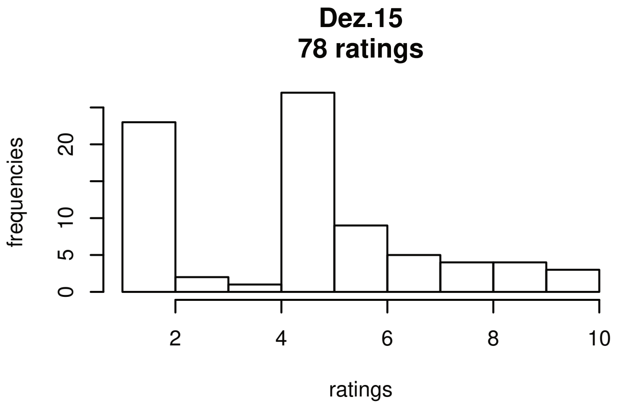




Rating - Cross-Sectional Plausibility Check



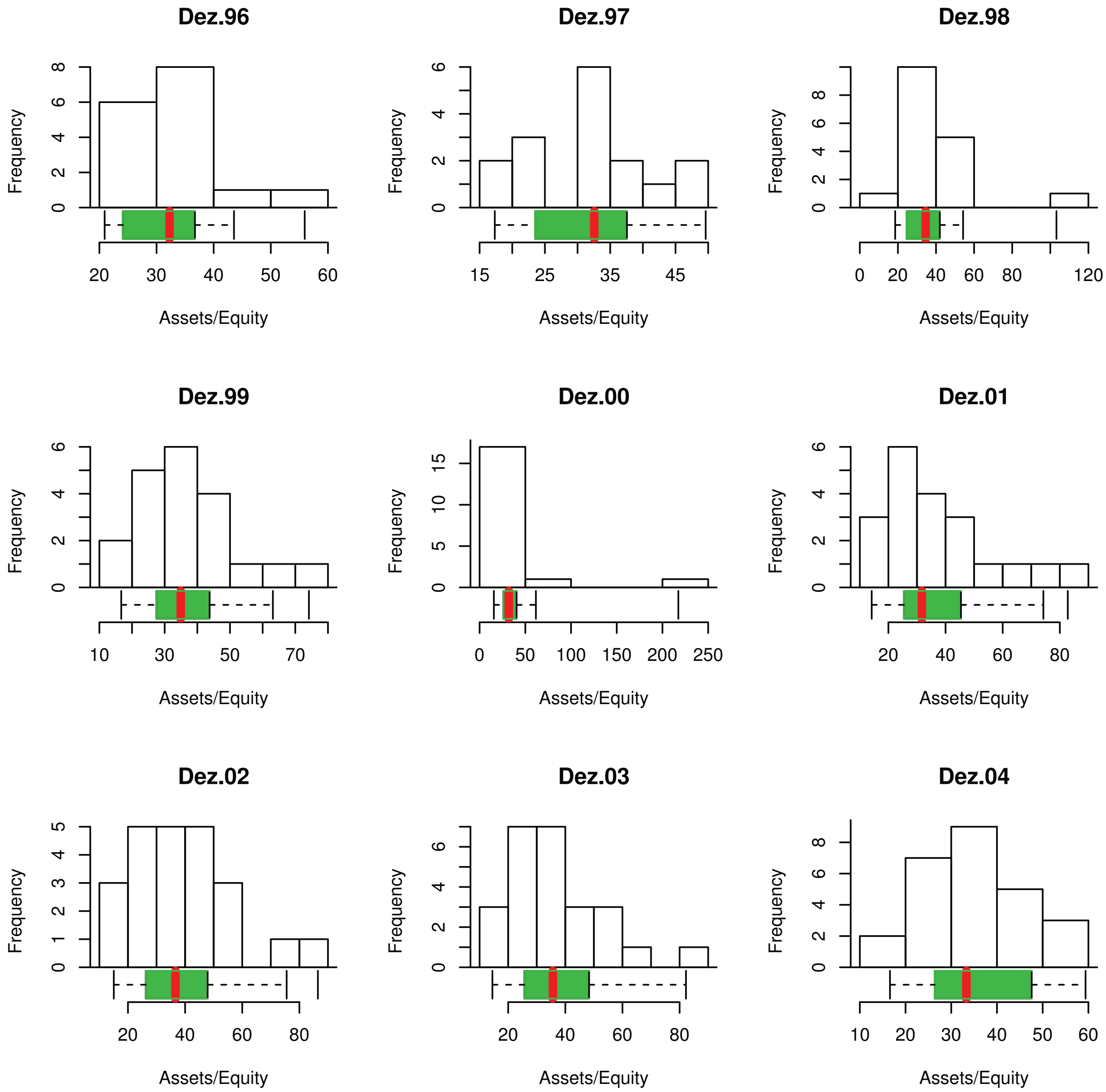




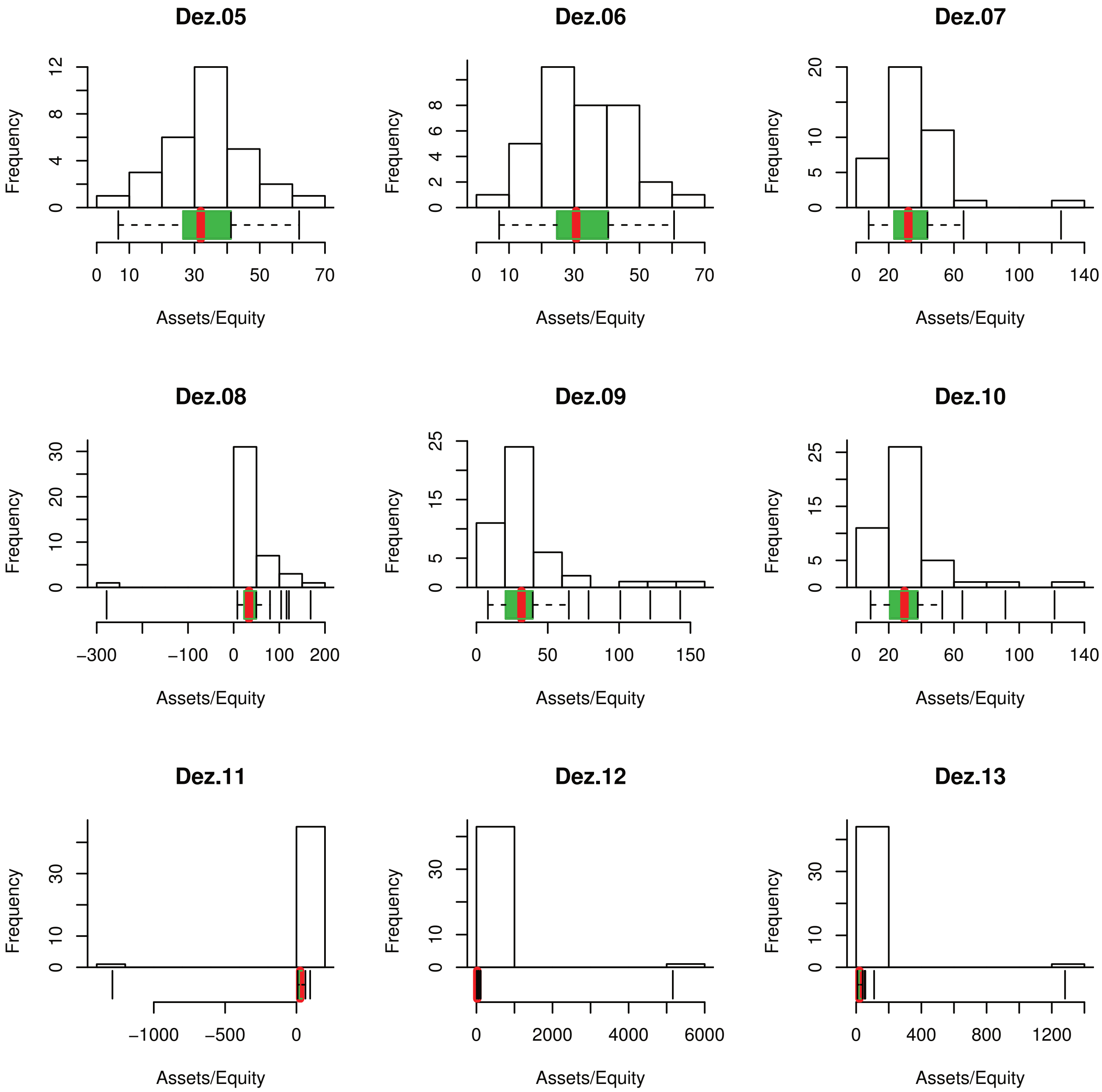
Appendix B

Appendix - Financial Information - Cross-Sectional Plausibility Check

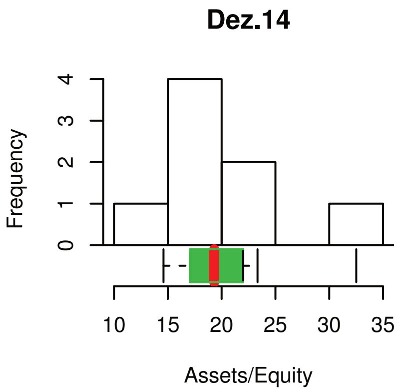
Assets/Equity



Assets/Equity



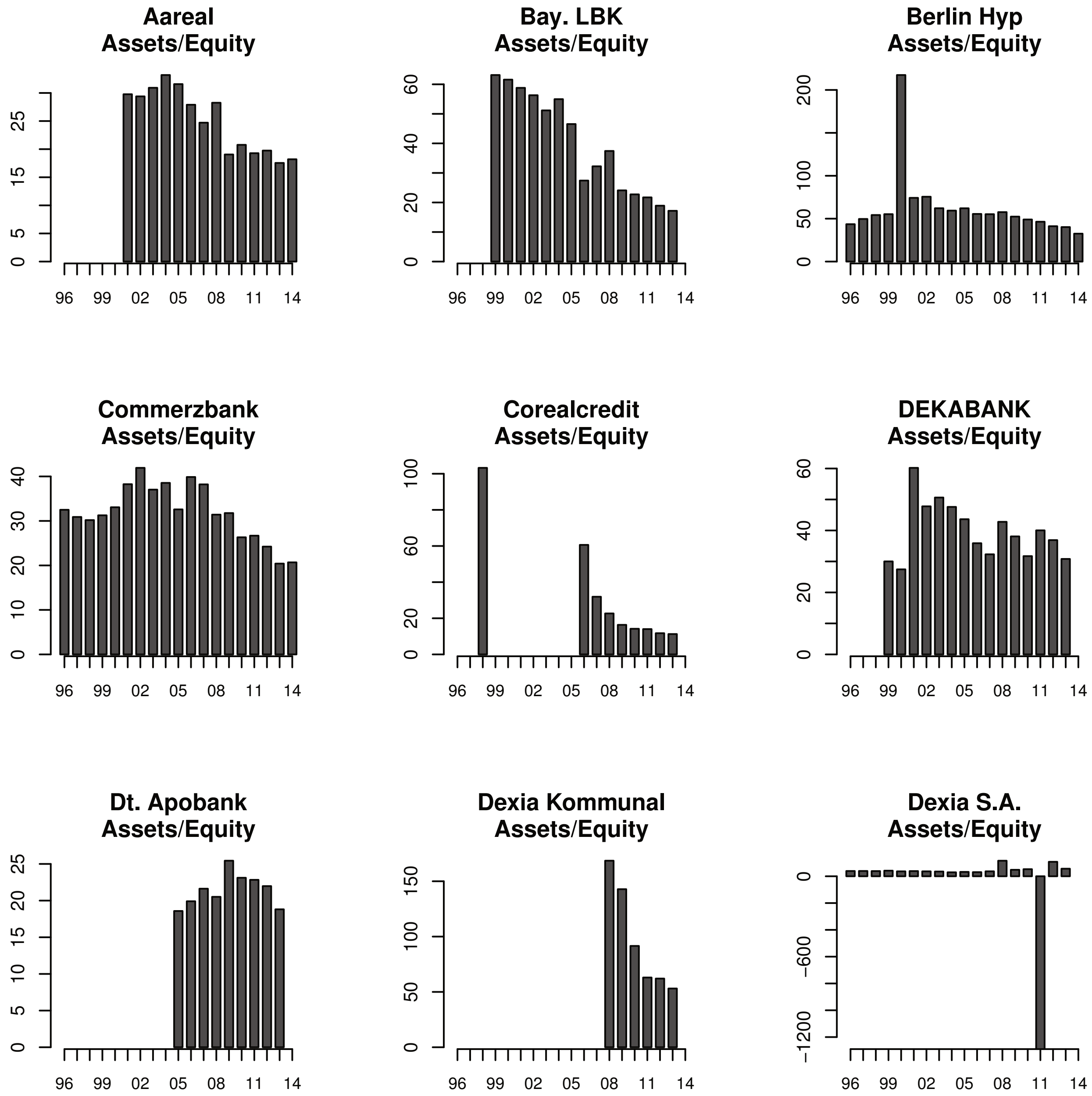
Assets/Equity



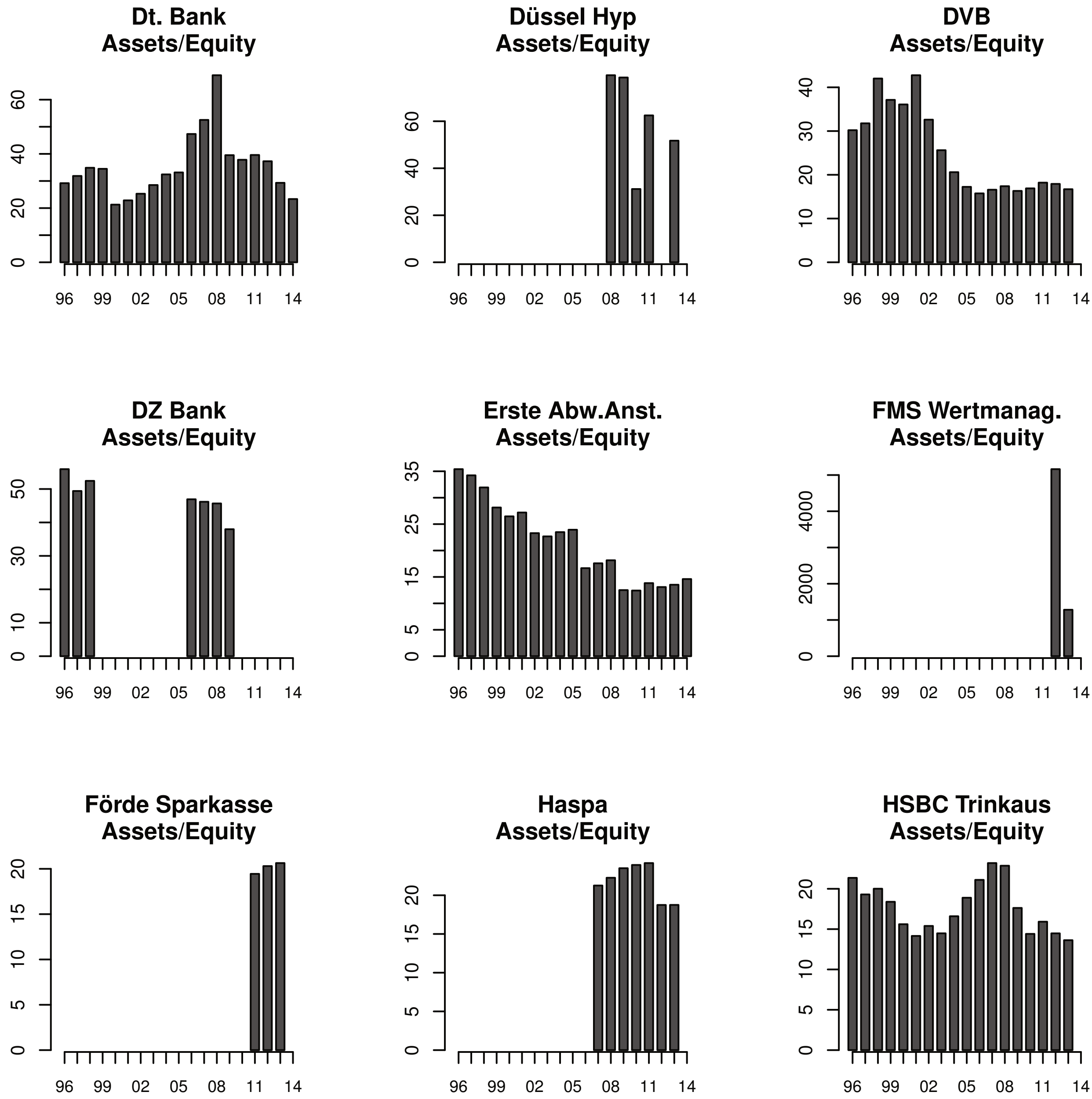
Appendix C

Appendix - Financial Information - Longitudinal Plausibility Check

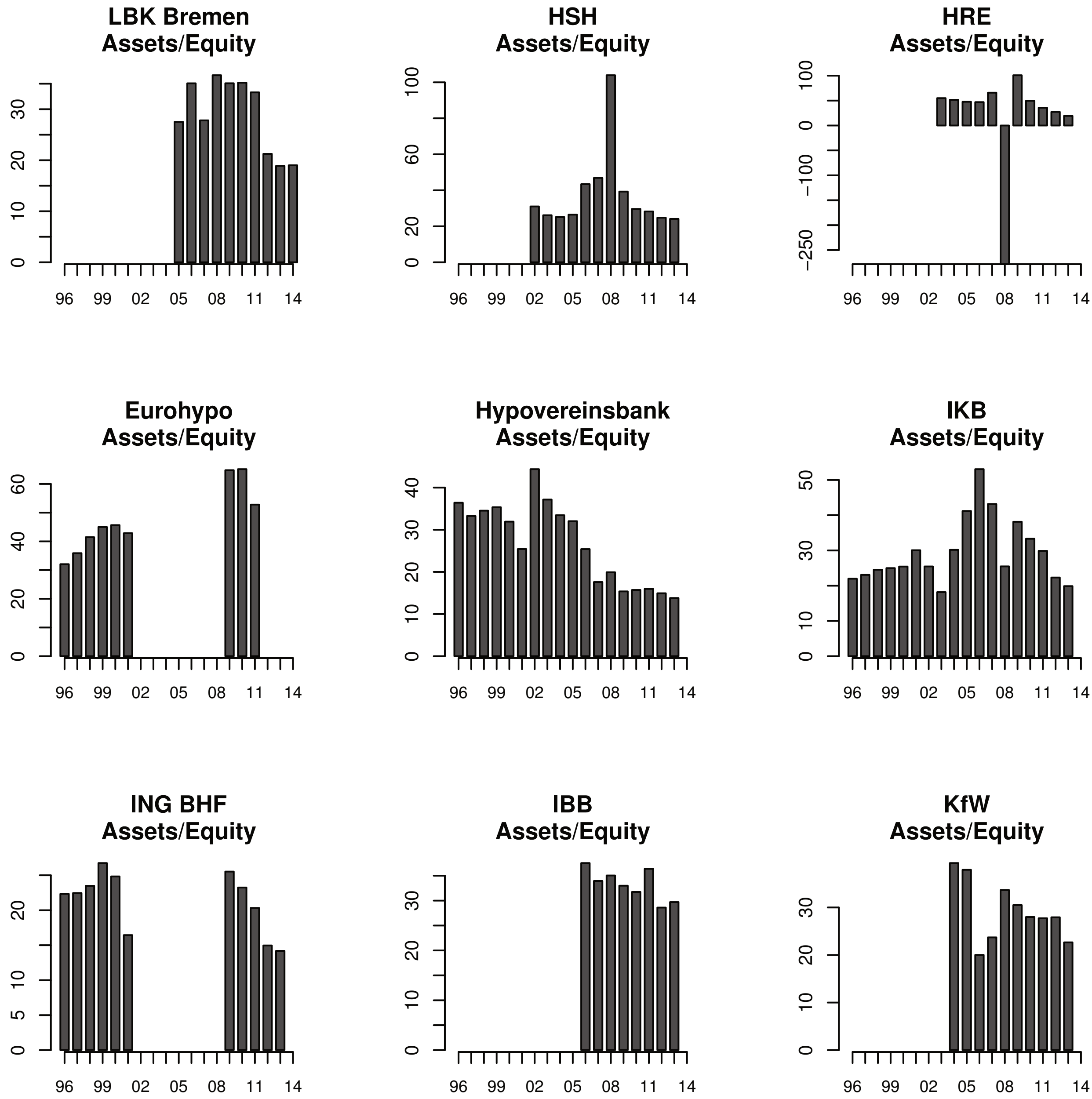
Assets/Equity



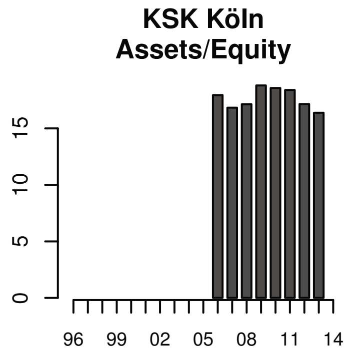
Assets/Equity

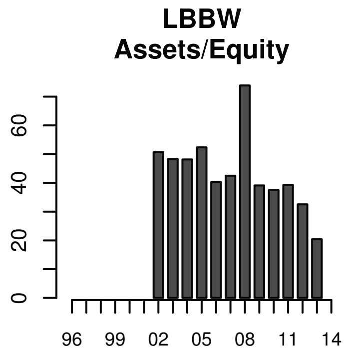


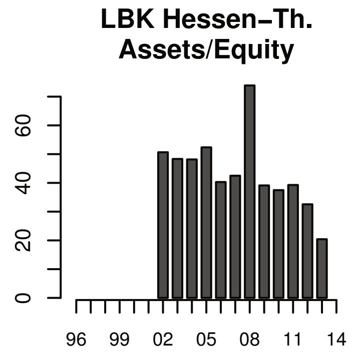
Assets/Equity

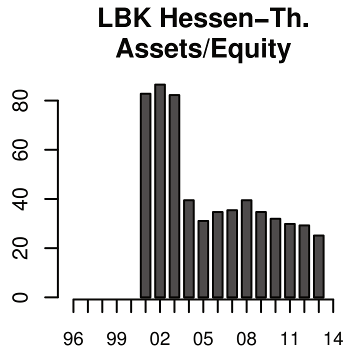


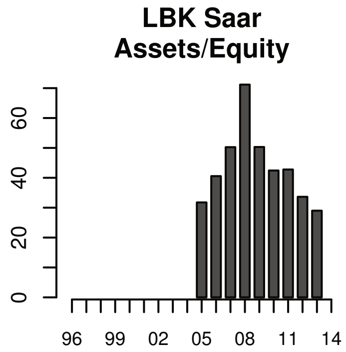
Assets/Equity

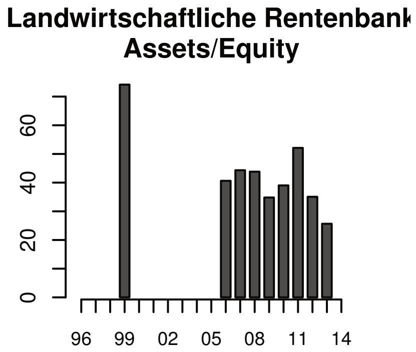


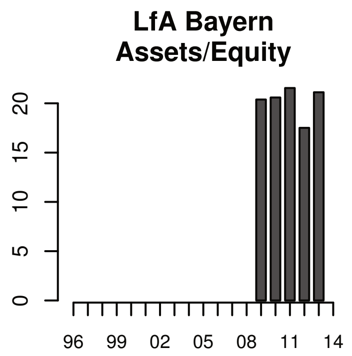


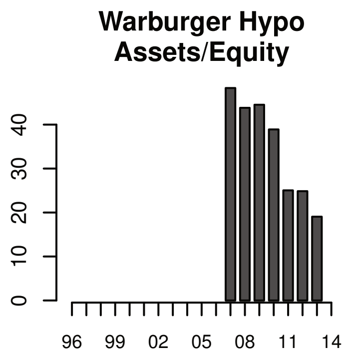


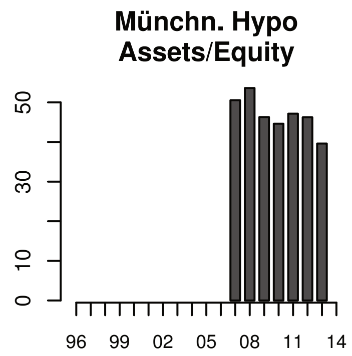




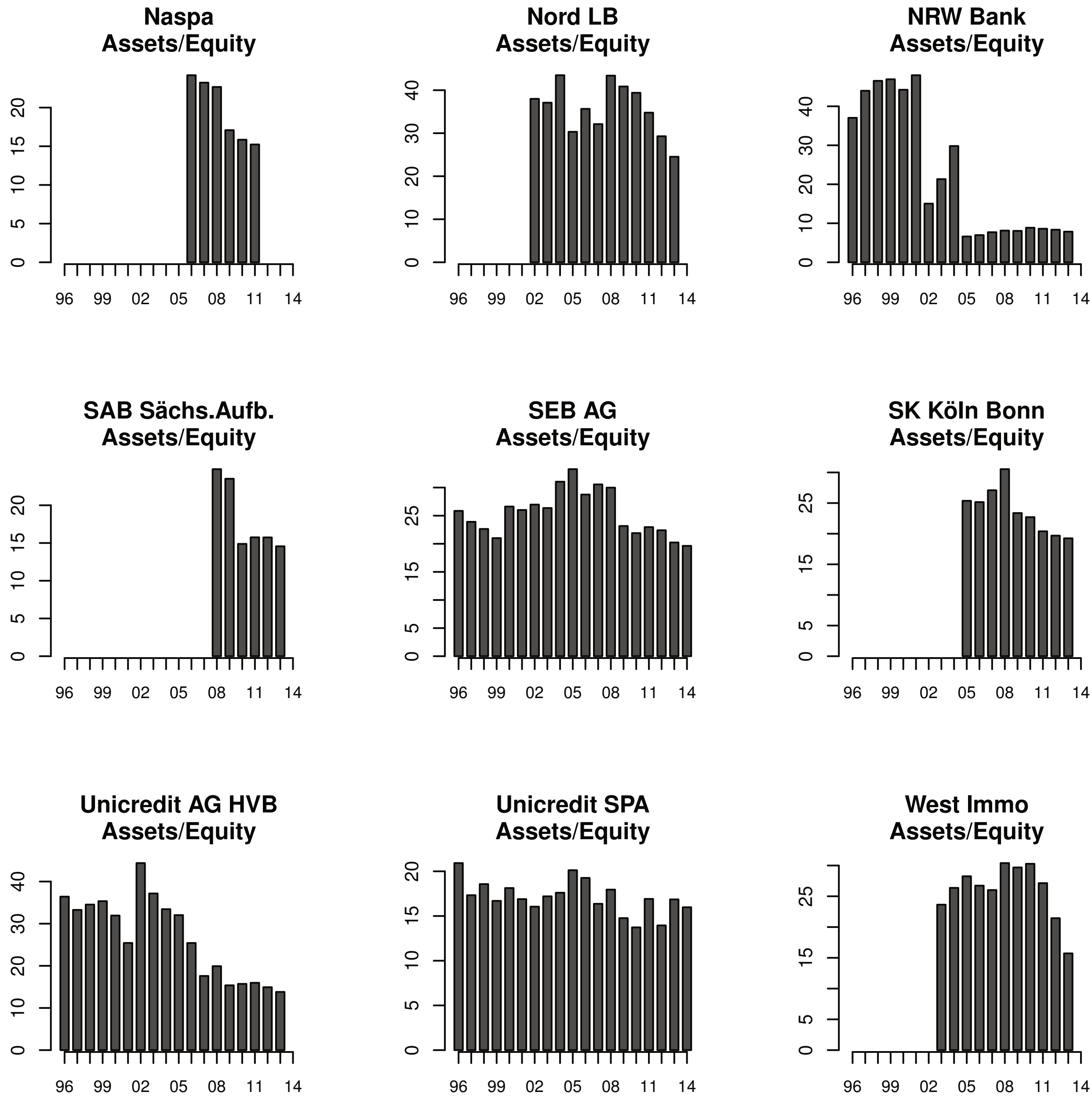




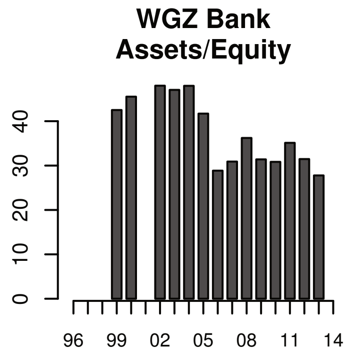


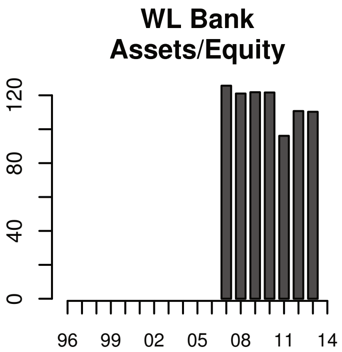


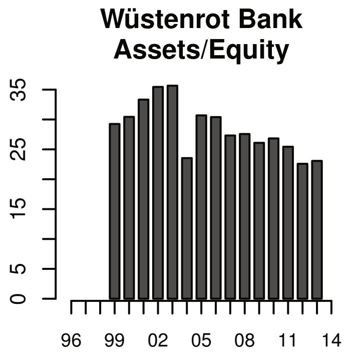
Assets/Equity



Assets/Equity







Appendix D

Appendix - Regression with Correlated Predictors Included

Variables	Coef	P-value	Signif. level
(Intercept_)	$-1.672e+01$	$7.04e-05$	***
Asset.Turnover_lag1	$1.124e+01$	0.155344	
Return.on.Assets_lag1	$-1.764e+00$	0.041144	*
Return.on.Common.Equity_lag1	$7.487e-02$	0.070455	.
Operating.ROE_lag1	$-5.639e-02$	0.107481	
Number.of.Employees_lag1	$3.581e-05$	$1.51e-05$	***
Net.Income.Per.1000.Employees_lag1	$3.728e-03$	$1.13e-05$	***
Actual.Sales.Per.Employee_lag1	$-1.053e-06$	$1.84e-05$	***
Assets.Per.1000.Employees_lag1	$4.491e-05$	$1.14e-06$	***
Total.Debt.Total.Capital_lag1	$1.611e-01$	0.002276	**
Assets.Equity_lag1	$-1.893e-01$	$9.30e-06$	***
LT.Debt.Common.Equity_lag1	$3.351e-03$	$5.78e-08$	***
LT.Debt.Total.Capital_lag1	$4.764e-02$	0.011593	*
LT.Debt.Total.Assets_lag1	$-1.397e-01$	0.000226	***
Total.Debt.Common.Equity_lag1	$-2.527e-03$	0.013119	*
Total.Debt.Total.Equity_lag1	$2.622e-03$	0.020747	*
Sales.Net.Fixed.Assets_lag1	$-2.067e-03$	0.000477	***
Tangible.Common.Equity_lag1	$-7.676e-05$	0.002480	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.445 on 194 degrees of freedom

Multiple R-squared: 0.3876, Adjusted R-squared: 0.3339

F-statistic: 7.223 on 17 and 194 DF, p-value: 1.592e-13

10-fold CV: 2.42, Number of Observations: 212

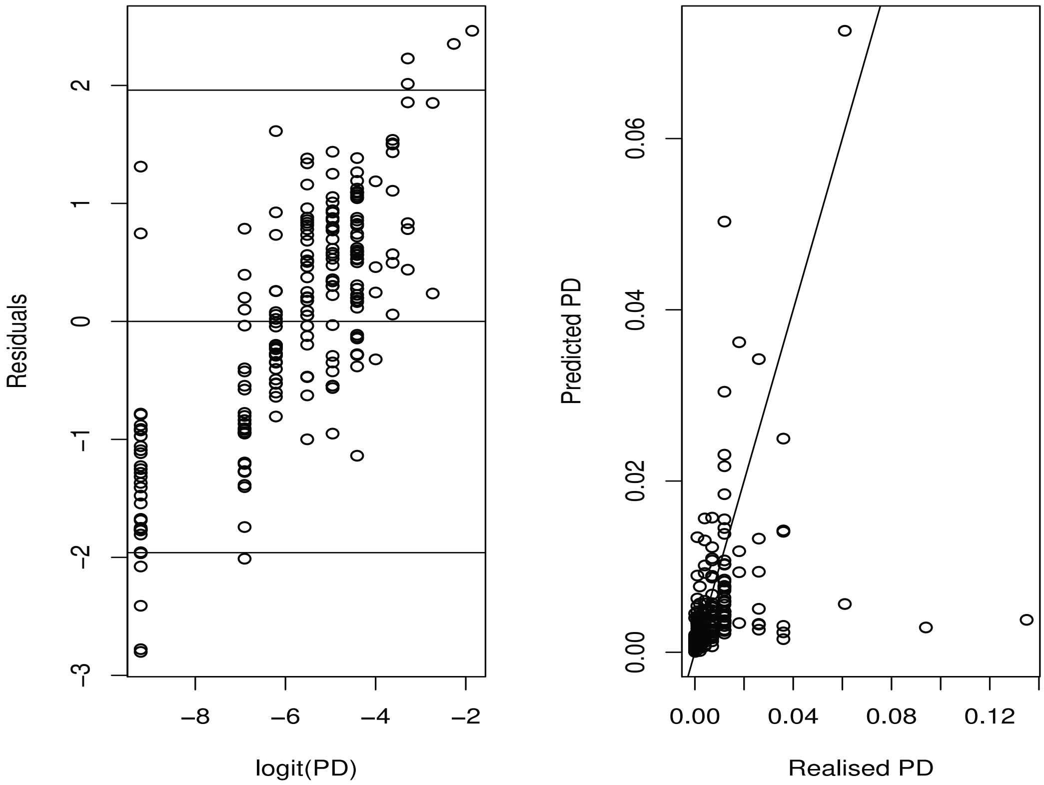


FIGURE D.1: Residuals and predictions plots for 1st step parameter estimation for correlated variables with less than 10% missing values

Variables	Coef	P-value	Signif. level
(Intercept)	$-1.519e+01$	$5.58e-05$	***
Return.on.Assets_lag1	$-1.050e+00$	0.170536	
Return.on.Common.Equity_lag1	$5.234e-02$	0.145701	
Operating.ROE_lag1	$-5.705e-02$	0.075291	.
Number.of.Employees_lag1	$3.793e-05$	$6.38e-07$	***
Net.Income.Per.1000.Employees_lag1	$3.528e-03$	$1.04e-05$	***
Actual.Sales.Per.Employee_lag1	$-7.922e-07$	$4.76e-06$	***
Assets.Per.1000.Employees_lag1	$3.462e-05$	$4.96e-08$	***
Total.Debt.Total.Capital_lag1	$1.420e-01$	0.002149	**
Assets.Equity_lag1	$-1.379e-01$	$1.03e-06$	***
LT.Debt.Common.Equity_lag1	$2.895e-03$	$6.81e-08$	***
LT.Debt.Total.Capital_lag1	$3.082e-02$	0.046695	*
LT.Debt.Total.Assets_lag1	$-9.923e-02$	0.000476	***
Total.Debt.Common.Equity_lag1	$-1.430e-03$	0.004297	**
Total.Debt.Total.Equity_lag1	$1.111e-03$	0.040093	*
Sales.Net.Fixed.Assets_lag1	$-2.075e-03$	0.000218	***
Tangible.Common.Equity_lag1	$-8.148e-05$	0.000399	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.414 on 221 degrees of freedom

Multiple R-squared: 0.3674, Adjusted R-squared: 0.3216

F-statistic: 8.021 on 16 and 221 DF, p-value: 5.507e-15

10-fold-CV: 2.19, Number of Observations: 238

Table D.2: 2nd step parameter estimation for correlated variables with less than 10% missing values

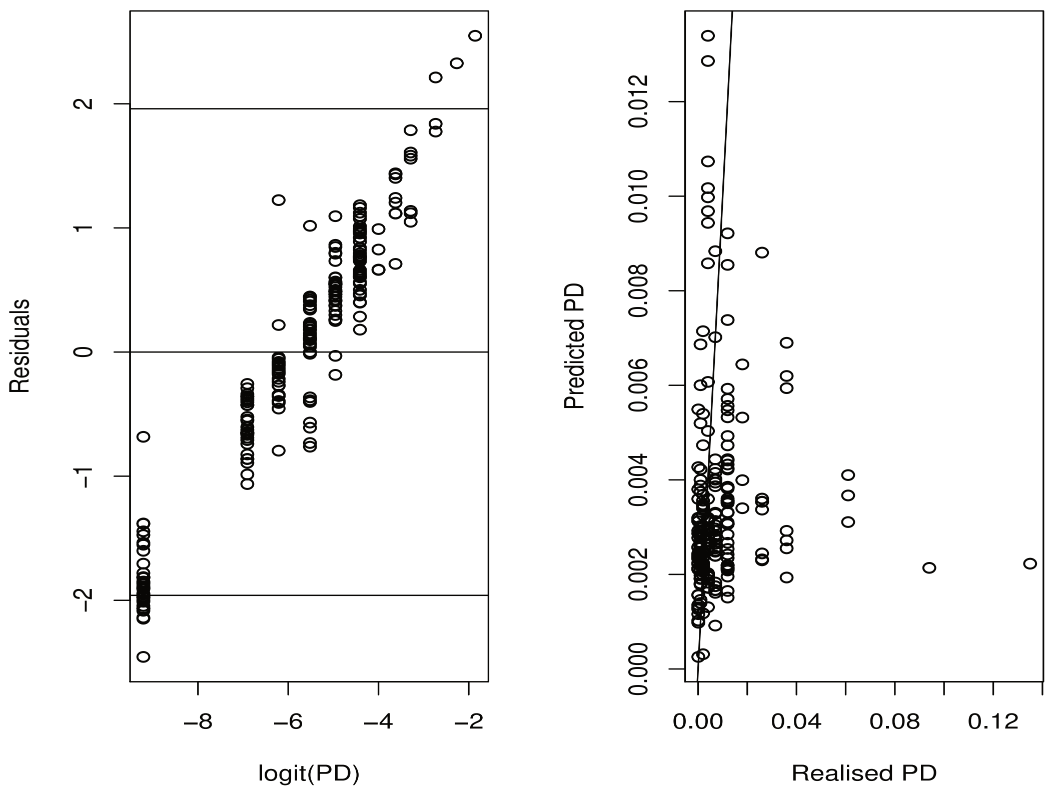


FIGURE D.2: Residuals and predictions plots for 2nd step parameter estimation for correlated variables with less than 10% missing values

Variables	Coef	P-value	Signif. level
(Intercept)	$-1.540e+01$	$9.20e-07$	***
Return.on.Assets_lag1	$-1.153e+00$	0.066244	.
Return.on.Common.Equity_lag1	$4.789e-02$	0.107977	
Operating.ROE_lag1	$-4.771e-02$	0.069400	.
Number.of.Employees_lag1	$3.039e-05$	$1.30e-06$	***
Net.Income.Per.1000.Employees_lag1	$3.212e-03$	$1.37e-06$	***
Actual.Sales.Per.Employee_lag1	$-1.000e-06$	$2.02e-11$	***
Assets.Per.1000.Employees_lag1	$3.646e-05$	$1.08e-11$	***
Total.Debt.Total.Capital_lag1	$1.412e-01$	0.000233	***
Assets.Equity_lag1	$-1.206e-01$	$2.62e-07$	***
LT.Debt.Common.Equity_lag1	$2.680e-03$	$1.51e-09$	***
LT.Debt.Total.Capital_lag1	$2.386e-02$	0.061119	.
LT.Debt.Total.Assets_lag1	$-7.707e-02$	0.001205	**
Total.Debt.Common.Equity_lag1	$-1.197e-03$	0.003791	**
Total.Debt.Total.Equity_lag1	$8.653e-04$	0.050460	.
Sales.Net.Fixed.Assets_lag1	$-2.091e-03$	$6.22e-06$	***
Tangible.Common.Equity_lag1	$-6.070e-05$	0.001446	**

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.137 on 203 degrees of freedom

Multiple R-squared: 0.4862, Adjusted R-squared: 0.4457

F-statistic: 12.01 on 16 and 203 DF, p-value: ¡ 2.2e-16

10-fold-CV: 1.56, Number of Observations: 220

Table D.3: 3rd step parameter estimation for correlated variables with less than 10% missing values

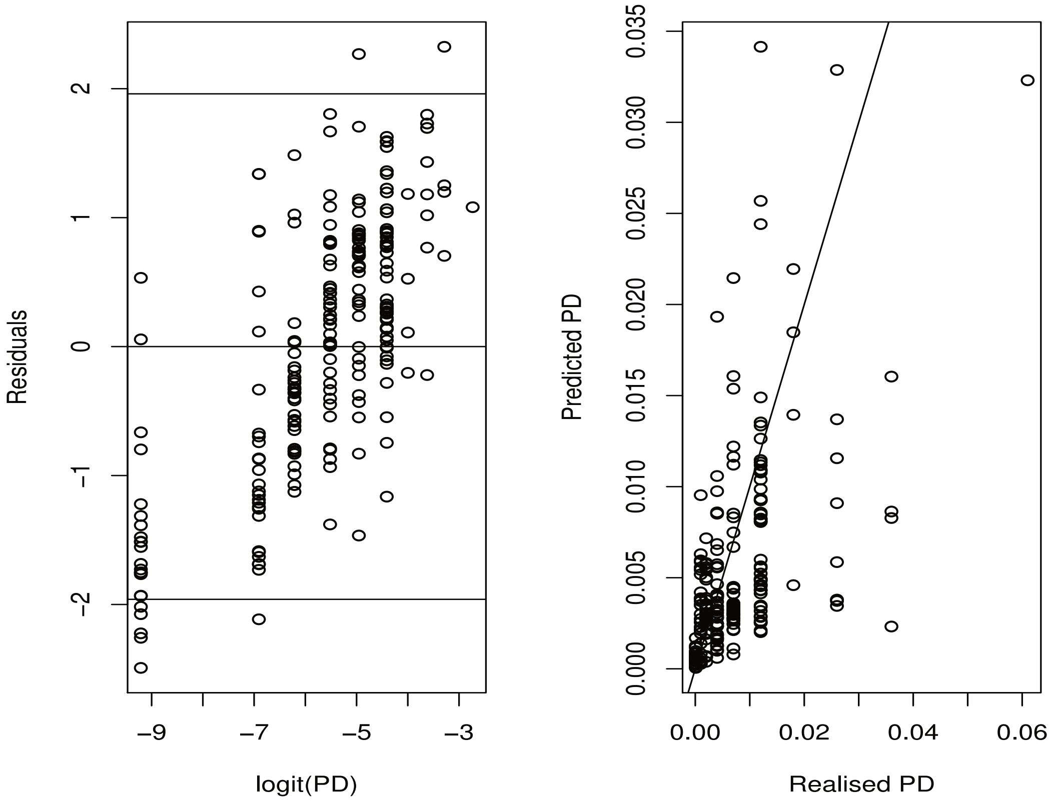


FIGURE D.3: Residuals and predictions plots for 3rd step parameter estimation for correlated variables with less than 10% missing values

Variables	Coef	P-value	Signif. level
(Intercept)	$-7.462e+00$	$3.03e-11$	***
Return.on.Assets_lag1	$-2.963e+00$	0.001059	**
Return.on.Common.Equity_lag1	$1.207e-01$	0.003214	**
Operating.ROE_lag1	$-7.256e-02$	0.023458	*
Number.of.Employees_lag1	$4.310e-05$	$3.25e-06$	***
Employees...1.Yr.Growth_lag1	$-8.579e-03$	0.070472	.
Net.Income.Per.1000.Employees_lag1	$2.719e-03$	0.088629	.
Actual.Sales.Per.Employee_lag1	$-6.773e-07$	0.000632	***
Assets.Per.1000.Employees_lag1	$3.912e-05$	$3.56e-08$	***
Tangible.Common.Equity.Ratio_lag1	$5.891e-01$	0.000220	***
LT.Debt.Common.Equity_lag1	$3.993e-03$	$7.18e-07$	***
LT.Debt.Total.Assets_lag1	$-1.005e-01$	0.001975	**
Total.Debt.Common.Equity_lag1	$-3.116e-03$	$6.62e-07$	***
Total.Debt.Tangible.Book.Value_lag1	$6.816e-02$	0.002362	**
Total.Debt.Total.Assets_lag1	$8.767e-02$	0.000365	***
Net.Change.in.Liabilities...of.Total_lag1	$1.965e-04$	0.011931	*
Sales.Marketable.Securities_lag1	$1.283e-01$	0.108465	
Sales.Net.Fixed.Assets_lag1	$-2.110e-03$	0.000903	***
Sales.LT.Investments_lag1	$-1.496e-04$	0.000393	***
Sales.Other.Assets_lag1	$2.555e-02$	0.083095	.
Tangible.Common.Equity_lag1	$-5.517e-05$	0.073196	.
int.inc....int...non.int.inc._lag1	$-2.367e+00$	0.038691	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.287 on 151 degrees of freedom

Multiple R-squared: 0.4732, Adjusted R-squared: 0.4

F-statistic: 6.459 on 21 and 151 DF, p-value: 1.049e-12

10-fold-CV: 2.08, Number of Observations: 173

Table D.4: 1st step parameter estimation for correlated variables with less than 15% missing values

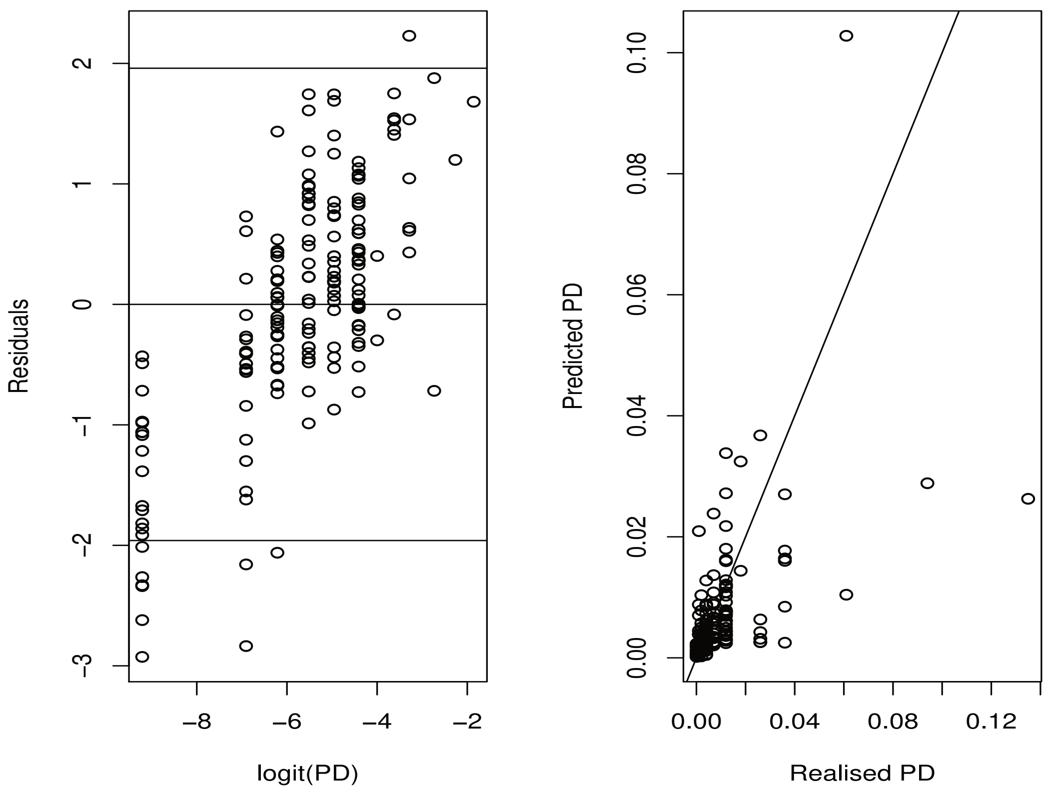


FIGURE D.4: Residuals and predictions plots for 1st step parameter estimation for variables with less than 15% missing values

Variables	Coef	P-value	Signif. level
(Intercept)	$-8.365e+00$	$< 2e-16$	***
Return.on.Assets_lag1	$-1.851e+00$	0.036478	*
Return.on.Common.Equity_lag1	$8.423e-02$	0.041007	*
Operating.ROE_lag1	$-5.912e-02$	0.069496	.
Number.of.Employees_lag1	$2.989e-05$	$1.17e-06$	***
Employees...1.Yr.Growth_lag1	$-5.837e-03$	0.179137	
Net.Income.Per.1000.Employees_lag1	$2.680e-03$	0.095385	.
Actual.Sales.Per.Employee_lag1	$-6.041e-07$	0.002659	**
Assets.Per.1000.Employees_lag1	$3.398e-05$	$2.59e-06$	***
Tangible.Common.Equity.Ratio_lag1	$3.747e-01$	0.006704	**
LT.Debt.Common.Equity_lag1	$3.597e-03$	$3.33e-06$	***
LT.Debt.Total.Assets_lag1	$-9.106e-02$	0.002181	**
Total.Debt.Common.Equity_lag1	$-2.958e-03$	$8.22e-07$	***
Total.Debt.Tangible.Book.Value_lag1	$6.961e-02$	0.001411	**
Total.Debt.Total.Assets_lag1	$7.423e-02$	0.000965	***
Net.Change.in.Liabilities...of.Total_lag1	$1.900e-04$	0.017867	*
Sales.Marketable.Securities_lag1	$1.330e-01$	0.109144	
Sales.Net.Fixed.Assets_lag1	$-1.748e-03$	0.008121	**
Sales.LT.Investments_lag1	$-1.149e-04$	0.003823	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.378 on 168 degrees of freedom

Multiple R-squared: 0.3701, Adjusted R-squared: 0.3027

F-statistic: 5.485 on 18 and 168 DF, p-value: 5.317e-10

10-fold-CV: 2.16, Number of Observations: 187

Table D.5: 2nd step parameter estimation for correlated variables with less than 15% missing values

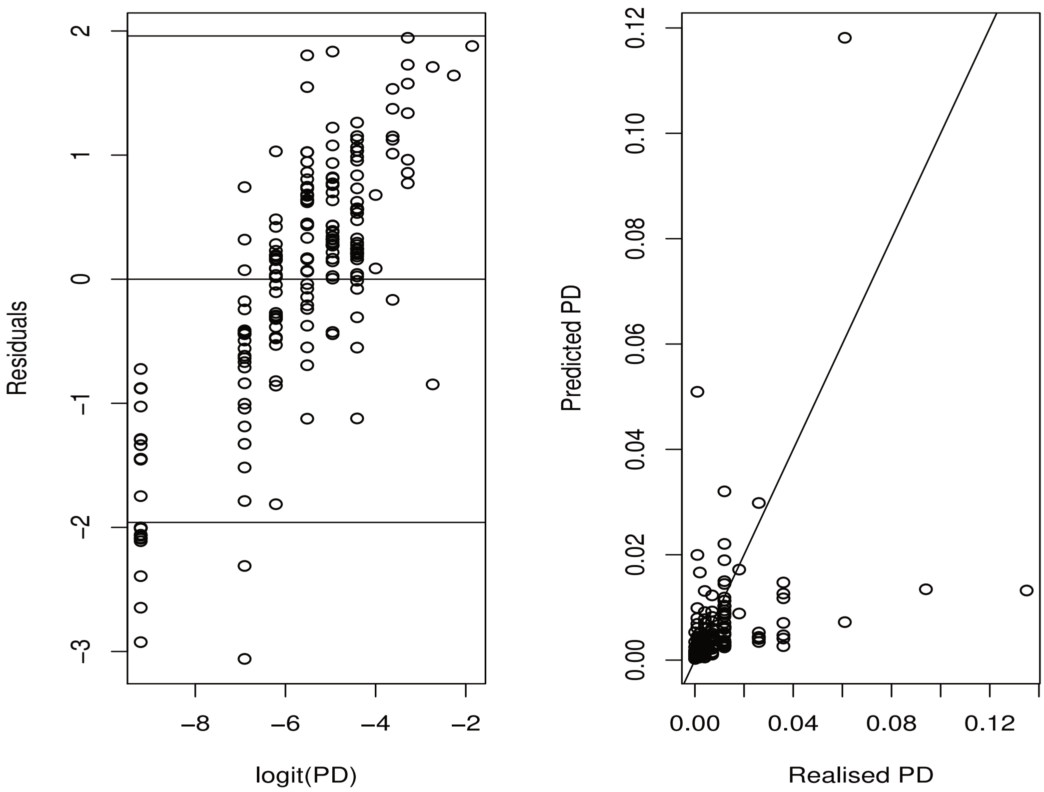


FIGURE D.5: Residuals and predictions plots for 2nd step parameter estimation for correlated variables with less than 15% missing values

Variables	Coef	P-value	Signif. level
(Intercept)	$-8.509e+00$	$< 2e-16$	***
Return.on.Assets_lag1	$-1.938e+00$	0.004873	**
Return.on.Common.Equity_lag1	$7.892e-02$	0.016682	*
Operating.ROE_lag1	$-6.566e-02$	0.009922	**
Number.of.Employees_lag1	$3.609e-05$	$2.42e-07$	***
Employees...1.Yr.Growth_lag1	$-8.835e-03$	0.008659	**
Net.Income.Per.1000.Employees_lag1	$3.071e-03$	0.015757	*
Actual.Sales.Per.Employee_lag1	$-8.274e-07$	$2.06e-07$	***
Assets.Per.1000.Employees_lag1	$3.777e-05$	$7.69e-11$	***
Tangible.Common.Equity.Ratio_lag1	$4.040e-01$	0.000208	***
LT.Debt.Common.Equity_lag1	$2.864e-03$	$4.97e-06$	***
LT.Debt.Total.Assets_lag1	$-4.905e-02$	0.050403	.
Total.Debt.Common.Equity_lag1	$-2.385e-03$	$1.63e-06$	***
Total.Debt.Tangible.Book.Value_lag1	$5.559e-02$	0.001207	**
Total.Debt.Total.Assets_lag1	$6.450e-02$	0.000831	***
Net.Change.in.Liabilities...of.Total_lag1	$1.854e-04$	0.002321	**
Sales.Net.Fixed.Assets_lag1	$-2.235e-03$	$1.49e-05$	***
Sales.LT.Investments_lag1	$-9.162e-05$	0.002880	**
Tangible.Common.Equity_lag1	$-2.906e-05$	0.168942	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.036 on 154 degrees of freedom

Multiple R-squared: 0.5337, Adjusted R-squared: 0.4792

F-statistic: 9.791 on 18 and 154 DF, p-value: ¡ 2.2e-16

10-fold-CV: 1.26, Number of Observations: 173

Table D.6: 3rd step parameter estimation for correlated variables with less than 15% missing values

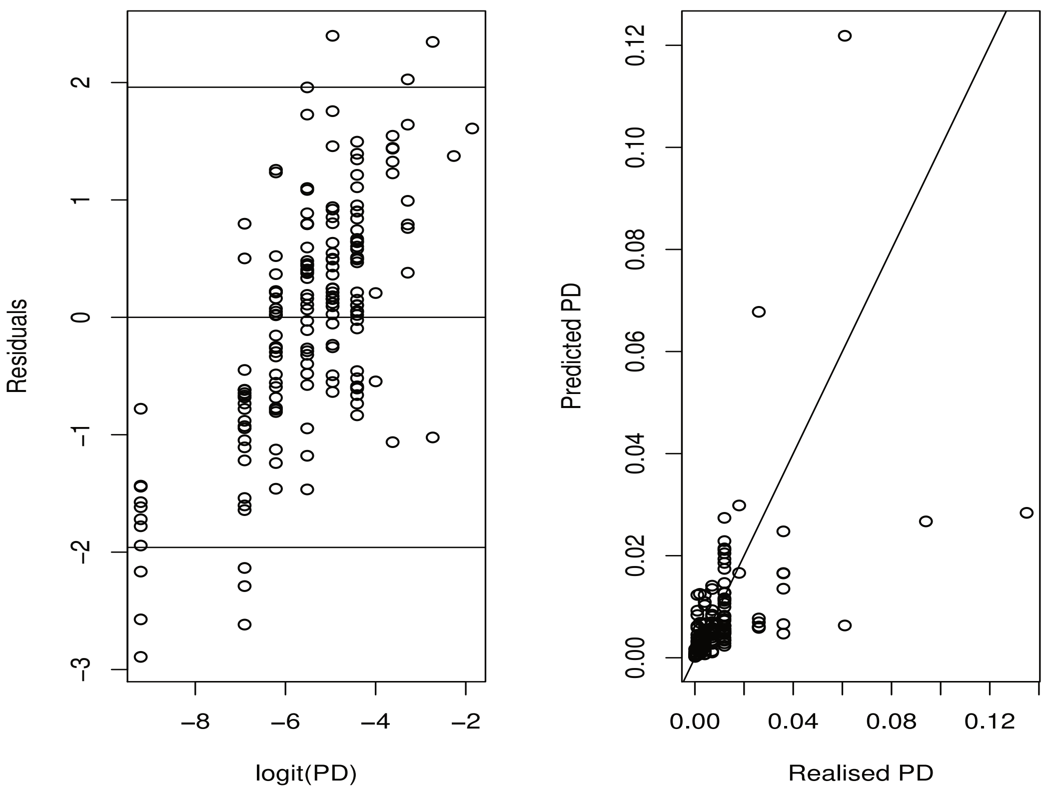


FIGURE D.6: Residuals and predictions plots for 3rd step parameter estimation for correlated variables with less than 15% missing values

Variables	Coef	P-value	Signif. level
(Intercept)	$-8.811e+00$	$< 2e-16$	***
Sales...1.Yr.Growth_lag1	$1.595e-02$	0.002322	**
Net.Income...1.Yr.Growth_lag1	$1.629e-03$	0.032486	*
Operating.Margin_lag1	$-5.315e-02$	$3.93e-07$	***
Pretax.Margin_lag1	$1.220e-01$	0.012306	*
Profit.Margin_lag1	$-1.855e-02$	0.122994	
Return.on.Assets_lag1	$-5.366e+00$	$3.18e-05$	***
Return.on.Common.Equity_lag1	$1.815e-01$	0.000816	***
Operating.ROE_lag1	$-1.064e-01$	0.042849	*
Number.of.Employees_lag1	$5.369e-05$	$4.49e-09$	***
Employees...1.Yr.Growth_lag1	$-1.628e-02$	0.001316	**
Net.Income.Per.1000.Employees_lag1	$9.020e-03$	$2.69e-06$	***
Actual.Sales.Per.Employee_lag1	$-8.129e-07$	0.000782	***
Assets.Per.1000.Employees_lag1	$4.516e-05$	$8.93e-08$	***
Tangible.Common.Equity.Ratio_lag1	$5.294e-01$	0.001933	**
LT.Debt.Common.Equity_lag1	$4.667e-03$	$2.64e-09$	***
LT.Debt.Total.Assets_lag1	$-1.280e-01$	$5.86e-05$	***
Total.Debt.Common.Equity_lag1	$-3.737e-03$	$5.06e-10$	***
Total.Debt.Tangible.Book.Value_lag1	$6.525e-02$	0.002566	**
Total.Debt.Total.Assets_lag1	$1.130e-01$	$5.70e-06$	***
Sales.Net.Fixed.Assets_lag1	$-1.364e-03$	0.077484	.
Sales.LT.Investments_lag1	$-1.459e-04$	0.003242	**
Sales.Other.Assets_lag1	$4.412e-02$	0.005735	**
Tangible.Common.Equity_lag1	$-9.615e-05$	0.001126	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.055 on 120 degrees of freedom

Multiple R-squared: 0.6616, Adjusted R-squared: 0.5967

F-statistic: 10.2 on 23 and 120 DF, p-value: ¡ 2.2e-16

10-fold-CV: 1.47, Number of Observations: 144

TABLE D.7: 1st step parameter estimation for variables with less than 25% missing values

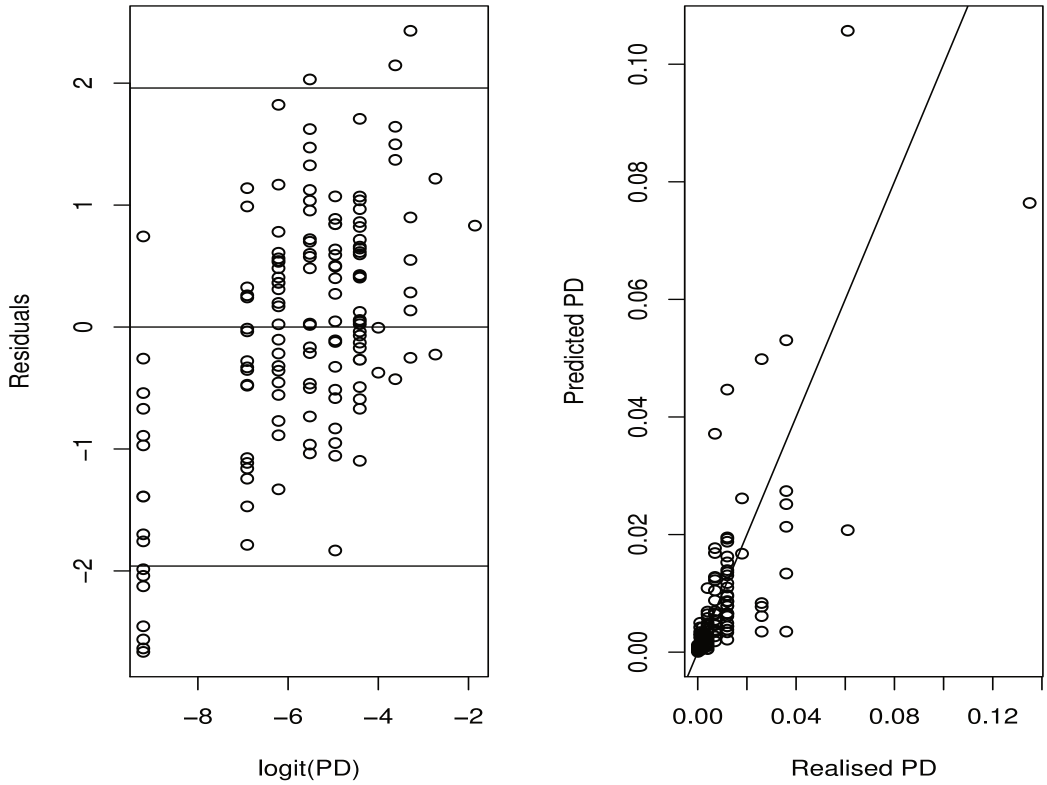


FIGURE D.7: Residuals and predictions plots for 1st step parameter estimation for correlated variables with less than 25% missing values

Variables	Coef	P-value	Signif. level
(Intercept)	$-4.542e+00$	$1.50e-10$	***
Net.Income...1.Yr.Growth_lag1	$1.599e-03$	0.078094	.
Operating.Margin_lag1	$-1.585e-02$	0.116423	
Pretax.Margin_lag1	$5.580e-02$	0.163748	
Profit.Margin_lag1	$-4.301e-02$	0.001172	**
Return.on.Common.Equity_lag1	$7.512e-02$	0.078969	.
Operating.ROE_lag1	$-1.371e-01$	0.010775	*
Number.of.Employees_lag1	$4.056e-05$	$9.25e-06$	***
Net.Income.Per.1000.Employees_lag1	$7.927e-03$	0.000116	***
Actual.Sales.Per.Employee_lag1	$-1.089e-06$	$5.75e-06$	***
Assets.Per.1000.Employees_lag1	$4.402e-05$	$1.84e-07$	***
Tangible.Common.Equity.Ratio_lag1	$-2.062e-01$	0.055392	.
LT.Debt.Common.Equity_lag1	$2.035e-03$	$3.84e-10$	***
Total.Debt.Common.Equity_lag1	$-1.751e-03$	$1.41e-09$	***
Total.Debt.Total.Assets_lag1	$2.528e-02$	0.013086	*
Sales.Net.Fixed.Assets_lag1	$-1.742e-03$	0.013077	*
Sales.Other.Assets_lag1	$3.325e-02$	0.093863	.
Tangible.Common.Equity_lag1	$-1.158e-04$	$1.87e-05$	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.378 on 153 degrees of freedom

Multiple R-squared: 0.4526, Adjusted R-squared: 0.3918

F-statistic: 7.441 on 17 and 153 DF, p-value: 4.05e-13

10-fold-CV: 2.36, Number of Observations: 171

TABLE D.8: 2nd step parameter estimation for correlated variables with less than 25% missing values

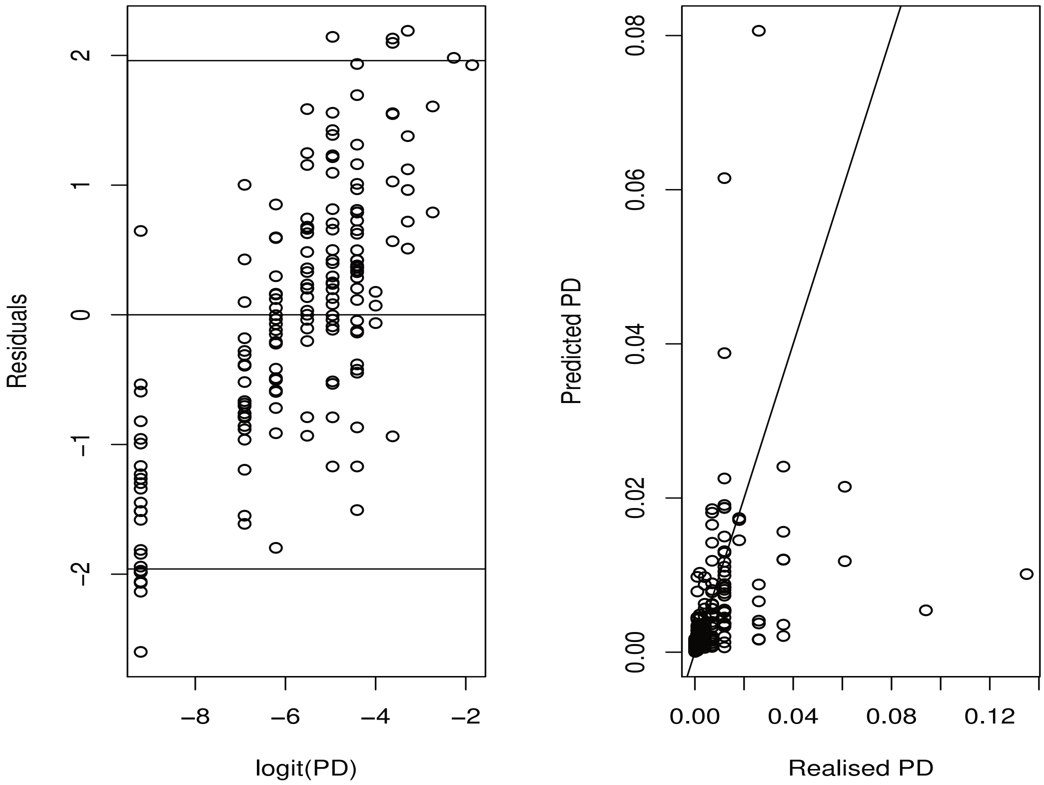


FIGURE D.8: Residuals and predictions plots for 2nd step parameter estimation for correlated variables with less than 25% missing values

Variables	Coef	P-value	Signif. level
(Intercept)	$-4.320e+00$	$1.29e-11$	***
Sales...1.Yr.Growth_lag1	$9.052e-03$	0.08477	.
Net.Income...1.Yr.Growth_lag1	$1.705e-03$	0.02752	*
Operating.Margin_lag1	$-1.559e-02$	0.07036	.
Pretax.Margin_lag1	$7.126e-02$	0.03442	*
Profit.Margin_lag1	$-4.424e-02$	$9.40e-05$	***
Return.on.Common.Equity_lag1	$6.176e-02$	0.09811	.
Operating.ROE_lag1	$-1.484e-01$	0.00117	**
Number.of.Employees_lag1	$3.613e-05$	$5.29e-06$	***
Employees...1.Yr.Growth_lag1	$-7.731e-03$	0.10069	
Net.Income.Per.1000.Employees_lag1	$8.715e-03$	$1.44e-06$	***
Actual.Sales.Per.Employee_lag1	$-1.149e-06$	$5.05e-08$	***
Assets.Per.1000.Employees_lag1	$4.190e-05$	$9.91e-09$	***
Tangible.Common.Equity.Ratio_lag1	$-2.437e-01$	0.00929	**
LT.Debt.Common.Equity_lag1	$1.995e-03$	$7.21e-13$	***
Total.Debt.Common.Equity_lag1	$-1.551e-03$	$3.12e-10$	***
Total.Debt.Total.Assets_lag1	$2.109e-02$	0.01548	*
Sales.Net.Fixed.Assets_lag1	$-2.015e-03$	0.00113	**
Sales.Other.Assets_lag1	$3.276e-02$	0.04498	*
Tangible.Common.Equity_lag1	$-1.043e-04$	$8.58e-06$	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.127 on 138 degrees of freedom

Multiple R-squared: 0.5659, Adjusted R-squared: 0.5061

F-statistic: 9.467 on 19 and 138 DF, p-value: ¡ 2.2e-16

10-fold-CV: 1.55, Number of Observations: 158

Table D.9: 3rd step parameter estimation for correlated variables with less than 25% missing values

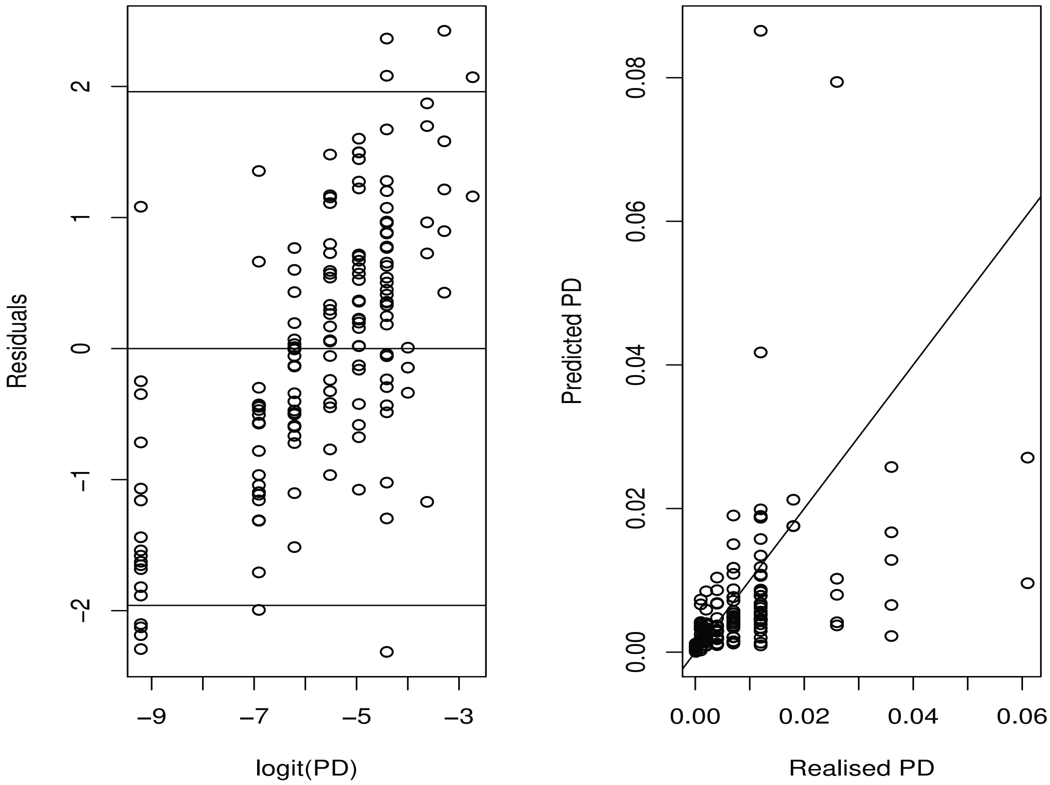


FIGURE D.9: Residuals and predictions plots for 3rd step parameter estimation for correlated variables with less than 25% missing values

Appendix E

Appendix - Principal Component Loadings ($< 25\%$ Missing)

Variables	PC1	PC2	PC3	PC4	PC5
Net.Income...1.Yr.Growth	0.06191	−0.083399	0.05350	−0.1045	0.042394
Assets...1.Yr.Growth	−0.00247	−0.183853	-0.37080	0.2436	−0.063447
Net.Worth...1.Yr.Growth	−0.00805	0.009502	-0.22487	0.1439	0.107342
Capital...1.Yr.Growth	−0.01791	−0.175359	-0.34627	0.2670	−0.057396
Net.Fixed.Asset.Turnover	−0.12371	−0.013057	0.26380	0.0957	0.021623
Operating.Margin	0.03106	−0.371377	0.07101	0.0031	0.000893
Pretax.Margin	0.15244	−0.307778	-0.01657	−0.0647	−0.034294
Profit.Margin	0.07909	−0.273369	0.12270	−0.1217	−0.057158
Return.on.Common.Equity	0.09688	−0.379844	0.00139	−0.0983	0.013049
Financial.Leverage	−0.25592	−0.000114	-0.14250	−0.1965	−0.000793
Annualized.Return.on.Common.Eq	0.09688	−0.379844	0.00139	−0.0983	0.013049
Operating.ROE	0.09140	−0.382691	0.01125	−0.0662	0.016184
Number.of.Employees	0.13511	0.022634	-0.31459	−0.1580	−0.014700
Employees...1.Yr.Growth	0.02237	−0.091371	-0.22796	0.3486	0.036969
Net.Income.Per.1000.Employees	0.01756	−0.347762	0.17276	0.0362	−0.004578
Actual.Sales.Per.Employee	−0.23667	−0.068249	0.14499	−0.0897	−0.046979
Assets.Per.1000.Employees	−0.22742	−0.060925	0.04820	−0.1214	−0.058676
Total.Debt.Total.Capital	−0.22974	−0.058414	-0.01244	0.1220	−0.018118
Assets.Equity	−0.26118	−0.031517	-0.12452	−0.1922	−0.034579
Tangible.Common.Equity.Ratio	0.16967	0.041816	0.25382	0.1132	−0.001602

Variables	PC1	PC2	PC3	PC4	PC5
LT.Debt.Common.Equity	−0.27983	−0.054045	-0.01277	−0.0514	0.034524
LT.Debt.Total.Capital	−0.16682	−0.041646	0.23409	0.0536	0.099861
LT.Debt.Total.Assets	−0.22212	−0.071390	0.16649	0.2003	0.099973
Total.Debt.Common.Equity	−0.27927	−0.047809	-0.10424	−0.0375	0.010712
Total.Debt.Tangible.Book.Value	−0.25748	−0.066117	-0.18406	−0.0127	−0.002674
Total.Debt.Total.Equity	−0.28124	−0.046740	-0.08751	−0.0364	0.006945
Total.Debt.Total.Assets	−0.21674	−0.071468	0.03283	0.2648	0.055300
Net.Debt.Shareholders.Equity	−0.27700	−0.031853	-0.06805	0.0108	0.042451
Net.Change.in.Liabilities.of.Total	−0.01887	0.016322	0.04105	0.0896	−0.678088
Increase.In.Equity...of.Total	0.01887	−0.016322	-0.04105	−0.0896	0.678088
Sales.Cash	−0.13785	−0.035072	-0.02597	−0.2471	−0.066471
Sales.Marketable.Securities	−0.00932	−0.030427	0.05269	0.1975	−0.023209
Sales.Net.Fixed.Assets	−0.11263	−0.014313	0.27657	0.1007	0.021100
Sales.LT.Investments	−0.14460	−0.061564	-0.03388	−0.3642	−0.106604
Sales.Other.Assets	−0.07082	−0.006571	0.08874	0.0746	−0.024786
Tangible.Common.Equity	0.10522	0.057352	-0.24750	−0.1727	−0.000696
Tot.loan.tot.dep	−0.09362	−0.082354	-0.01377	0.3147	0.043441

Variables	PC6	PC7	PC8	PC9	PC10
Net.Income...1.Yr.Growth	−0.2270	0.3523	-0.17882	0.28602	−0.43321
Assets...1.Yr.Growth	−0.1190	−0.2517	0.02245	−0.07090	−0.06850
Net.Worth...1.Yr.Growth	−0.3482	−0.0973	-0.01913	0.18162	−0.51708
Capital...1.Yr.Growth	−0.0589	−0.2807	0.06697	−0.10388	−0.03124
Net.Fixed.Asset.Turnover	−0.4752	−0.0362	0.01208	−0.36545	0.04270
Operating.Margin	0.0266	−0.0480	0.05592	−0.00753	0.02308
Pretax.Margin	0.0222	−0.0502	0.02567	−0.04252	0.03074
Profit.Margin	−0.0864	0.0635	-0.09586	−0.05703	−0.29865
Return.on.Common.Equity	0.0228	0.1199	0.03594	−0.00362	0.02544
Financial.Leverage	−0.0328	0.0422	-0.02386	−0.07666	−0.08814
Annualized.Return.on.Common.Eq	0.0228	0.1199	0.03594	−0.00362	0.02544
Operating.ROE	0.0767	0.0512	0.07815	−0.06302	0.10542
Number.of.Employees	−0.2576	0.2439	-0.10972	−0.04129	0.26551
Employees...1.Yr.Growth	−0.0810	−0.0450	-0.34607	0.08413	0.09328
Net.Income.Per.1000.Employees	−0.0249	−0.0741	0.04443	0.09082	0.07127
Actual.Sales.Per.Employee	−0.1037	−0.1577	-0.13201	0.09421	0.07722
Assets.Per.1000.Employees	−0.0528	−0.2359	-0.09990	0.12854	0.06819
Total.Debt.Total.Capital	0.1099	0.2008	0.23014	0.00759	−0.09828
Assets.Equity	0.0271	0.0196	-0.02973	−0.09776	0.00986
Tangible.Common.Equity.Ratio	0.0210	−0.3223	-0.13294	0.30586	−0.06786

Variables	PC6	PC7	PC8	PC9	PC10
LT.Debt.Common.Equity	−0.0171	0.0341	-0.03771	0.06025	0.04444
LT.Debt.Total.Capital	−0.1243	0.0831	-0.00889	0.25420	0.11088
LT.Debt.Total.Assets	−0.0189	0.1043	0.06464	0.26693	0.06971
Total.Debt.Common.Equity	0.0607	0.0576	0.03815	−0.02504	−0.01160
Total.Debt.Tangible.Book.Value	0.0439	0.0330	0.07186	−0.08439	−0.01645
Total.Debt.Total.Equity	0.0565	0.0475	0.03072	−0.01717	−0.01492
Total.Debt.Total.Assets	0.1279	0.1063	0.17360	0.17899	−0.06038
Net.Debt.Shareholders.Equity	0.0290	0.0629	0.10307	0.02556	−0.01624
Net.Change.in.Liabilities.of.Total	−0.0458	0.0895	0.04651	0.07078	−0.01122
Increase.In.Equity...of.Total	0.0458	−0.0895	-0.04651	−0.07078	0.01122
Sales.Cash	−0.0302	−0.1792	-0.29527	0.08216	−0.02636
Sales.Marketable.Securities	0.2655	0.1355	-0.50188	−0.25918	0.03167
Sales.Net.Fixed.Assets	−0.4729	−0.0197	0.01870	−0.36354	0.08667
Sales.LT.Investments	0.0133	−0.2861	-0.29630	0.08685	0.01006
Sales.Other.Assets	0.2078	0.2252	-0.29113	−0.33245	−0.36965
Tangible.Common.Equity	−0.2844	0.3293	-0.02319	0.17952	0.21493
Tot.loan.tot.dep	−0.0349	0.2002	-0.36979	0.16916	0.31828

Variables	PC11	PC12	PC13	PC14	PC15
Net.Income...1.Yr.Growth	0.38905	−0.01393	-0.23669	0.31680	−0.24623
Assets...1.Yr.Growth	−0.06093	0.17471	0.02300	0.17594	−0.13855
Net.Worth...1.Yr.Growth	−0.04511	−0.06909	0.27937	−0.43379	0.33756
Capital...1.Yr.Growth	−0.09714	0.14998	0.02734	0.19190	−0.13568
Net.Fixed.Asset.Turnover	0.07465	−0.06587	-0.00565	0.02761	−0.07840
Operating.Margin	0.01081	0.11846	-0.10334	−0.23010	−0.27093
Pretax.Margin	−0.01683	−0.19427	0.02679	−0.04494	0.14965
Profit.Margin	−0.06595	0.10349	0.26384	0.21242	0.10385
Return.on.Common.Equity	0.01312	−0.08144	0.01418	0.01827	0.12898
Financial.Leverage	0.02543	−0.07934	0.03873	−0.04660	0.12065
Annualized.Return.on.Common.Eq	0.01312	−0.08144	0.01418	0.01827	0.12898
Operating.ROE	−0.02213	−0.11810	-0.02598	0.03416	0.10252
Number.of.Employees	−0.07389	0.12131	0.01127	−0.16238	−0.08349
Employees...1.Yr.Growth	−0.03556	−0.24640	-0.26604	0.28270	−0.02916
Net.Income.Per.1000.Employees	−0.02072	0.18928	-0.04797	−0.31525	−0.16510
Actual.Sales.Per.Employee	0.12361	0.30326	-0.08412	0.01316	0.09481
Assets.Per.1000.Employees	0.11501	0.36191	-0.10993	0.10982	0.19808
Total.Debt.Total.Capital	0.04306	0.08417	0.08970	0.03643	−0.21923
Assets.Equity	0.00877	0.01737	0.00191	0.07150	0.00674
Tangible.Common.Equity.Ratio	0.01829	0.08590	-0.08400	−0.15025	−0.15537

Variables	PC11	PC12	PC13	PC14	PC15
LT.Debt.Common.Equity	−0.09275	−0.09599	0.03921	0.01224	0.01087
LT.Debt.Total.Capital	−0.49214	0.00537	0.26548	0.35856	0.09996
LT.Debt.Total.Assets	−0.20876	−0.01962	0.14256	0.04208	−0.04389
Total.Debt.Common.Equity	0.07460	−0.03747	-0.03041	−0.05387	−0.02183
Total.Debt.Tangible.Book.Value	0.10034	−0.08560	-0.04382	0.00566	0.04694
Total.Debt.Total.Equity	0.06972	−0.01450	-0.02655	−0.05375	−0.03351
Total.Debt.Total.Assets	0.11468	0.02918	0.01970	−0.18765	−0.18204
Net.Debt.Shareholders.Equity	0.06356	−0.04722	-0.08667	−0.06399	0.00801
Net.Change.in.Liabilities.of.Total	−0.01688	−0.05749	0.01487	−0.02154	0.04401
Increase.In.Equity.of.Total	0.01688	0.05749	-0.01487	0.02154	−0.04401
Sales.Cash	−0.23932	−0.49148	0.07885	−0.16538	−0.48059
Sales.Marketable.Securities	0.26270	0.15339	0.59466	−0.02343	−0.13693
Sales.Net.Fixed.Assets	0.06385	−0.07126	-0.04589	−0.06505	−0.05969
Sales.LT.Investments	0.02613	0.07432	0.01410	−0.00175	0.05908
Sales.Other.Assets	−0.52116	0.23934	-0.41401	−0.14660	0.04397
Tangible.Common.Equity	−0.17787	0.34875	0.08446	−0.11661	−0.20278
Tot.loan.tot.dep	0.12959	−0.09932	-0.16926	−0.21939	0.32933

Variables	PC16	PC17	PC18	PC19	PC20
Net.Income...1.Yr.Growth	−0.2909	0.02015	0.084473	−0.13538	−0.0304
Assets...1.Yr.Growth	−0.0657	0.11003	-0.016858	−0.07125	−0.1898
Net.Worth...1.Yr.Growth	−0.0707	−0.15415	-0.063630	−0.03547	0.0922
Capital...1.Yr.Growth	−0.1561	0.25205	0.000154	−0.15074	0.0554
Net.Fixed.Asset.Turnover	0.0129	0.02480	-0.055027	0.09154	0.0124
Operating.Margin	0.0509	−0.26593	0.528546	0.01663	−0.0153
Pretax.Margin	−0.3117	0.17829	0.392500	0.15378	0.2325
Profit.Margin	0.6578	0.22849	0.202307	0.11444	−0.1943
Return.on.Common.Equity	−0.0320	0.07423	-0.352222	0.10953	−0.0149
Financial.Leverage	−0.0122	−0.17867	-0.006111	0.02449	0.0219
Annualized.Return.on.Common.Eq	−0.0320	0.07423	-0.352222	0.10953	−0.0149
Operating.ROE	−0.1362	0.01030	-0.174686	−0.08189	0.2033
Number.of.Employees	−0.0216	0.09426	0.162508	0.18794	0.1900
Employees...1.Yr.Growth	0.2762	−0.44636	-0.097478	0.28462	0.1634
Net.Income.Per.1000.Employees	0.0201	−0.41232	-0.157062	−0.22237	−0.2742
Actual.Sales.Per.Employee	−0.0485	0.03316	-0.095996	−0.03731	0.0805
Assets.Per.1000.Employees	−0.0643	0.11144	-0.047688	−0.00487	−0.0613
Total.Debt.Total.Capital	0.2427	0.00267	-0.148729	−0.27842	0.5132
Assets.Equity	0.0221	−0.09593	-0.052674	0.04253	−0.0221
Tangible.Common.Equity.Ratio	−0.0340	0.19884	-0.162260	0.44110	0.0617

Variables	PC16	PC17	PC18	PC19	PC20
LT.Debt.Common.Equity	−0.0887	−0.00963	0.105345	0.19681	−0.1012
LT.Debt.Total.Capital	−0.1958	−0.15017	0.073766	−0.09984	−0.0114
LT.Debt.Total.Assets	−0.0972	0.06996	0.116511	0.12500	0.0603
Total.Debt.Common.Equity	0.0127	0.02884	0.012996	0.19375	−0.0746
Total.Debt.Tangible.Book.Value	−0.0625	−0.04424	0.091411	−0.00896	−0.2076
Total.Debt.Total.Equity	0.0147	0.04502	0.004216	0.19070	−0.0862
Total.Debt.Total.Assets	0.1270	0.21638	-0.006392	0.14534	0.2061
Net.Debt.Shareholders.Equity	−0.0572	0.05269	-0.048878	0.25034	−0.1709
Net.Change.in.Liabilities.of.Total	−0.0369	−0.04441	-0.003834	0.03274	−0.0140
Increase.In.Equity...of.Total	0.0369	0.04441	0.003834	−0.03274	0.0140
Sales.Cash	0.0313	0.23898	-0.132058	−0.21872	−0.1282
Sales.Marketable.Securities	−0.2275	−0.09849	-0.051820	0.09052	−0.0145
Sales.Net.Fixed.Assets	−0.0276	0.05723	-0.039081	−0.00254	0.1316
Sales.LT.Investments	0.0928	−0.08827	0.072525	−0.07940	0.4547
Sales.Other.Assets	−0.1143	0.07948	-0.042666	0.02592	0.0600
Tangible.Common.Equity	0.0475	0.03088	-0.158243	0.09733	−0.0457
Tot.loan.tot.dep	0.1569	0.30063	0.119600	−0.37159	−0.1123

Variables	PC21	PC22	PC23	PC24	PC25
Net.Income...1.Yr.Growth	−0.054060	0.07452	0.06742	−0.02068	−0.043924
Assets...1.Yr.Growth	−0.063883	0.15642	-0.18184	0.20550	0.092389
Net.Worth...1.Yr.Growth	0.085839	−0.06959	-0.07987	−0.07172	0.000257
Capital...1.Yr.Growth	−0.158606	0.00953	0.19212	−0.30055	−0.142505
Net.Fixed.Asset.Turnover	−0.022035	−0.04885	0.18264	−0.22870	0.622843
Operating.Margin	0.004780	−0.08660	-0.46784	−0.24370	0.063457
Pretax.Margin	−0.059595	−0.39430	0.24361	0.38446	0.095219
Profit.Margin	0.005135	−0.00523	0.15223	0.01917	−0.101142
Return.on.Common.Equity	0.049169	0.11760	-0.21014	−0.01075	0.077544
Financial.Leverage	−0.198924	−0.09743	0.11374	−0.03543	−0.084874
Annualized.Return.on.Common.Eq	0.049169	0.11760	-0.21014	−0.01075	0.077544
Operating.ROE	−0.013855	−0.04867	0.08435	−0.24520	−0.141976
Number.of.Employees	0.342447	0.31701	0.29840	−0.08227	−0.078848
Employees...1.Yr.Growth	0.138753	−0.16647	0.02884	0.07266	−0.041687
Net.Income.Per.1000.Employees	−0.054489	0.10441	0.51377	0.18357	−0.079776
Actual.Sales.Per.Employee	0.309707	−0.17720	-0.03721	−0.24248	−0.041931
Assets.Per.1000.Employees	0.354842	−0.27545	-0.02787	0.19912	−0.030929
Total.Debt.Total.Capital	−0.032875	−0.20261	0.00820	0.13953	0.140486
Assets.Equity	−0.288001	−0.13358	0.04547	−0.05976	−0.096811
Tangible.Common.Equity.Ratio	−0.314198	0.03769	0.03358	0.07000	0.065476

Variables	PC21	PC22	PC23	PC24	PC25
LT.Debt.Common.Equity	−0.067642	0.03599	-0.00300	−0.19860	−0.111609
LT.Debt.Total.Capital	−0.076552	0.05430	-0.00314	0.09358	0.060862
LT.Debt.Total.Assets	0.122939	0.18493	-0.02082	−0.08975	−0.048824
Total.Debt.Common.Equity	−0.080553	0.00789	0.05357	−0.00705	−0.073129
Total.Debt.Tangible.Book.Value	0.169469	0.26599	-0.08379	0.40960	0.300012
Total.Debt.Total.Equity	−0.107170	−0.05286	0.04032	−0.03427	−0.082698
Total.Debt.Total.Assets	0.168118	0.12372	0.07884	0.09314	−0.031388
Net.Debt.Shareholders.Equity	−0.124794	−0.01851	0.05769	−0.05357	0.039512
Net.Change.in.Liabilities.of.Total	0.000903	0.01326	-0.01443	−0.02661	−0.006878
Increase.In.Equity...of.Total	−0.000903	−0.01326	0.01443	0.02661	0.006878
Sales.Cash	0.221625	−0.16241	-0.03302	0.01577	−0.005568
Sales.Marketable.Securities	0.023589	−0.03623	-0.02824	0.01896	−0.007414
Sales.Net.Fixed.Assets	−0.097403	0.09809	-0.23987	0.31858	−0.552076
Sales.LT.Investments	−0.255422	0.44390	-0.05038	0.07221	0.151246
Sales.Other.Assets	0.029765	0.02239	0.00409	0.03567	0.039293
Tangible.Common.Equity	−0.250336	−0.30088	-0.17923	0.14056	0.102870
Tot.loan.tot.dep	−0.252353	0.00109	-0.02628	−0.04129	0.069996

Variables	PC26	PC27	PC28	PC29	PC30
LT.Debt.Common.Equity	−0.067642	0.03599	-0.00300	−0.19860	−0.111609
Net.Income...1.Yr.Growth	−0.004631	−0.00899	-2.56e-05	0.020503	$2.35e-02$
Assets...1.Yr.Growth	−0.315493	0.27212	-3.21e-01	−0.067549	$1.92e-01$
Net.Worth...1.Yr.Growth	0.025384	−0.02496	-4.68e-02	−0.000516	$3.35e-02$
Capital...1.Yr.Growth	0.220762	−0.18453	3.33e-01	0.086396	$-2.21e-01$
Net.Fixed.Asset.Turnover	0.017540	−0.03437	8.81e-03	0.112280	$1.39e-01$
Operating.Margin	−0.046272	−0.16706	-6.51e-02	0.073113	$-7.00e-02$
Pretax.Margin	−0.022623	0.22474	9.43e-02	0.013369	$-7.34e-02$
Profit.Margin	0.080218	−0.01107	-2.59e-02	−0.116005	$-2.11e-02$
Return.on.Common.Equity	−0.139971	0.01906	1.61e-01	0.148381	$-1.18e-01$
Financial.Leverage	−0.167996	−0.26275	6.26e-02	−0.008905	$-1.37e-01$
Annualized.Return.on.Common.Eq	−0.139971	0.01906	1.61e-01	0.148381	$-1.18e-01$
Operating.ROE	0.329034	−0.22373	-4.28e-01	−0.319978	$3.48e-01$
Number.of.Employees	−0.298550	−0.14753	-1.75e-01	−0.029635	$-1.01e-01$
Employees...1.Yr.Growth	0.087605	0.02418	4.36e-02	0.010112	$-8.81e-03$
Net.Income.Per.1000.Employees	−0.030930	0.10030	9.97e-02	0.013954	$2.36e-02$
Actual.Sales.Per.Employee	0.056871	0.31199	-2.76e-02	−0.280353	$-4.79e-01$
Assets.Per.1000.Employees	−0.075584	−0.29623	5.37e-03	0.207470	$3.82e-01$
Total.Debt.Total.Capital	−0.342696	−0.04332	1.85e-02	−0.248894	$-3.18e-02$
Assets.Equity	−0.157651	−0.10516	-1.11e-01	0.104208	$2.51e-02$
Tangible.Common.Equity.Ratio	−0.176106	−0.34235	-8.50e-02	−0.251334	$-1.21e-01$

Variables	PC26	PC27	PC28	PC29	PC30
LT.Debt.Common.Equity	-0.202105	0.08788	3.58e-01	-0.284126	$2.87e-01$
LT.Debt.Total.Capital	-0.021157	-0.16377	-2.74e-01	0.188523	$-2.74e-01$
LT.Debt.Total.Assets	0.000957	0.14797	2.62e-01	-0.163937	$2.91e-01$
Total.Debt.Common.Equity	-0.012130	-0.03032	-2.41e-02	0.162884	$-3.64e-02$
Total.Debt.Tangible.Book.Value	0.290962	-0.34787	1.15e-01	-0.398356	$-1.85e-01$
Total.Debt.Total.Equity	-0.012168	0.00349	-6.37e-02	0.191232	$-8.62e-03$
Total.Debt.Total.Assets	0.271018	-0.02181	-9.68e-02	0.394559	$4.07e-05$
Net.Debt.Shareholders.Equity	0.131598	0.34541	-3.82e-01	-0.116100	$-1.08e-01$
Net.Change.in.Liabilities...of.Total	0.010434	-0.01293	-9.63e-03	0.008005	$6.41e-03$
Increase.In.Equity.of.Total	-0.010434	0.01293	9.63e-03	-0.008005	$-6.41e-03$
Sales.Cash	0.009255	-0.01193	-4.83e-02	0.021551	$-1.41e-02$
Sales.Marketable.Securities	0.032396	-0.01306	-2.68e-02	-0.011080	$1.42e-02$
Sales.Net.Fixed.Assets	0.049352	-0.00471	1.34e-02	-0.037329	$-4.53e-02$
Sales.LT.Investments	0.197318	0.19883	4.04e-02	0.107118	$1.12e-01$
Sales.Other.Assets	0.021777	0.01232	1.11e-02	0.000818	$2.35e-02$
Tangible.Common.Equity	0.342582	0.05324	1.30e-01	-0.054770	$8.17e-02$
Tot.loan.tot.dep	-0.091661	-0.05154	-3.10e-02	0.000132	$-2.17e-02$

Variables	PC31	PC32	PC33	PC34	PC35
Net.Income...1.Yr.Growth	0.00103	0.010257	0.00216	0.000805	0.002967
Assets...1.Yr.Growth	−0.31237	−0.069897	-0.01430	−0.013467	−0.040718
Net.Worth...1.Yr.Growth	0.07613	0.138445	0.00843	0.010046	0.051659
Capital...1.Yr.Growth	0.22360	−0.031637	0.01805	0.011834	0.006429
Net.Fixed.Asset.Turnover	−0.05796	−0.002523	-0.00760	0.003235	0.001578
Operating.Margin	0.09103	−0.047735	-0.00374	−0.000147	0.006435
Pretax.Margin	−0.05035	0.030843	-0.03735	−0.014814	0.017258
Profit.Margin	0.01368	−0.017439	-0.00368	0.003116	0.001191
Return.on.Common.Equity	0.05825	−0.022081	-0.01763	−0.018162	−0.002194
Financial.Leverage	−0.31800	−0.664153	-0.10844	0.010603	−0.235816
Annualized.Return.on.Common.Eq	0.05825	−0.022081	-0.01763	−0.018162	−0.002194
Operating.ROE	−0.15874	0.034598	0.05503	0.046903	−0.020788
Number.of.Employees	0.11157	0.033504	-0.01051	−0.005182	−0.000231
Employees...1.Yr.Growth	0.02624	0.000917	-0.00648	0.002292	−0.003532
Net.Income.Per.1000.Employees	0.02506	0.027890	0.01503	0.002384	0.002090
Actual.Sales.Per.Employee	−0.29287	0.089282	-0.05359	−0.012830	0.017533
Assets.Per.1000.Employees	0.24989	−0.115721	0.05151	0.006330	−0.003337
Total.Debt.Total.Capital	0.19462	0.045546	0.12364	0.075904	−0.008485
Assets.Equity	0.01527	0.371627	-0.57082	−0.303464	0.336780
Tangible.Common.Equity.Ratio	−0.01168	0.025862	-0.01432	0.018616	0.009239

Variables	PC31	PC32	PC33	PC34	PC35
LT.Debt.Common.Equity	−0.05443	0.128598	0.44400	−0.451237	−0.034656
LT.Debt.Total.Capital	0.01565	0.062029	0.18298	−0.062135	−0.018196
LT.Debt.Total.Assets	0.00372	−0.141092	-0.50489	0.349240	0.041952
Total.Debt.Common.Equity	−0.19537	0.035501	0.37241	0.512991	0.597479
Total.Debt.Tangible.Book.Value	0.00317	0.135775	-0.07016	−0.034987	0.011347
Total.Debt.Total.Equity	−0.03192	0.447503	0.01508	0.341505	−0.681199
Total.Debt.Total.Assets	−0.29533	−0.079871	-0.00745	−0.422798	−0.003157
Net.Debt.Shareholders.Equity	0.59009	−0.306123	0.02217	−0.085483	0.000655
Net.Change.in.Liabilities.of.Total	0.00439	−0.002267	0.00172	0.000803	0.000060
Increase.In.Equity...of.Total	−0.00439	0.002267	-0.00172	−0.000803	−0.000060
Sales.Cash	0.02949	−0.031875	-0.02111	0.017162	0.007576
Sales.Marketable.Securities	0.05506	−0.032866	0.01516	−0.003156	−0.001946
Sales.Net.Fixed.Assets	0.06824	−0.009165	0.01501	−0.005289	−0.001283
Sales.LT.Investments	0.07252	−0.015943	0.06103	0.031821	−0.026787
Sales.Other.Assets	−0.01654	−0.006386	-0.00298	−0.004557	−0.000300
Tangible.Common.Equity	−0.08335	−0.033006	0.02688	0.006180	−0.007804
Tot.loan.tot.dep	0.00260	−0.001734	-0.00213	0.000830	0.005474

Variables	PC36	PC37
Net.Income...1.Yr.Growth	$0.00e+00$	$0.00e+00$
Assets...1.Yr.Growth	$-1.32e-16$	$1.29e-16$
Net.Worth...1.Yr.Growth	$9.63e-17$	$4.78e-17$
Capital...1.Yr.Growth	$-5.64e-17$	$-3.61e-17$
Net.Fixed.Asset.Turnover	$8.87e-16$	$-6.83e-18$
Operating.Margin	$-2.07e-16$	$-1.69e-16$
Pretax.Margin	$6.62e-16$	$1.07e-16$
Profit.Margin	$8.53e-17$	$-2.41e-16$
Return.on.Common.Equity	$-6.30e-01$	$3.21e-01$
Financial.Leverage	$-1.70e-15$	$1.13e-16$
Annualized.Return.on.Common.Equity	$6.30e-01$	$-3.21e-01$
Operating.ROE	$-3.62e-16$	$-2.96e-17$
Number.of.Employees	$4.22e-16$	$2.09e-16$
Employees...1.Yr.Growth	$1.30e-16$	$-1.22e-17$
Net.Income.Per.1000.Employees	$1.33e-16$	$5.66e-18$
Actual.Sales.Per.Employee	$-2.18e-16$	$1.01e-16$
Assets.Per.1000.Employees	$3.20e-16$	$1.45e-16$
Total.Debt.Total.Capital	$2.88e-16$	$1.24e-16$
Assets.Equity	$7.86e-16$	$-2.32e-16$
Tangible.Common.Equity.Ratio	$-3.70e-16$	$2.53e-17$

Variables	PC36	PC37
LT.Debt.Common.Equity	$-1.01e-16$	$-1.06e-16$
LT.Debt.Total.Capital	$1.78e-16$	$1.65e-16$
LT.Debt.Total.Assets	$3.77e-16$	$1.05e-16$
Total.Debt.Common.Equity	$1.00e-15$	$-4.15e-16$
Total.Debt.Tangible.Book.Value	$3.14e-16$	$-4.63e-17$
Total.Debt.Total.Equity	$-8.25e-16$	$2.92e-16$
Total.Debt.Total.Assets	$3.73e-16$	$-5.66e-17$
Net.Debt.Shareholders.Equity	$-2.70e-16$	$1.58e-16$
Net.Change.in.Liabilities...of.Total	$3.21e-01$	$6.30e-01$
Increase.In.Equity...of.Total	$3.21e-01$	$6.30e-01$
Sales.Cash	$1.54e-16$	$1.17e-16$
Sales.Marketable.Securities	$-1.16e-16$	$-8.02e-17$
Sales.Net.Fixed.Assets	$-5.75e-16$	$-1.06e-16$
Sales.LT.Investments	$4.55e-16$	$-1.57e-16$
Sales.Other.Assets	$1.25e-16$	$1.43e-16$
Tangible.Common.Equity	$-7.80e-17$	$9.10e-17$
Tot.loan.tot.dep	$3.23e-17$	$1.25e-16$

Appendix F

Appendix - R-Programm

```
#   Titel:          Data preparation and modeling
#   Description:    Data cleaning, transformation, missing imputation and
    modeling
#   Creation date:  27.05.2015
#   Last change:    27.08.2015
#   Autor:          Lavri Labi


#loading required libraries####
library(gdata)
library(sfsmisc)
library(MASS)
library(glmnet)
library(DAAG)
library(lme4)
library(nlme)
library(forecast)
library(lmerTest)

#reading data####

# Initialise default directory
default.dir <- "D:\\Lavri document\\phd thesis\\data\\psvag\\Data"

# reading BB Financial in the RAM
data.BBfinancial <- read.csv2(choose.files(default=paste(default.dir,"\\BBG
    Financials_20150715.csv"),
```

```
                                                    caption="BBG Financials"), header = TRUE, as.is = T)
View(data.BBfinancial)

# reading Emittenten Historische Ratings in the RAM
data.histRating <- read.csv2(choose.files(default=paste(default.dir,"\\Emittenten Historische Ratings.csv"),
                                                    caption="Emittenten Historische Ratings"), header = TRUE, as.is = T)
View(data.histRating)

# reading rating order of the rating agency
ratingOrder.moodys <- read.csv2(choose.files(default=paste(default.dir,"\\ratingOrder_moodys.csv"),
                                                    caption="choose ratingOrder_moodys"),
                                      header = TRUE, as.is = T)
ratingOrder.sp <- read.csv2(choose.files(default=paste(default.dir,"\\ratingOrder_sp.csv"),
                                                    caption="choose ratingOrder_sp"),
                                  header = TRUE, as.is = T)
ratingOrder.fitch <- read.csv2(choose.files(default=paste(default.dir,"\\ratingOrder_fitch.csv"),
                                                    caption="choose ratingOrder_fitch"),
                                      header = TRUE, as.is = T)

# reading PD to rating-grade of Moodys
PD2RatingGrade.moodys <- read.csv2(choose.files(default=paste(default.dir,"\\moodys_PD_to_ratingGrade.csv"),
                                                    caption="choose moodys_PD_to_ratingGrade"),
                                        header = TRUE, as.is = T)
# saving PD2RatingGrade.moodys
save(PD2RatingGrade.moodys, file=paste(prog.path, "\\PD2RatingGradeMoodys.RData", sep=""))

# reading PD to rating-grade of Moodys
Real.GDP <- read.csv2(choose.files(default=default.dir,
                            caption="Real_GDP_Forecast"), header = TRUE, as.is = T, sep=",")
```

```

#declaring paths####
prog.path <- "D:\\Lavri document\\phd thesis\\data\\psvag\\programm\\PSVaG_R_Prog"

#declaring functions####

# function to select last rating
selectLastRating <- function(rating){
  if (is.na(rating)) {
    lastRating <- rating
  }else{
    vect.ratings <- trim(gsub("\\s+|-", "", strsplit(rating, "/")[[1]]))
    vect.ratings.rel <- vect.ratings[which(vect.ratings != "")]
    n.ratings <- length(vect.ratings.rel)
    if (n.ratings == 0){
      lastRating <- NA
      return(lastRating)
    }
    if (n.ratings == 1) {
      lastRating <- vect.ratings.rel[n.ratings]
    }else{
      order.index.moodys.fun <- which(tolower(ratingOrder.moodys$Grade) %in% tolower(vect.ratings.rel))
      order.index.sp.fun <- which(tolower(ratingOrder.sp$Grade) %in% tolower(vect.ratings.rel))
      order.index.fitch.fun <- which(tolower(ratingOrder.fitch$Grade) %in% tolower(vect.ratings.rel))
      # joining all relevant rating-grades with their related orders
      ratingOrder.all.rel <- rbind(ratingOrder.moodys[order.index.moodys.fun,], ratingOrder.sp[order.index.sp.fun,],
                                   ratingOrder.fitch[order.index.fitch.fun,])
      # Order for selection of the selection in the next command
      ratingOrder.all.rel.ordered <- ratingOrder.all.rel[order(ratingOrder.all.rel$Order),]
      ratingOrder.all.rel.ordered$Grade <- tolower(ratingOrder.all.rel.ordered$Grade)
      ratingOrder.all.rel.ordered.unik <- ratingOrder.all.rel.ordered[!duplicated(ratingOrder.all.rel.ordered),]
      # selecting the second best rating as required by BaFin
```

```
      lastRating  <- ratingOrder.all.rel.ordered.unik$Grade[2]
    }
  }
  return(lastRating)
}

# function to substract from date
my.datediff <- function(datum, add.num){
  datum.only.num <- as.integer(strsplit(datum, "\\.")[[1]][2])
  datum.add <- as.character(ifelse(datum.only.num == 0, 100, datum.only.num
    ) + add.num)
  if (nchar(datum.add) < 2){
    datum.add.format <- paste("0", datum.add, sep="")
  }else{
    datum.add.format <- substring(datum.add, nchar(datum.add) - 1, nchar(
    datum.add))
  }
  return(paste("Dez.", datum.add.format, sep=""))
}

# function to extract date part from row.names
extract.date <- function(string2extract){
  strsplit(string2extract, "__")[[1]][2]
}

# function to transform dates
trans.date <- function(date){
  date.part <- as.integer(strsplit(as.character(date), "\\.")[[1]][2])
  ifelse(date.part > 50, date.part + 1900, date.part + 2000)
}

conv.pd.rating <- function(pd2Rating){
  which(abs(PD2RatingGrade.moodys$Year_10 - pd2Rating) == min(abs(
    PD2RatingGrade.moodys$Year_10 - pd2Rating)))
}

#data preparation - financials####

## selecting all variables in the basis table
unique.var <- unique(data.BBfinancial$var)

```

```
## selecting all emitters in the basis table
unique.emit <- unique(data.BBfinancial$emitter)

# transforming data.BBfinancial in a suitable matrix
all.var <- gsub("\\s+|[[:punct:]]", ".",trim(unique(data.BBfinancial$var)))
all.dates <- colnames(data.BBfinancial)[-(1:3)]
all.emitters <- unique(data.BBfinancial$emitter)

## creating an empty data frame
data.BBfinancial.trans <- read.csv(text=paste(c("emitter","date", all.var), collapse=","),
                                   colClasses=c(rep("character",2),rep("numeric",length(all.var))))

## Filling data.BBfinancial.trans with value from data.BBfinancial
for (i in 1:length(all.emitters)){
  for (j in 1:length(all.dates)){
    data.BBfinancial.trans[((i-1)*length(all.dates)+j), "emitter"] <- all.emitters[i]
    data.BBfinancial.trans[((i-1)*length(all.dates)+j), "date"] <- all.dates[j]
    for (k in all.var){
      rel.index <- which((data.BBfinancial$emitter == all.emitters[i]) & (gsub("\\s+|[[:punct:]]", ".",trim(data.BBfinancial$var)) == k))
        data.BBfinancial.trans[((i-1)*length(all.dates)+j), k] <- data.BBfinancial[rel.index[1], all.dates[j]]
    }
  }
}

# Index of variables to exclude
index.var.rm <- which(colnames(data.BBfinancial.trans) == gsub("\\s+|[[:punct:]]", ".","Est Basel III Tier 1 CE Ratio Fully Phased In")
                      | colnames(data.BBfinancial.trans) == gsub("\\s+|[[:punct:]]", ".","Est Basel III RWA")
                      #| colnames(data.BBfinancial.trans) == gsub("\\s+|[[:punct:]]", ".","Number of Employees")
                      )

# Excluding variables
data.BBfinancial.trans.rm <- data.BBfinancial.trans[,-index.var.rm]
```

```
#plots - financials####

# set work directory for pdf-files
wd <- choose.dir(default = default.dir, caption = "analysis")
setwd(wd)

# settings for pdf-outputs
mfrow.vec <- c(3,3)

var.path <- paste(wd, "\\","analysis by var", sep="")
if (file.exists(var.path)) unlink(var.path, recursive=TRUE)
dir.create(var.path, showWarnings = TRUE)

## iteration over all variables in order to create related diagramm for
    each emitter
for (i.var in 1:length(unique.var)){
  data.BBfinancial.var <- data.BBfinancial[data.BBfinancial$var == unique.
    var[i.var], ]
  pdf(paste(var.path, "\\", trim(gsub("\\s+|[[:punct:]]", "",unique.var[i.
    var])), ".pdf", sep=""))
  op <- par(mfrow = mfrow.vec)
  for (i.emit in 1:nrow(data.BBfinancial.var)){
    data.BBfinancial.var.matrix <- as.matrix(data.BBfinancial.var[i.emit,-c
    (1,2,3)])
    xaxis <- substr(colnames(data.BBfinancial.var.matrix), nchar(colnames(
    data.BBfinancial.var.matrix))-1,
                    nchar(colnames(data.BBfinancial.var.matrix)))
    barplot(data.BBfinancial.var.matrix, names.arg = xaxis, cex.names =
    .85, space = .5, axis.lty = 1,
            main = paste(trim(data.BBfinancial.var[i.emit,1]), trim(data.
    BBfinancial.var[i.emit,3]), sep="\n"))
    substr(colnames(data.BBfinancial.var.matrix), 5, 7)
  }
  par(op)
  dev.off()
}

## creating folder for "analysis by emitter"
var.path <- paste(wd, "\\","analysis by emitter", sep="")
if (file.exists(var.path)) unlink(var.path, recursive=TRUE)
```

```
dir.create(var.path, showWarnings = TRUE)

## iteration over all variables in order to create related diagramm for each emitter
for (i.emit in 1:length(unique.emit)){
  data.BBfinancial.emit <- data.BBfinancial[data.BBfinancial$emit == unique.emit[i.emit], ]
  pdf(paste(var.path, "\\", trim(gsub("\\s+|[[:punct:]]", "",unique.emit[i.emit])), ".pdf", sep=""))
  op <- par(mfrow = mfrow.vec)
  for (i.emit in 1:nrow(data.BBfinancial.emit)){
    data.BBfinancial.emit.matrix <- as.matrix(data.BBfinancial.emit[i.emit,-c(1,2,3)])
    xaxis <- substr(colnames(data.BBfinancial.emit.matrix), nchar(colnames(data.BBfinancial.emit.matrix))-1,
                    nchar(colnames(data.BBfinancial.emit.matrix)))
    barplot(data.BBfinancial.emit.matrix, names.arg = xaxis, cex.names = .85, space = .5, axis.lty = 1,
            main = paste(trim(data.BBfinancial.emit[i.emit,1]), trim(data.BBfinancial.emit[i.emit,3]), sep="\n"))
    substr(colnames(data.BBfinancial.emit.matrix), 5, 7)
  }
  par(op)
  dev.off()
}

## creating folder for "analysis of hist and boxplot"
var.path <- paste(wd, "\\","analysis of hist and boxplot", sep="")
if (file.exists(var.path)) unlink(var.path, recursive=TRUE)
dir.create(var.path, showWarnings = TRUE)

## creating histogram and boxplot for the variables
## iteration over all variables in order to create related diagramm for each emitter
for (i.var in 1:length(unique.var)){
  data.BBfinancial.var <- data.BBfinancial[data.BBfinancial$var == unique.var[i.var], ]
  data.BBfinancial.var.matrix <- as.matrix(data.BBfinancial.var[,-c(1,2,3)])
  pdf(paste(var.path, "\\", trim(gsub("\\s+|[[:punct:]]", "",unique.var[i.var])), ".pdf", sep=""))
```

```
  op <- par(mfrow = mfrow.vec)
  for (t.serie in 1:ncol(data.BBfinancial.var.matrix)){
    if (all(is.na(data.BBfinancial.var.matrix[,t.serie]))){
      plot.new()
      title(paste(trim(data.BBfinancial.var[i.var,3]), colnames(data.
    BBfinancial.var.matrix)[t.serie], sep="\n"))
      next
    }
    histBxp(data.BBfinancial.var.matrix[,t.serie], main = colnames(data.
    BBfinancial.var.matrix)[t.serie],
            xlab = trim(data.BBfinancial.var[i.var,3]))
  }
  par(op)
  dev.off()
}

#data preparation - ratings####

data.histRating.lastR <- data.histRating
# selecting the last rating
for (i in 4:ncol(data.histRating.lastR)){
  data.histRating.lastR[,i] <- unlist(lapply(data.histRating.lastR[,i],
    selectLastRating))
}
View(data.histRating.lastR)

## selecting all rating agency in data.histRating.lastR
unique.rating_Ag <- unique(data.histRating.lastR$rating_Ag)

unique.Schuldner <- unique(data.histRating.lastR$Schuldner)

## creating subset for each rating agency
data.histRating.Moodys <- subset(data.histRating.lastR, data.histRating.
    lastR$rating_Ag == "Moodys")
data.histRating.SP <- subset(data.histRating.lastR, data.histRating.lastR$
    rating_Ag == "S+P")
data.histRating.Fitch <- subset(data.histRating.lastR, data.histRating.
    lastR$rating_Ag == "Fitch")

## replacing rating by order
### Moodys
```

```r
for (i in 1:nrow(data.histRating.Moodys)){
  for (j in 4:ncol(data.histRating.Moodys)){
    #print(data.histRating.Moodys[i,j]); print(i); print(j)
    if (is.na(data.histRating.Moodys[i,j])) next
    if (trim(data.histRating.Moodys[i,j])=="") {
      data.histRating.Moodys[i,j] <- NA
      next
    }
    order.index.moodys <- which(tolower(ratingOrder.moodys$Grade) == tolower(trim(data.histRating.Moodys[i,j])))
    order.index.sp <- which(tolower(ratingOrder.sp$Grade) == tolower(trim(data.histRating.SP[i,j])))
    order.index.fitch <- which(tolower(ratingOrder.fitch$Grade) == tolower(trim(data.histRating.Fitch[i,j])))
    order.index <- ifelse(length(order.index.moodys) != 0, order.index.moodys,
                          ifelse(length(order.index.sp) != 0, order.index.sp, order.index.fitch))
    data.histRating.Moodys[i,j] <- ratingOrder.moodys$Order[order.index]
  }
}
### S and P
for (i in 1:nrow(data.histRating.SP)){
  for (j in 4:ncol(data.histRating.SP)){
    #print(data.histRating.SP[i,j]); print(i); print(j)
    if (is.na(data.histRating.SP[i,j])) next
    if (trim(data.histRating.SP[i,j])=="") {
      data.histRating.SP[i,j] <- NA
      next
    }
    order.index.moodys <- which(tolower(ratingOrder.moodys$Grade) == tolower(trim(data.histRating.Moodys[i,j])))
    order.index.sp <- which(tolower(ratingOrder.sp$Grade) == tolower(trim(data.histRating.SP[i,j])))
    order.index.fitch <- which(tolower(ratingOrder.fitch$Grade) == tolower(trim(data.histRating.Fitch[i,j])))
    order.index <- ifelse(length(order.index.moodys) != 0, order.index.moodys,
                          ifelse(length(order.index.sp) != 0, order.index.sp, order.index.fitch))
    data.histRating.SP[i,j] <- ratingOrder.sp$Order[order.index]
```

```
    }
}
### Fitch
for (i in 1:nrow(data.histRating.Fitch)){
  for (j in 4:ncol(data.histRating.Fitch)){
    #print(data.histRating.Fitch[i,j]); print(i); print(j)
    if (is.na(data.histRating.Fitch[i,j])) next
    if (trim(data.histRating.Fitch[i,j])=="") {
      data.histRating.Fitch[i,j] <- NA
      next
    }
    order.index.moodys <- which(tolower(ratingOrder.moodys$Grade) ==
      tolower(trim(data.histRating.Moodys[i,j])))
    order.index.sp <- which(tolower(ratingOrder.sp$Grade) == tolower(trim(
      data.histRating.SP[i,j])))
    order.index.fitch <- which(tolower(ratingOrder.fitch$Grade) == tolower(
      trim(data.histRating.Fitch[i,j])))
    order.index <- ifelse(length(order.index.moodys) != 0, order.index.
      moodys,
                          ifelse(length(order.index.sp) != 0, order.index.
      sp, order.index.fitch))
    data.histRating.Fitch[i,j] <- ratingOrder.fitch$Order[order.index]
  }
}

## create basis table
data.histRating.basisTable<- character(ncol(data.histRating.Moodys))
data.histRating.basisTable<- NULL

## combining all ratings from all agencies
for (i.schuldner in unique.Schuldner){
  i.schuldner.index.Moodys <- which(data.histRating.Moodys$Schuldner == i.
    schuldner)
  i.schuldner.index.SP <- which(data.histRating.SP$Schuldner == i.schuldner
    )
  i.schuldner.index.Fitch <- which(data.histRating.Fitch$Schuldner == i.
    schuldner)
  if (length(i.schuldner.index.Moodys) > 0){
    i.schuldner.vect <- data.histRating.Moodys[i.schuldner.index.Moodys ,]
  }else{
    if(length(i.schuldner.index.SP) > 0){
```

```r
      i.schuldner.vect <- data.histRating.SP[i.schuldner.index.SP ,]
    }else{
      i.schuldner.vect <- data.histRating.Fitch[i.schuldner.index.Fitch ,]
    }
  }
  for (i.year in 4:ncol(data.histRating.Moodys)){
    if (!is.na(i.schuldner.vect[i.year])) next
    if(length(i.schuldner.index.SP) > 0){
      i.schuldner.vect[i.year] <- data.histRating.SP[i.schuldner.index.SP , i.year]
    }else{
      i.schuldner.vect[i.year] <- data.histRating.Fitch[i.schuldner.index.Fitch ,i.year]
    }
  }
  #print(i.schuldner.vect)
  data.histRating.basisTable <- rbind(data.histRating.basisTable, i.schuldner.vect)
}

all.dates.rtg <- colnames(data.histRating.basisTable)[-(1:3)]
all.emitters.rtg <- unique(data.histRating.basisTable$Schuldner_Abkzg)

## creating an empty data frame
data.histRating.basisTable.trans <- read.csv(text=paste(c("emitter","date","Order"), collapse=","),
                                             colClasses=c(rep("character",2),"integer"))

## Filling data.histRating.basisTable.trans with value from data.histRating.basisTable
for (i in 1:length(all.emitters.rtg)){
  for (j in 1:length(all.dates.rtg)){
    data.histRating.basisTable.trans[((i-1)*length(all.dates)+j), "emitter"] <- all.emitters.rtg[i]
    data.histRating.basisTable.trans[((i-1)*length(all.dates)+j), "date"] <- all.dates.rtg[j]

    rel.index <- which(data.histRating.basisTable$Schuldner_Abkzg == all.emitters.rtg[i])
```

```
    data.histRating.basisTable.trans[((i-1)*length(all.dates.rtg)+j), "
    Order"] <- data.histRating.basisTable[rel.index[1], all.dates.rtg[j]]
  }
}

## Transforming order in PD
data.histRating.PD <- data.histRating.basisTable.trans
colnames(data.histRating.PD)[3] <- "PD";
class(data.histRating.PD$PD) <- "numeric"
data.histRating.all <- list()
for (k in 1:10){
  for(i in 1:nrow(data.histRating.basisTable.trans)){
    if (is.na(data.histRating.basisTable.trans$Order[i])) next
    data.histRating.PD$PD[i] <- PD2RatingGrade.moodys[which(PD2RatingGrade.
    moodys$Order == data.histRating.basisTable.trans$Order[i]), (k+2)]
  }
  data.histRating.all[[k]] <- data.histRating.PD
}

## Giving names to list elements) <-
names(data.histRating.all) <- colnames(PD2RatingGrade.moodys)[-c(1,2)]

#plots - ratings####

# set work directory for pdf-files
wd <- choose.dir(default = default.dir, caption = "analysis")
setwd(wd)

# settings for pdf-outputs
mfrow.vec <- c(3,3)

# creating barplot for visualisation
## selecting all rating agency in data.histRating.lastR
unique.rating_Ag <- unique(data.histRating.lastR$rating_Ag)

## creating folder for "analysis by rating agency"
var.path <- paste(wd, "\\","analysis by rating agency", sep="")
if (file.exists(var.path)) unlink(var.path, recursive=TRUE)
dir.create(var.path, showWarnings = TRUE)

```

```
## iteration over all variables in order to create related diagramm for
    each emitter
for (i.var in 1:length(unique.rating_Ag)){
  data.histRating.lastR.Ag <- data.histRating.lastR[data.histRating.lastR$
    rating_Ag == unique.rating_Ag[i.var], ]
  pdf(paste(var.path, "\\", trim(gsub("[[:punct:]]", " ",unique.rating_Ag[i
    .var])), ".pdf", sep=""))
  op <- par(mfrow = mfrow.vec)
  for (i.emit in 4:ncol(data.histRating.lastR.Ag)){
    if (all(is.na((data.histRating.lastR.Ag[,i.emit])))) next
    non.na.ratings <- sum(!is.na((data.histRating.lastR.Ag[,i.emit])))
    data.histRating.lastR.Ag.factor <- as.factor(data.histRating.lastR.Ag[,
    i.emit])
    plot(data.histRating.lastR.Ag.factor
         , main = paste(unique.rating_Ag[i.var], colnames(data.histRating.
    lastR.Ag)[i.emit],
                        paste(non.na.ratings, "ratings", sep=" "), sep="\n"
    ))
  }
  par(op)
  dev.off()
}

# settings for pdf-outputs
mfrow.vec <- c(3,2)

# creating barplot for visualisation

## creating folder for "analysis by rating agency"
var.path <- paste(wd, "\\","analysis for all rating combined", sep="")
if (file.exists(var.path)) unlink(var.path, recursive=TRUE)
dir.create(var.path, showWarnings = TRUE)

## iteration over all variables in order to create related balkendiagramm
    for each emitter
pdf(paste(var.path, "\\", "allRatingsCombined.pdf", sep=""))
op <- par(mfrow = mfrow.vec)
for (i.emit in 4:ncol(data.histRating.basisTable)){
  if (all(is.na((data.histRating.basisTable[,i.emit])))) next
  non.na.ratings <- sum(!is.na((data.histRating.basisTable[,i.emit])))
```

```
  data.histRating.basisTable.factor <- as.factor(data.histRating.basisTable
    [,i.emit])
  plot(data.histRating.basisTable.factor
       , main = paste(colnames(data.histRating.basisTable)[i.emit],
                      paste(non.na.ratings, "ratings", sep=" "), sep="\n"))
}
par(op)
dev.off()

## iteration over all variables in order to create related histogramm for
    each emitter
pdf(paste(var.path, "\\", "allRatingsCombinedHist.pdf", sep=""))
op <- par(mfrow = mfrow.vec)
for (i.emit in 4:ncol(data.histRating.basisTable)){
  if (all(is.na((data.histRating.basisTable[,i.emit])))) next
  non.na.ratings <- sum(!is.na((data.histRating.basisTable[,i.emit])))
  hist(as.numeric(data.histRating.basisTable[,i.emit]), ylab = "frequencies
    ", xlab="ratings"
       , main = paste(colnames(data.histRating.basisTable)[i.emit],
                      paste(non.na.ratings, "ratings", sep=" "), sep="\n"))
}
par(op)
dev.off()

#merging ratings and financials####

# creating vector with laged date
date.lag.1 <- unlist(lapply(data.BBfinancial.trans.rm$date, my.datediff, 1)
    )
date.lag.2 <- unlist(lapply(data.BBfinancial.trans.rm$date, my.datediff, 2)
    )
date.lag.3 <- unlist(lapply(data.BBfinancial.trans.rm$date, my.datediff, 3)
    )
date.lag.4 <- unlist(lapply(data.BBfinancial.trans.rm$date, my.datediff, 4)
    )
date.lag.5 <- unlist(lapply(data.BBfinancial.trans.rm$date, my.datediff, 5)
    )

# creating basis table with laged date for financial data
data.BBfinancial.trans.lag.0 <- data.BBfinancial.trans.rm;
```

```
colnames(data.BBfinancial.trans.lag.0) <- paste(colnames(data.BBfinancial.trans.lag.0), "lag0", sep="_")
colnames(data.BBfinancial.trans.lag.0)[1:2] <- c("emitter", "date")
data.BBfinancial.trans.lag.1 <- data.BBfinancial.trans.rm; data.BBfinancial.trans.lag.1$date <- date.lag.1
colnames(data.BBfinancial.trans.lag.1) <- paste(colnames(data.BBfinancial.trans.lag.1), "lag1", sep="_")
colnames(data.BBfinancial.trans.lag.1)[1:2] <- c("emitter", "date")
data.BBfinancial.trans.lag.2 <- data.BBfinancial.trans.rm; data.BBfinancial.trans.lag.2$date <- date.lag.2
colnames(data.BBfinancial.trans.lag.2) <- paste(colnames(data.BBfinancial.trans.lag.2), "lag2", sep="_")
colnames(data.BBfinancial.trans.lag.2)[1:2] <- c("emitter", "date")
data.BBfinancial.trans.lag.3 <- data.BBfinancial.trans.rm; data.BBfinancial.trans.lag.3$date <- date.lag.3
colnames(data.BBfinancial.trans.lag.3) <- paste(colnames(data.BBfinancial.trans.lag.3), "lag3", sep="_")
colnames(data.BBfinancial.trans.lag.3)[1:2] <- c("emitter", "date")
data.BBfinancial.trans.lag.4 <- data.BBfinancial.trans.rm; data.BBfinancial.trans.lag.4$date <- date.lag.4
colnames(data.BBfinancial.trans.lag.4) <- paste(colnames(data.BBfinancial.trans.lag.4), "lag4", sep="_")
colnames(data.BBfinancial.trans.lag.4)[1:2] <- c("emitter", "date")
data.BBfinancial.trans.lag.5 <- data.BBfinancial.trans.rm; data.BBfinancial.trans.lag.5$date <- date.lag.5
colnames(data.BBfinancial.trans.lag.5) <- paste(colnames(data.BBfinancial.trans.lag.5), "lag5", sep="_")
colnames(data.BBfinancial.trans.lag.5)[1:2] <- c("emitter", "date")

# creating a suitable for the regression
data4reg.year10.lag.0 <- merge(data.histRating.all[["Year_10"]], data.BBfinancial.trans.lag.0, by = c("emitter", "date"))
data4reg.year10.lag.1 <- merge(data.histRating.all[["Year_10"]], data.BBfinancial.trans.lag.1, by = c("emitter", "date"))
data4reg.year10.lag.2 <- merge(data.histRating.all[["Year_10"]], data.BBfinancial.trans.lag.2, by = c("emitter", "date"))
data4reg.year10.lag.3 <- merge(data.histRating.all[["Year_10"]], data.BBfinancial.trans.lag.3, by = c("emitter", "date"))
data4reg.year10.lag.4 <- merge(data.histRating.all[["Year_10"]], data.BBfinancial.trans.lag.4, by = c("emitter", "date"))
```

```
data4reg.year10.lag.5 <- merge(data.histRating.all[["Year_10"]], data.
    BBfinancial.trans.lag.5, by = c("emitter", "date"))

#start stepwise linear regression####

# variables initialisation
step.all.lag <- list()
# iteration
for (i.lag in 0:5){

  # choose the right table
  string2exec <- paste("data4reg.year10 <- data4reg.year10.lag.", i.lag,
    sep="")
  # code ausf hren
  eval(parse(text = string2exec))
  # adapting name ending
  lag.ini <- paste("lag", i.lag, sep="")

#cleaning after merging####

  # removing observations without ratings
  data4reg.year10.step1 <- data4reg.year10[!is.na(data4reg.year10$PD),]

  # Index of variable to exclude
  index.row.rm <- NULL
  for (i in 1:nrow(data4reg.year10.step1)){
    nr.non.missing <- sum(!is.na(data4reg.year10.step1[i,-c(1,2,3)]))
    if(nr.non.missing < 4){
      index.row.rm <- c(index.row.rm, i)
    }
  }

  # adapting the data basis
  data4reg.y10.rm <- data4reg.year10.step1[-index.row.rm,]

#missng value analysis####

  # filling level
  fill.rate <- numeric(ncol(data4reg.y10.rm[,-c(1,2,3)]))
  for (i in 1:ncol(data4reg.y10.rm[,-c(1,2,3)])){
    fill.rate[i] <- sum(is.na(data4reg.y10.rm[,(i+3)]))
```

```
    #fill.rate[i] <- (sum(is.na(data4reg.y10.rm[,(i+3)])) / nrow(data4reg.
      y10.rm[,(i+3)])) * 100
    names(fill.rate)[i] <- colnames(data4reg.y10.rm)[i+3]
  }
  fill.rate <- (fill.rate/nrow(data4reg.y10.rm)) * 100
  fill.rate.sel <- fill.rate[fill.rate < 50]
  #write.csv2(fill.rate, file = paste(cor.path, "\\fill_rate.csv", sep=""))

#correlation analysis####

  # calculating correlation
  cor.meth1 <- cor(data4reg.year10.step1[,names(fill.rate.sel)], use = "
    pairwise.complete.obs")
  cor.meth2 <- cor(data4reg.year10.step1[,names(fill.rate.sel)], use = "na.
    or.complete")

  # exporting correlation report
  #write.csv2(cor.meth1, file = paste(cor.path, "\\cor_less10_meth1.csv",
    sep=""))
  #write.csv2(cor.meth2, file = paste(cor.path, "\\cor_less10_meth2.csv",
    sep=""))

  # removing correlated variables
  vor.corr.rm <- c("Capital...1.Yr.Growth", "Operating.Margin", "Pretax.
    Margin", "Return.on.Common.Equity"
                      , "Annualized.Return.on.Common.Equity", "Operating.ROE",
     "Actual.Sales.Per.Employee"
                      , "LT.Debt.Common.Equity", "Total.Debt.Common.Equity", "
    Total.Debt.Tangible.Book.Value"
                      , "Total.Debt.Total.Equity", "Net.Debt.Shareholders.
    Equity", "Total.Debt.Total.Assets"
                      , "Tangible.Common.Equity")
  vor.corr.rm <- paste(vor.corr.rm, lag.ini, sep="_")
  corr.sel.var <- setdiff(names(fill.rate.sel), vor.corr.rm)

#data transformation and additional cleaning#####

  # transforming depending variable
  logit.PD <- log(data4reg.year10.step1$PD/(1-data4reg.year10.step1$PD))
  data4reg.last <- data4reg.year10.step1[,c("PD", corr.sel.var)]
```

```
  rownames(data4reg.last) <- paste(data4reg.year10.step1$emitter, data4reg.year10.step1$date, sep="__")
  data4reg.last$PD <- logit.PD; colnames(data4reg.last)[colnames(data4reg.last) == "PD"] <- "logit.PD"

  # removing rows without observation
  row2rm.all <- NULL
  row2rm.any <- NULL
  for (i in 1:nrow(data4reg.last)){
    if (all(is.na(data4reg.last[i,-1]))){
      row2rm.all <- c(row2rm.all, i)
    }
    if (any(is.na(data4reg.last[i,-1]))){
      row2rm.any <- c(row2rm.any, i)
    }
  }
  data4reg.last <- data4reg.last[-row2rm.any,]

# Missing value Imputation using Nearest Neighbour with Mahalanobis Distance

  # Function to calculate Imputation value based on Paiwise Mahalanobis Ditance
  imputation.value <- function (imp.index, basis.mat = data4reg.last[,-c(1,2,3)]){
    imp.rowIndex <- imp.index%%nrow(basis.mat)
    imp.colIndex <- as.integer(imp.index/nrow(basis.mat)) + ifelse(imp.index%%nrow(basis.mat) > 0, 1, 0)
    basis.mat.target.rm <- basis.mat[-imp.rowIndex,-imp.colIndex]

    # removing rows without observation
    row2rm.any <- NULL
    for (i in 1:nrow(basis.mat.target.rm)){
      if (any(is.na(basis.mat.target.rm[i,-1]))){
        row2rm.any <- c(row2rm.any, i)
      }
    }
    basis.mat.target.rm.last <- basis.mat.target.rm[-row2rm.any,]

    maha.dist <- mahalanobis(basis.mat.target.rm.last, basis.mat[imp.rowIndex,], cov(basis.mat.target.rm.last, na.rm=T))
```

```
    maha.dist.min.rowN <- row.names(basis.mat.target.rm.last[which.min(maha
    .dist),])
    value2impute <- basis.mat[which(row.names(basis.mat)==maha.dist.min.
   rowN),imp.colIndex]
    return(value2impute)
  }

  # Checking the number of missing in observations
  na.number.check <- function(obs){
    na.number <- sum(is.na(obs))
  }

  # Removing all observations with more than 2 missing values
  na.number.vec <- apply(data4reg.last, 1, na.number.check)
  data4reg <- data4reg.last[-which(na.number.vec > 5),]
  index2impute <- which(is.na(data4reg))
  data4reg[index2impute] <- sapply(index2impute, imputation.value)


#linear regression####

  # fiting linear regression model
  fit <- lm(logit.PD ~ ., data = data4reg, na.action=na.omit)

  # performing stewise regression based on the fitted linear regression
    model
  step <- stepAIC(fit, direction="both", na.action=na.omit, trace = 0)
  summary(step)
  # adding result in list
  step.all.lag[[paste("lag", i.lag, sep="")]] <- step

  # printing result analysis plots
  op <- par(mfrow=c(1,2))
    # ploting residuals
    plot(step$residuals)
    abline(h=0);abline(h=1.96);abline(h=-1.96)
    # computing predictions
    preditc.fit <- predict(fit, se.fit=T)
    pd.pred <- exp(preditc.fit$fit)/(exp(preditc.fit$fit) + 1)
    pd <- exp(data4reg$logit.PD)/(exp(data4reg$logit.PD) + 1)
    # ploting predictions against realisations
```

```
    plot(pd,pd.pred)
    abline(a=0, b=1, type="l")
  par(op)

}

#including lag-variables in linear regression####

# removing lagX in variable names
lag0.model.var <- substr(names(step.all.lag$lag0$coefficients[-1]), 1
                          , (nchar(names(step.all.lag$lag0$coefficients[-1])
    )-5))
lag1.model.var <- substr(names(step.all.lag$lag1$coefficients[-1]), 1
                          , (nchar(names(step.all.lag$lag1$coefficients[-1])
    )-5))
lag2.model.var <- substr(names(step.all.lag$lag2$coefficients[-1]), 1
                          , (nchar(names(step.all.lag$lag2$coefficients[-1])
    )-5))
lag3.model.var <- substr(names(step.all.lag$lag3$coefficients[-1]), 1
                          , (nchar(names(step.all.lag$lag3$coefficients[-1])
    )-5))
lag4.model.var <- substr(names(step.all.lag$lag4$coefficients[-1]), 1
                          , (nchar(names(step.all.lag$lag4$coefficients[-1])
    )-5))
lag5.model.var <- substr(names(step.all.lag$lag5$coefficients[-1]), 1
                          , (nchar(names(step.all.lag$lag5$coefficients[-1])
    )-5))

# finding additional relevant variable in each lagX-regression compared to
    lag-regression
diff.lag0.lag1 <- paste(setdiff(lag1.model.var, lag0.model.var), "_lag1",
    sep="")
diff.lag0.lag2 <- paste(setdiff(lag2.model.var, union(lag1.model.var, lag0.
    model.var)), "_lag2", sep="")
diff.lag0.lag3 <- paste(setdiff(lag3.model.var, union(lag2.model.var, union
    (lag1.model.var, lag0.model.var)))
                         , "_lag3", sep="")
diff.lag0.lag4 <-  paste(setdiff(lag4.model.var,
                                  union(lag3.model.var, union(lag2.model.var
    ,
```

```
                                                                    union(lag1.
    model.var, lag0.model.var))))
                        , "_lag4", sep="")
# comment since setdiff is empty
# diff.lag0.lag5 <- paste(setdiff(lag5.model.var,
#                          union(lag4.model.var, union(lag3.model.
    var, union(lag2.model.var,
#                                                                   union(lag1.
    model.var, lag0.model.var)))))
#                       , "_lag4", sep="")

# crating table including relevant lag-variable
data4reg.year10.lag.0.1 <- merge(data4reg.year10.lag.0, data4reg.year10.lag
    .1[,c("emitter", "date", diff.lag0.lag1)]
                                 , by = c("emitter", "date"))
data4reg.year10.lag.0.1 <- data4reg.year10.lag.0.1[,-
                                                     which(colnames(
    data4reg.year10.lag.0.1)
                                                           %in% paste(
    substr(diff.lag0.lag1

      , 1, (nchar(diff.lag0.lag1)-1)),"0",
                                                                    sep="
    "))]

data4reg.year10.lag.0.1.2 <- merge(data4reg.year10.lag.0.1, data4reg.year10
    .lag.2[,c("emitter"

             , "date", diff.lag0.lag2)]
                                 , by = c("emitter", "date"))
data4reg.year10.lag.0.1.2 <- data4reg.year10.lag.0.1.2[,-
                                                     which(colnames(
    data4reg.year10.lag.0.1.2)
                                                           %in% paste(
    substr(diff.lag0.lag2

      , 1, (nchar(diff.lag0.lag2)-1)),"0",
                                                                    sep="
    "))]


```

```
data4reg.year10.lag.0.1.2.3 <- merge(data4reg.year10.lag.0.1.2, data4reg.
    year10.lag.3[,c("emitter"

                        , "date", diff.lag0.lag3)]
                                                   , by = c("emitter", "date"))
data4reg.year10.lag.0.1.2.3 <- data4reg.year10.lag.0.1.2.3[,-
                                                                        which(names(data4reg
    .year10.lag.0.1.2.3)
                                                                              %in% paste(
    substr(diff.lag0.lag3

          , 1, (nchar(diff.lag0.lag3)-1)),"0",
                                                                                          sep
    ="")))]

data4reg.year10.lag.0.1.2.3.4 <- merge(data4reg.year10.lag.0.1.2.3,
                                           data4reg.year10.lag.4[,c("emitter",
    "date", diff.lag0.lag4)]
                                                   , by = c("emitter", "date"))
data4reg.year10.lag.0.1.2.3.4 <- data4reg.year10.lag.0.1.2.3.4[,-
                                                                              which(names(
    data4reg.year10.lag.0.1.2.3.4)
                                                                                    %in%
    paste(substr(diff.lag0.lag4

                 , 1, (nchar(diff.lag0.lag4)-1)),"0",

        sep="")))]

# comment since setdiff is empty
# data4reg.year10.lag.0.1.2.3.4.5 <- merge(data4reg.year10.lag.0.1.2.3.4,
#                                           data4reg.year10.lag.5[,c("emitter
    ", "date", diff.lag0.lag5)]
#                                                  , by = c("emitter", "date"))

#lag-variables - stepwise linear regression####

  # initialisation
  data4reg.year10 <- data4reg.year10.lag.1
  lagName <- "lag1"
  missingPercent <- 25
```

```
  outlierFlag <- "mitOutlier"

  #cleaning after merging####

  # removing observations without ratings
  data4reg.year10.step1 <- data4reg.year10[!is.na(data4reg.year10$PD),]

  # Index of row to exclude
  index.row.rm <- NULL
  for (i in 1:nrow(data4reg.year10.step1)){
    nr.non.missing <- sum(!is.na(data4reg.year10.step1[i,-c(1,2,3)]))
    if(nr.non.missing < 4){
      index.row.rm <- c(index.row.rm, i)
    }
  }

  # adapting the data basis
  data4reg.y10.rm <- data4reg.year10.step1[-index.row.rm,]

# # for tex
# nrow(data4reg.y10.rm)
# unique(data4reg.y10.rm$emitter)
# unique(data4reg.y10.rm$date)
# setdiff(unique(data4reg.year10.lag.0$emitter), unique(data4reg.y10.rm$
    emitter))

  #missng value analysis####

  # filling level
  fill.rate <- numeric(ncol(data4reg.y10.rm[,-c(1,2,3)]))
  for (i in 1:ncol(data4reg.y10.rm[,-c(1,2,3)])){
    fill.rate[i] <- sum(is.na(data4reg.y10.rm[,(i+3)]))
    #fill.rate[i] <- (sum(is.na(data4reg.y10.rm[,(i+3)])) / nrow(data4reg.
    y10.rm[,(i+3)])) * 100
    names(fill.rate)[i] <- colnames(data4reg.y10.rm)[i+3]
  }
  fill.rate <- (fill.rate/nrow(data4reg.y10.rm)) * 100
  fill.rate.sel <- fill.rate[fill.rate < missingPercent]
#   cor.path <- "D:\\Lavri document\\phd thesis\\data\\psvag\\programm\\
    PSVaG_R_Prog\\results\\missing"
```

```
#   write.csv2(fill.rate, file = paste(cor.path, "\\fill_rate.csv", sep=""))

  #correlation analysis####

  # calculating correlation
  cor.meth1 <- cor(data4reg.year10.step1[,names(fill.rate.sel)], use = "pairwise.complete.obs")
  cor.meth2 <- cor(data4reg.year10.step1[,names(fill.rate.sel)], use = "na.or.complete")

#   # exporting correlation report
#   cor.path <- "D:\\Lavri document\\phd thesis\\data\\psvag\\programm\\PSVaG_R_Prog\\results\\corr"
#   write.csv2(cor.meth1, file = paste(cor.path, "\\cor_lag1_less10_meth1.csv", sep=""))
#   write.csv2(cor.meth2, file = paste(cor.path, "\\cor_lag1_less10_meth2.csv", sep=""))

  # removing correlated variables
  vor.corr.rm <- c("Capital...1.Yr.Growth", "Operating.Margin", "Pretax.Margin", "Return.on.Common.Equity"
                   , "Annualized.Return.on.Common.Equity", "Operating.ROE", "Actual.Sales.Per.Employee"
                   , "LT.Debt.Common.Equity", "Total.Debt.Common.Equity", "Total.Debt.Tangible.Book.Value"
                   , "Total.Debt.Total.Equity", "Net.Debt.Shareholders.Equity", "Total.Debt.Total.Assets"
                   , "Tangible.Common.Equity")
  vor.corr.rm <- paste(vor.corr.rm, "lag0", sep="_")
  corr.sel.var <- setdiff(names(fill.rate.sel), vor.corr.rm)
  # model.var <- names(step$coefficients)[-1]
  # corr.sel.var <- model.var

  # transforming depending variable
  logit.PD <- log(data4reg.year10.step1$PD/(1-data4reg.year10.step1$PD))
  data4reg <- data4reg.year10.step1[,c("PD", corr.sel.var)]
  rownames(data4reg) <- paste(data4reg.year10.step1$emitter, data4reg.year10.step1$date, sep="__")
  data4reg$PD <- logit.PD; colnames(data4reg)[colnames(data4reg) == "PD"] <- "logit.PD"
```

```

# removing rows without observation
row2rm.all <- NULL
row2rm.any <- NULL
for (i in 1:nrow(data4reg)){
  if (all(is.na(data4reg[i,-1]))){
    row2rm.all <- c(row2rm.all, i)
  }
  if (any(is.na(data4reg[i,-1]))){
    row2rm.any <- c(row2rm.any, i)
  }
}
data4reg <- data4reg[-row2rm.any,]
#  # calculating correlation
#  cor.meth1.after <- cor(data4reg, use = "pairwise.complete.obs")
#  cor.meth2.after <- cor(data4reg, use = "na.or.complete")

#  #linear regression####
#
#  # selecting outlier to exclude
#  outlierFlag <- "ohneOutlier"
#  outliers <- names(step$residuals[step$residuals < -2.50 | step$
   residuals > 2.50])
#  #exluding outliers
#  data4reg <- data4reg[-which(rownames(data4reg) %in% outliers),]
#
#  # fiting linear regression model
#  fit <- lm(logit.PD ~ ., data = data4reg, na.action=na.omit)
#
#  # performing stewise regression based on the fitted linear regression
   model
#  step <- stepAIC(fit, direction="both", na.action=na.omit, trace = 0)
#  summary(step)
#  print("Number of observations: "); nrow(step$model)
#  step.cvlm.10 <- cv.lm(data4reg, step, m=10, seed = 123456789, printit=F
   )
#  print("CV 10 mean squared error: "); attr(step.cvlm.10, "ms")
#  step.cvlm.20 <- cv.lm(data4reg, step, m=20, seed = 987654321, printit=F
   )
#  print("CV 20 mean squared error: "); attr(step.cvlm.20, "ms")
#
```

```
#   # saving plot in file
#   plot.path <- "D:\\Lavri document\\phd thesis\\data\\psvag\\programm\\PSVaG_R_Prog\\results\\plot"
#   pdf(paste(plot.path, "\\regPlot", lagName, outlierFlag, missingPercent, "OhneCorr.pdf", sep=""))
#   #pdf(paste(plot.path, "\\regPlot", lagName, "orig", missingPercent, "OhneCorr.pdf", sep=""))
#   # printing result analysis plots
#   op <- par(mfrow=c(1,2))
#   # ploting residuals
#   plot(step$model$logit.PD, stdres(step), xlab = "logit(PD)", ylab="Residuals")
#   abline(h=0);abline(h=1.96);abline(h=-1.96)
#
#
#   # computing predictions
#   preditc.fit <- predict(fit, se.fit=T)
#   pd.pred <- exp(preditc.fit$fit)/(exp(preditc.fit$fit) + 1)
#   pd <- exp(data4reg$logit.PD)/(exp(data4reg$logit.PD) + 1)
#   # ploting predictions against realisations
#   plot(pd,pd.pred, xlab="Realised PD", ylab="Predicted PD")
#   abline(a=0, b=1, type="l")
#   par(op)
#   # closing pdf.file
#   dev.off()
#
# # to visualise
# # printing result analysis plots
# op <- par(mfrow=c(1,2))
# # ploting residuals
# plot(step$model$logit.PD, stdres(step), xlab = "logit(PD)", ylab="Residuals")
# abline(h=0);abline(h=1.96);abline(h=-1.96)
#
#
# # computing predictions
# preditc.fit <- predict(fit, se.fit=T)
# pd.pred <- exp(preditc.fit$fit)/(exp(preditc.fit$fit) + 1)
# pd <- exp(data4reg$logit.PD)/(exp(data4reg$logit.PD) + 1)
# # ploting predictions against realisations
# plot(pd,pd.pred, xlab="Realised PD", ylab="Predicted PD")
```

```
# abline(a=0, b=1, type="l")
# par(op)


#principal component analysis####

# # analysisng variance of the variable
# library(caret)
# nzv <- nearZeroVar(data4reg.y10.rm, saveMetrics=T)
# print(range(nzv$percentUnique))
# head(nzv)

# cor.pca <- cor(data4reg[,-1])
#
# find.big.corr <- function(corCol){
#   which(corCol == 1)
# }
#
# #find.big.corr(.7, cor.pca[,3])
#
# apply(cor.pca, 2, find.big.corr)
#
# sum(!is.na(data4reg$Return.on.Common.Equity_lag1))
# sum(!is.na(data4reg$Annualized.Return.on.Common.Equity_lag1))

# remove Annualized.Return.on.Common.Equity_lag1 because completely
    correlated with Return.on.Common.Equity_lag1
remove.cor <- which(colnames(data4reg) == paste("Annualized.Return.on.
    Common.Equity", lagName, sep="_"))

#var.vec <- apply(data4reg[,-c(1,2,3)], 2, var)
var.vec <- apply(data4reg[,-c(1,2,3,remove.cor)], 2, var)
#var.vec <- apply(data4reg[,-c(1,2,remove.cor)], 2, var)
#var.vec <- apply(data4reg[,-c(1,remove.cor)], 2, var)
var.vec.names <- names(var.vec[var.vec > 1])

# performing PCA
pred.pca <- prcomp(data4reg[,var.vec.names], scale = TRUE)
save(pred.pca, file=paste(prog.path, "\\predPca.RData", sep=""))

# saving scales
```

```
mean.all.pca <- apply(data4reg[,var.vec.names], 2, mean)
sd.all.pca <- apply(data4reg[,var.vec.names], 2, sd)
# saveRDS(mean.all.pca, paste(prog.path, "\\meanAllPca.rds", sep=""))
# saveRDS(sd.all.pca, paste(prog.path, "\\sdAllPca.rds", sep=""))
save(mean.all.pca, file=paste(prog.path, "\\meanAllPca.RData", sep=""))
save(sd.all.pca, file=paste(prog.path, "\\sdAllPca.RData", sep=""))

#print(pred.pca, digits=4)
summary(pred.pca)

# printing biplot
plot.path <- "D:\\Lavri document\\phd thesis\\data\\psvag\\programm\\PSVaG_R_Prog\\results\\plot"
pdf(paste(plot.path, "\\Biplot", missingPercent, "PCA.pdf", sep=""))
pred.pca.4.biplot <- pred.pca
attr(pred.pca.4.biplot$rotation, "dimnames")[[1]] <- rep("",length(attr(pred.pca.4.biplot$rotation, "dimnames")[[1]]))
biplot(pred.pca.4.biplot, expand = 1.3)
legend("topright", legend = "Vector of \n Input-Variables", bty = "n", col = "red", lty="solid", cex = .7)
dev.off()

# printing cum variances
pdf(paste(plot.path, "\\cumVariance", missingPercent, "PCA.pdf", sep=""))
screeplot(pred.pca, type="lines", npcs = length(attr(pred.pca$rotation, "dimnames")[[1]]), main="")
title(xlab="Principal Components")
dev.off()


# creating input-data for regression based on PCA
data4reg.pca <- merge(data4reg[,c(1,2)], as.data.frame(pred.pca$x), by="row.names")
row.names(data4reg.pca) <- data4reg.pca$Row.names
data4reg.pca <- data4reg.pca[,-c(1,3)]

# # selecting outlier to exclude
# outlierFlag.pca <- "ohneOutlier"
# outliers.pca <- names(step.pca$residuals[step.pca$residuals < -2.50 | step.pca$residuals > 2.50])
# #exluding outliers
```

```r
# data4reg.pca <- data4reg.pca[-which(rownames(data4reg.pca) %in% outliers.pca),]

# fiting linear regression model
fit.pca <- lm(logit.PD ~ ., data = data4reg.pca, na.action=na.omit)
#fit.pca <- lm(logit.PD ~ PC1+PC2+PC3+PC4+PC5+PC6+PC7+PC8, data = data4reg.pca, na.action=na.omit)

# saving pca-lm-model
save(fit.pca, file=paste(prog.path, "\\pcaLmModel.RData", sep=""))

# performing stewise regression based on the fitted linear regression model
step.pca <- stepAIC(fit.pca, direction="both", na.action=na.omit, trace = 0)
save(step.pca, file=paste(prog.path, "\\stepPca.RData", sep=""))
summary(step.pca)
print("Number of observations: "); nrow(step.pca$model)
step.pca.cvlm.10 <- cv.lm(data4reg.pca, step.pca, m=10, seed = 123456789, printit=F)
print("CV 10 mean squared error: "); attr(step.pca.cvlm.10, "ms")
step.pca.cvlm.20 <- cv.lm(data4reg.pca, step.pca, m=20, seed = 987654321, printit=F)
print("CV 20 mean squared error: "); attr(step.pca.cvlm.20, "ms")

# to visualise
plot.path <- "D:\\Lavri document\\phd thesis\\data\\psvag\\programm\\PSVaG_R_Prog\\results\\plot"
pdf(paste(plot.path, "\\regPlot", lagName, outlierFlag, missingPercent, "PCA.pdf", sep=""))
#pdf(paste(plot.path, "\\regPlot", lagName, "orig", missingPercent, "PCA.pdf", sep=""))
# printing result analysis plots
op <- par(mfrow=c(1,2))
# ploting residuals
plot(step.pca$model$logit.PD, stdres(step.pca), xlab = "logit(PD)", ylab="Residuals")
abline(h=0);abline(h=1.96);abline(h=-1.96)


# computing predictions
preditc.fit <- predict(fit.pca, se.fit=T)
```

```
pd.pred <- exp(preditc.fit$fit)/(exp(preditc.fit$fit) + 1)
pd <- exp(data4reg.pca$logit.PD)/(exp(data4reg.pca$logit.PD) + 1)
# ploting predictions against realisations
plot(pd,pd.pred, xlab="Realised PD", ylab="Predicted PD")
abline(a=0, b=1, type="l")
par(op)
# closing pdf.file
dev.off()

# -- # -- # -- # -- # -- # -- # -- #

# printing result analysis plots
op <- par(mfrow=c(1,2))
# ploting residuals
plot(step.pca$model$logit.PD, stdres(step.pca), xlab = "logit(PD)", ylab="
    Residuals")
abline(h=0);abline(h=1.96);abline(h=-1.96)


# computing predictions
preditc.fit <- predict(fit.pca, se.fit=T)
pd.pred <- exp(preditc.fit$fit)/(exp(preditc.fit$fit) + 1)
pd <- exp(data4reg.pca$logit.PD)/(exp(data4reg.pca$logit.PD) + 1)
# ploting predictions against realisations
plot(pd,pd.pred, xlab="Realised PD", ylab="Predicted PD")
abline(a=0, b=1, type="l")
par(op)


#glmm####


# create table for lmm
date.vec <- unlist(lapply(row.names(data4reg.pca), extract.date))
data4reg.lmm <- cbind(date.vec,data4reg.pca)

# # creating table without outliers
# data4reg.lmm.ohneOutlier <- data4reg.lmm[which(!(data4reg.lmm$date.vec %
    in% c("Dez.97", "Dez.98"))),]

# overview of obs in the time series
```

```
table(data4reg.lmm$date.vec)

# selecting relevant variables for regression and paste into lme-command
paste(names(step.pca$coefficients[-1]), collapse = " + ")

## performing lmm
#(lmer.model <- lmer(logit.PD ~ PC1 + PC3 + PC4 + PC5 + PC6 + PC8 + PC9 +
    PC10 + PC11 + PC12 + PC15 + PC16 + PC17 + PC18 + PC20
#                      + PC21 + PC22 + PC23 + PC24 + PC25 + PC26 + PC28 +
    PC30 + PC31 + PC32 + PC33 + PC34 + PC36 + PC37
#                      + (1|date.vec), data = data4reg.lmm))

# performing lmm less 25%
(lmer.model <- lmer(logit.PD ~ PC3 + PC5 + PC9 + PC11 + PC12 + PC16 + PC17
    + PC20 + PC21 + PC22 + PC23 + PC24 + PC26 + PC27
                       + PC28 + PC30 + PC33 + PC34 + PC36 + PC7 + (1|date.
    vec), data = data4reg.lmm))

# # performing lmm less 25% with outlier
# (lmer.model <- lmer(logit.PD ~ PC3 + PC5 + PC9 + PC11 + PC12 + PC16 +
    PC17 + PC20 + PC21 + PC22 + PC23 + PC24 + PC26 + PC27 +
#                          PC28 + PC30 + PC34 + PC35 + PC36 + PC7 + PC8 + (1|
    date.vec), data = data4reg.lmm))


summary(lmer.model)

# saving lmm-model
save(lmer.model, file=paste(prog.path, "\\lmerModel.RData", sep=""))

# estimating r-squared
r.squared.lm <- 1-var(residuals(step.pca))/(var(model.response(model.frame(
    step.pca))))
r.squared.lmm <- 1-var(residuals(lmer.model))/(var(model.response(model.
    frame(lmer.model))))
r.squared.lm ; r.squared.lmm

# estimating AIC and BIC
logLik(step.pca) ; logLik(lmer.model)
AIC(step.pca, lmer.model) ; BIC(step.pca, lmer.model)
```

```
# printing result analysis plots
plot.path <- "D:\\Lavri document\\phd thesis\\data\\psvag\\programm\\PSVaG_R_Prog\\results\\plot"
pdf(paste(plot.path, "\\lmmPlotResidPred", missingPercent, ".pdf", sep=""))
op <- par(mfrow=c(1,2))
  # ploting residuals
  plot(data4reg.lmm$logit.PD, residuals(lmer.model), xlab = "logit(PD)", ylab="Residuals")
  abline(h=0);abline(h=1.96);abline(h=-1.96)

  # computing predictions
  preditc.fit <- predict(lmer.model, se.fit=T)
  pd.pred <- exp(preditc.fit)/(exp(preditc.fit) + 1)
  pd <- exp(data4reg.lmm$logit.PD)/(exp(data4reg.lmm$logit.PD) + 1)
  # ploting predictions against realisations
  plot(pd,pd.pred, xlab="Realised PD", ylab="Predicted PD")
  abline(a=0, b=1, type="l")
par(op)
dev.off()

# extract random effects
lmer.model.ranef <- random.effects(lmer.model)$date.vec

# replacing rownames by date as integer
row.names(lmer.model.ranef) <- unlist(lapply(row.names(lmer.model.ranef), trans.date))

# sorting random effects
lmer.model.ranef.4.ts <- lmer.model.ranef[order(row.names(lmer.model.ranef)),]

# cretaing time series
lmer.model.ranef.ts <- ts(lmer.model.ranef.4.ts, start=as.integer(min(row.names(lmer.model.ranef))))

# ploting time series
plot.ts(lmer.model.ranef.ts)

# fitting exponential smoothing
expsmooth.ranef.ts <- HoltWinters((100*lmer.model.ranef.ts), gamma=FALSE)
expsmooth.ranef.ts <- ets(lmer.model.ranef.ts, model="AAN")
```

```
# ploting forecast
plot(expsmooth.ranef.ts)

# creating real GDP table for Germany
Real.GDP.DEU <- subset(Real.GDP, ..LOCATION=="DEU", select=c(TIME, Value))
Real.GDP.DEU$Value <- as.numeric(Real.GDP.DEU$Value)
Real.GDP.DEU.ts <- ts(Real.GDP.DEU[-nrow(Real.GDP.DEU), "Value"], start=1997)

# creating real GDP table for Germany
Real.GDP.OECD <- subset(Real.GDP, ..LOCATION=="OECD", select=c(TIME, Value))
Real.GDP.OECD$Value <- as.numeric(Real.GDP.OECD$Value)
Real.GDP.OECD.ts <- ts(Real.GDP.OECD[-nrow(Real.GDP.OECD), "Value"], start=1997)

op <- par(mfrow=c(1,2))
  # ploting forecast
  plot(lmer.model.ranef.ts, axes =F)
  axis(side=1, at=seq(1997,2013, by=2))
  axis(side=2)
  # ploting real GDP
  plot(Real.GDP.DEU.ts, axes =F)
  lines(Real.GDP.OECD$TIME[-nrow(Real.GDP.OECD)],
        Real.GDP.OECD$Value[-nrow(Real.GDP.OECD)], col = "blue")
  axis(side=1, at=seq(1997,2013, by=2))
  axis(side=2)
par(op)

plot(lmer.model.ranef.4.ts[2:length(lmer.model.ranef.4.ts)],
    Real.GDP.DEU[-c(nrow(Real.GDP.DEU)-1,nrow(Real.GDP.DEU)), "Value"])

# performing forecast ahead in time
expsmooth.ranef.ts.ahead <- forecast.HoltWinters(expsmooth.ranef.ts, h=1)
expsmooth.ranef.ts.ahead <- forecast.ets(expsmooth.ranef.ts, h=1)

# printing result analysis plots
plot.path <- "D:\\Lavri document\\phd thesis\\data\\psvag\\programm\\PSVaG_R_Prog\\results\\plot"
```

```
pdf(paste(plot.path, "\\expSmoothPlotOrigPred", missingPercent, ".pdf", sep=""))
  # ploting the forecast
  plot.forecast(expsmooth.ranef.ts.ahead, main = "", ylab = "Random Effects"
                    , xlab = "Years", axes = F, frame.plot = T)
  legend(1997,.422,c("Original Time Series", "Predicted Time Series"), col=c("black","blue"),  lwd = c(4,4), text.col = c("black","blue"))
  axis(side=1, at=seq(1997,2013, by=1))
  axis(side=2)
  # adding forecast line in the plot
  #lines(attr(expsmooth.ranef.ts$fitted, "tsp")[1]:attr(expsmooth.ranef.ts$fitted, "tsp")[2], expsmooth.ranef.ts$fitted[,1], col="blue")
  lines(1997:2013, expsmooth.ranef.ts$fitted, col="blue")

dev.off()

# comparison of lm and glmm####

# comparing predictions
plot.path <- "D:\\Lavri document\\phd thesis\\data\\psvag\\programm\\PSVaG_R_Prog\\results\\plot"
pdf(paste(plot.path, "\\lmmVSlmPlotPred", missingPercent, ".pdf", sep=""))
op <- par(mfrow=c(1,2))
# computing predictions
preditc.fit <- predict(fit.pca, se.fit=T)
pd.pred <- exp(preditc.fit$fit)/(exp(preditc.fit$fit) + 1)
pd <- exp(data4reg.pca$logit.PD)/(exp(data4reg.pca$logit.PD) + 1)
# ploting predictions against realisations
plot(pd,pd.pred, xlab="Realised PD", ylab="Predicted PD", main = "Linear Regression", ylim=c(-.003,.09))
abline(a=0, b=1, type="l")

# computing predictions
preditc.fit <- predict(lmer.model, se.fit=T)
pd.pred <- exp(preditc.fit)/(exp(preditc.fit) + 1)
pd <- exp(data4reg.lmm$logit.PD)/(exp(data4reg.lmm$logit.PD) + 1)
# ploting predictions against realisations
plot(pd,pd.pred, xlab="Realised PD", ylab="Predicted PD", main = "Linear Mixed Model", ylim=c(-.003,.09))
abline(a=0, b=1, type="l")
```

```
par(op)
dev.off()

# comparing residuals
plot.path <- "D:\\Lavri document\\phd thesis\\data\\psvag\\programm\\PSVaG_R_Prog\\results\\plot"
pdf(paste(plot.path, "\\lmmVSlmPlotResid", missingPercent, ".pdf", sep=""))
op <- par(mfrow=c(1,2))
# ploting residuals
plot(step.pca$model$logit.PD, residuals(step.pca), xlab = "logit(PD)", ylab
    ="Residuals", main = "Linear Regression", ylim=c(-2.5,2.5))
abline(h=0);abline(h=1.96);abline(h=-1.96)

# ploting residuals
plot(data4reg.lmm$logit.PD, residuals(lmer.model), xlab = "logit(PD)", ylab
    ="Residuals", main = "Linear Mixed Model", ylim=c(-2.5,2.5))
abline(h=0);abline(h=1.96);abline(h=-1.96)
par(op)
dev.off()


# glmm with macroeco ####

# #creating real GDP table for Germany
# Real.GDP.DEU <- subset(Real.GDP, ..LOCATION=="DEU", select=c(TIME, Value
    ))
# Real.GDP.DEU$Value <- as.numeric(Real.GDP.DEU$Value)
# Real.GDP.DEU$TIME <- Real.GDP.DEU$TIME + 5

# # creating real GDP table for Germany
# Real.GDP.OECD <- subset(Real.GDP, ..LOCATION=="OECD", select=c(TIME,
    Value))
# Real.GDP.OECD$Value <- as.numeric(Real.GDP.OECD$Value)
# Real.GDP.OECD$TIME <- Real.GDP.OECD$TIME + 5
#
# creating real GDP table for Germany
Real.GDP.WLD <- subset(Real.GDP, ..LOCATION=="WLD", select=c(TIME, Value))
Real.GDP.WLD$Value <- as.numeric(Real.GDP.WLD$Value)
Real.GDP.WLD$TIME <- Real.GDP.WLD$TIME + 5

#Real.GDP.select <- Real.GDP.DEU
```

```
#Real.GDP.select <- Real.GDP.OECD
Real.GDP.select <- Real.GDP.WLD

# replacing rownames by date as integer
data4reg.lmm.macro <- data4reg.lmm
data4reg.lmm.macro$date.vec <- unlist(lapply(data4reg.lmm.macro$date.vec,
    trans.date))

# merging with macroeco
data4reg.lmm.macro <- merge(data4reg.lmm.macro, Real.GDP.select, by.x="date
    .vec", by.y="TIME")

# performing lmm
(lmer.macro.model <- lmer(logit.PD ~  PC3 + PC5 + PC9 + PC11 + PC12 + PC16
    + PC17 + PC20 + PC21 + PC22 + PC23 + PC24 + PC26 + PC27
                                        + PC28 + PC30 + PC33 + PC34 + PC36 + PC7
                              + Value + (1|date.vec), data = data4reg.lmm.macro))

summary(lmer.macro.model)

# performing lmm
(lmer.macro.model.lm <- lm(logit.PD ~  PC3 + PC5 + PC9 + PC11 + PC12 + PC16
     + PC17 + PC20 + PC21 + PC22 + PC23 + PC24 + PC26 + PC27
                                        + PC28 + PC30 + PC33 + PC34 + PC36 + PC7
                                       + Value , data = data4reg.lmm.macro))

summary(lmer.macro.model.lm)


# estimating r-squared
r.squared.lm <- 1-var(residuals(step.pca))/(var(model.response(model.frame(
    step.pca))))
r.squared.lmm <- 1-var(residuals(lmer.macro.model))/(var(model.response(
    model.frame(lmer.macro.model))))
r.squared.lm ; r.squared.lmm

### estimating AIC and BIC
#  logLik(step.pca) ; logLik(lmer.macro.model)
#  AIC(step.pca, lmer.macro.model) ; BIC(step.pca, lmer.macro.model)


```

```
#Performance related to rating-grades####
#----fit pca
preditc.fit.pca <- predict(fit.pca, se.fit=T)
pd.pred.pca <- exp(preditc.fit.pca$fit)/(exp(preditc.fit.pca$fit) + 1)

conv.pd.rating(pd.pred.pca[10])
pred.ratings <- unlist(lapply(pd.pred.pca, conv.pd.rating))
pred.ratings.df <- as.data.frame(pred.ratings)

data4reg.4.ratingcalc <- data4reg.y10.rm[,1:3]
rownames(data4reg.4.ratingcalc) <- paste(data4reg.4.ratingcalc$emitter, data4reg.4.ratingcalc$date, sep="__")

data4reg.4.ratingcalc.rating <- merge(data4reg.4.ratingcalc, PD2RatingGrade.moodys, by.x = "PD", by.y="Year_10", all.x)
rownames(data4reg.4.ratingcalc.rating) <- paste(data4reg.4.ratingcalc.rating$emitter, data4reg.4.ratingcalc.rating$date, sep="__")

merged.pred.and.orig <- merge(pred.ratings.df, data4reg.4.ratingcalc.rating, by="row.names")

error.vec <- abs(merged.pred.and.orig$pred.ratings - merged.pred.and.orig$Order)
table(error.vec)

#----fit lmm
preditc.fit <- predict(lmer.model, se.fit=T)
pd.pred <- exp(preditc.fit)/(exp(preditc.fit) + 1)

pred.ratings <- unlist(lapply(pd.pred, conv.pd.rating))
pred.ratings.df <- as.data.frame(pred.ratings)

data4reg.4.ratingcalc <- data4reg.y10.rm[,1:3]
rownames(data4reg.4.ratingcalc) <- paste(data4reg.4.ratingcalc$emitter, data4reg.4.ratingcalc$date, sep="__")

data4reg.4.ratingcalc.rating <- merge(data4reg.4.ratingcalc, PD2RatingGrade.moodys, by.x = "PD", by.y="Year_10", all.x)
rownames(data4reg.4.ratingcalc.rating) <- paste(data4reg.4.ratingcalc.rating$emitter, data4reg.4.ratingcalc.rating$date, sep="__")

```

```
merged.pred.and.orig <- merge(pred.ratings.df, data4reg.4.ratingcalc.rating
    , by="row.names")

error.vec.lmm <- abs(merged.pred.and.orig$pred.ratings - merged.pred.and.
    orig$Order)
table(error.vec.lmm)
```

Programm/PSVaGRPropV012.r

Bibliography

[1] Ifrs. URL http://www.ifrs.org.

[2] Basel Committee on Banking Supervision. Basel ii: International convergence of capital measurement and capital standards. *Bank of International Settlement*, June 2006.

[3] L. Mueller. *Bank-Raeuber: Wie kriminelle Manager und Politiker uns in den Ruin treiben.* Econ, 2010.

[4] B. Engelmann and R. Rauhmeier. *The Basel II Risk Parameters Estimation, Validation, and Stress Testing.* Springer Berlin Heidelberg, 2006.

[5] V. S. Cardoso, A. L. S. Guimaraes, H. F. Macedo, and J. C. C. O. Lima. Assessing corporate risk: a pd model based on credit rating. *ACRN Journal of Finance and Risk Perspectives*, Vol. 2:p. 51 – 58, 2013. URL www.acrn.eu/ressources/journals/jfrp201301e.pdf. shadow rating, logistic regression.

[6] D. Porath. Estimating probabilities of default for german savings banks and credit cooperatives. *Deutsche Bundesbank Discussion Paper*, 2(No 06/2004), 2004.

[7] D. A. Nuxoll. The contribution of economic data to bank-failure models. *FDIC Working Paper*, August 2003.

[8] A. Hamerle, T. Liebig, and H. Scheule. Forecasting credit portfolio risk. *Deutsche Bundesbank Discussion Paper*, 1(No 01/2004), 2004.

[9] W. Drobetz and S. Heller. What factors drive corporate credit ratings? evidence from german smes and large corporates. *Social Science Research Network*, March 2014. URL http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2392377.

[10] H. Andersen. Failure prediction of norwegian banks: A logit approach. *Norges Banks working papers*, (ANO 2008/2), February 2008. URL http://www.researchgate.net/publication/4921642.

[11] S. Canbas, A. Cabuk, and S. B. Kilic. Prediction of commercial bank failure via multivariate statistical analysis of financial structures: The turkish case. *European Journal of Operational Research*, 166:528–546, 2005.

[12] P. Espahbodi. Identification of problem banks and binary choice models. *Journal of Banking and Finance*, 15:53–71, 1991.

[13] J. Tatom. Predicting failure in the commercial banking industry. *MPRA Munich Personal RePEc Archiv*, (No. 34608), August 2011. URL http://mpra.ub.uni-muenchen.de/34608/.

[14] M. A. Boyacioglu, Y. Kara, and O. K. Baykan. Predicting bank financial failures using neural networks, support vector machines and multivariate statistical methods: A comparative analysis in the sample of savings deposit insurance fund (sdif) transferred banks in turkey. *Expert Systems with Applications*, 36:3355–3366, 2009. URL www.elsevier.com/locate/eswa.

[15] P. R. Kumar and V. Ravi. Bankruptcy prediction in banks and firms via statistical and intelligent techniques - a review. *European Journal of Operational Research*, 180:1–28, 2007.

[16] Bafin. URL https://www.bafin.de/SharedDocs/Veroeffentlichungen/DE/Rundschreiben/rs_1104_va_anlagers.html?nn=2818068#doc2675992bodyText10.

[17] Bad bank. URL http://www.investopedia.com/terms/badbank.asp.

[18] Kwg. URL http://dejure.org/gesetze/KWG/1.html.

[19] Basel Committee on Banking Supervision. Basel iii: A global regulatory framework for more resilient banks and banking systems. *Bank for International Settlements*, 2010.

[20] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics, 2001.

[21] H. Toutenburg. *Lineare Modelle*. Physica-Verlag, 1992.

[22] Alan Agresti. *Categorical Data Analysis*. New York: Wiley, 1990.

[23] L. Fahrmeir and G. Tutz. *Multivariate Statistical ModeModel Based on Generalised Linear Models*. New York: Springer-Verlag, 2. ed. edition, 2001.

[24] R. J. Hyndman, A. B. Koehler, R. D. Snyder, and S. Grose. A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, 18:439 – 454, 2002.

[25] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL http://www.R-project.org/.

[26] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0.

[27] Gregory R. Warnes, Ben Bolker, Gregor Gorjanc, Gabor Grothendieck, Ales Korosec, Thomas Lumley, Don MacQueen, Arni Magnusson, Jim Rogers, and others. *gdata: Various R Programming Tools for Data Manipulation*, 2015. URL http://CRAN.R-project.org/package=gdata. R package version 2.16.1.

[28] John H. Maindonald and W. John Braun. *DAAG: Data Analysis and Graphics Data and Functions*, 2015. URL http://CRAN.R-project.org/package=DAAG. R package version 1.22.

[29] Jose Pinheiro, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, and R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*, 2014. URL http://CRAN.R-project.org/package=nlme. R package version 3.1-117.

[30] Alexandra Kuznetsova, Per Bruun Brockhoff, and Rune Haubo Bojesen Christensen. *lmerTest: Tests in Linear Mixed Effects Models*, 2015. URL http://CRAN.R-project.org/package=lmerTest. R package version 2.0-29.

[31] R. J. Hyndman and Y. Khandakar. Automatic time series forecasting: the forecast package for r. *Journal of Statistical Software*, 26:1–22, 2008.

[32] J. A. Nelder P. McCullagh. *Generalized Linear Models*. Chapman and Hall/CRC, 1989.