TECHNISCHE UNIVERSITÄT WIEN
Vienna University of Technology

Diplomarbeit

Modelling and stress testing of the probability of default with the approach of robust regression

Ausgeführt am Institut für

Stochastik und Wirtschaftsmathematik

der Technischen Universität Wien

unter Anleitung von

Ao.Univ.-Prof.Dipl.-Ing.Dr.techn. Peter Filzmoser

durch

Elisabeth Jakupec

Otto-Herschmanngasse 4/2/11
1110 Wien

Danksagung

Nachdem Sie der erste waren, der diese Arbeit gelesen hat möchte ich mich bei Ihnen auch als erstes bedanken. Vielen Dank Herr Prof. Filzmoser für die Betreuung, die Unterstützung und die Geduld, für die Aufgeschlossenheit und für die Freiheit, die ich beim Schreiben hatte.

Danke liebe BAWAG P.S.K. für die Daten, die Begleitung durch das Studium und die Möglichkeit dieses Thema aufzugreifen. Vielen Dank Georg für die Idee, die Unterstützung und alles was ich lernen durfte.

An meine Familie und an meine Freunde, die sich wundern was ich eigentlich die ganze Zeit mache: Danke für euer Verständnis und die Freude während meines Studiums. Danke Mama und Papa für euren Rückhalt und das mitfiebern bei jeder Prüfung. Danke Tulio für deine Unterstützung und die Motivation.

Danke Caro, Lia, Ivanna und Lili für das durchhalten und die gemeinsame Studienzeit.

Abstract

The potential impact of macroeconomic developments and crises scenarios on credit risk is not only from regulatory perspective, but also from an entrepreneurial point of view crucial for a rigorously and sales-oriented risk management. Basis of this thesis is a logistic regression model for the probability of default (PD) of a retail portfolio. The PD model on product level is evaluated and improved in terms of robustness and information criterion. The effect of different macroeconomic scenarios on the estimated PD is evaluated, using multivariate regression. The aim of the thesis is to identify and interpret macroeconomic risk drivers to recognize and manage credit risk in a timely manner.

Zusammenfassung

Sowohl aus aufsichtsrechtlicher, als auch aus unternehmerischer Sicht ist die Auswirkung makroökonomischer Szenarien auf das Kreditrisiko einer Bank wesentlich für ein gründliches Risikomanagement. Basis dieser Arbeit ist ein logistisches Regressionsmodell für die Ausfallwahrscheinlichkeit (PD[1]) eines Retail Portfolios. Das PD-Modell wird auf Produktebene hinsichtlich seiner Robustheit und Informationskriterien untersucht und optimiert. Mittels multivariater Regression wird anschließend die Auswirkung verschiedener makroökonomischer Szenarien auf die geschätzte PD analysiert und interpretiert. Ziel ist es die makroökonomischen Treiber zu erkennen und mit Hilfe dieser Information das Risiko frühzeitig erkennen und steuern zu können.

[1]Probability of Default

Contents

1 Introduction

1.1 Default, defaulted, has been defaulted...

Cheap money. It is the outcome from the low interest rate environment in which we operate at the moment and at the same time it is the cause of the financial crisis we are still fighting our way out. Eight years and more than 3 trillion euro later it can be said that the latest crisis of the financial markets wasn't a liquidity crisis, but rather the result of aggravated credit risk and misjudgment of counterparty risk. Due to the low interest rate policy pursued by the US FED it was possible to take out a loan for low-income households with poor credit rating. The effects are now known: an oversupply of money and finally an economic misery.

The financial industry and especially the banking authorities drew their conclusions and turned the focus to credit ratings and counterparty risk. Assessing the creditworthiness and the models in connection with it have become considerably more importance. The models used have to be more reliable than ever, which makes the probability of default (PD) a key indicator not only within the credit approval process but for the assessment of survivability of a bank. Meanwhile, credit institutions owe almost ninety percent of the deposited equity to credit risk thus a significant part to counterparty risk.

A further consequence was implemented by the European Banking Authority (EBA). 124 of the major European banks were committed to participate in a EU-wide stress test, where the effect of macro economic stress scenarios on capital thresholds and related key figures was investigated and verified. On the test stand: the PD. From this obvious and right consequence can therefore be concluded that the reliability of the data and the resulting models are of central significance for the banking industry and everything or everyone it effects.

Since models and data play a major role for business decisions and their consequences, the assumptions and limitations under which statistical models evolve must not be made injudicious. Even if a model performs the most complex calculations, misguided assumptions lead to wrong results. Despite the fact that the applicable legislative texts and regulations contain hundreds of articles concerning the development and the validation of the applied methods and models, the problem of robustness is not attended and certainly not overcome. The consequences are not negligible, as the sensitivity to outliers affects the ability of a model to resist the modifications and adjustments of the initially stable structure. The question is, how reliable can a result be if the calculation is not robust?

1.2 What would we like to know?

The EBA stress test is no longer just a compulsory exercise. It developed into a fixed component of the modern risk management. In the performance, PD's resulting from existing models are exposed to prescribed macroeconomic stress scenarios. Using

PD's deriving from models which aren't robust gives rise to the assumption that the stress test results forfeit reliability - which leads to the following question:

How does the application of robust estimation affect the influence of macroeconomic risk factors on the PD of a specific Austrian retail portfolio during a stress test?

1.3 The journey is the reward

The generous provision of real anonymized data containing the private customer segment of a portfolio, by BAWAG P.S.K. makes it possible to base the analyses and conclusions in this thesis on genuine values. The aim is to develop a robust model for forecasting the probability of default with practically applied methods of the banking industry. Building on that, a robust approach should demonstrate an appropriate alternative for estimating a stressed PD with macroeconomic risk factors. The profound theoretical basics in conjunction with the results of the practical application should give information on the implementation of robust estimators.

1.4 The methodological approach and the structure of the thesis

In the second chapter the used methods are worked up theoretically.

The topic of modeling the probability of default is handled in subsection 2.1 with the concept of default risk and a description of developing a scoring model. The section containing the regulatory requirements shows the delimitation of the topic within the regulatory framework.

Section 2.2 explains the requirements and preconditions of the EBA stress test and how macroeconomic variables can define a specific scenario.

In Section 2.3 the used methods are described in theoretical context with the help of some very general examples. References to practical applications which don't derive from the stated literature, can be attributed to the business applications and methodological guidelines of BAWAG P.S.K.

The third chapter shows the described methods applied to a real data sample provided by BAWAG P.S.K. The data underly real-life restrictions and are therefore not the ideal choice for the applied methods. Nevertheless, the most significant differences between the evaluated methods can be seen.

Chapter 4 gives a short conclusion of the results from the previous chapter. The applied R-Codes are provided in the Appendix.

2 Modeling the probability of default

One key growth area of the last 30 years, both in science and management is the prediction of financial risk and thus the default risk. In consumer lending this forecasting is applied in the form of credit scoring. This chapter should give an overview of the basic terms in conjunction with the probability of default as a ratio in risk management and as target variable of the underlying models in consumer lending.

2.1 Default risk - definition and concept

Generally, default risk describes the risk that a consumer is not able to meet his obligation of paying his debts. The risk management is interested in measuring and steering this hazard. To do so, a precise definition of "defaulting" is necessary.

Let

$$c = (c_1, \ldots, c_n)$$

be the vector of the customers of the considered portfolio with n describing the number of all customers. The customer c_i is part of the portfolio, if he has at least one product p_{ij} with $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, m\}$ at the considered point in time. The vector

$$p_i = (p_{i1}, \ldots, p_{im})$$

therefore contains all products of customer i. The products can be of different types, but all of them are credit products, i.e. loans, credit cards or current accounts.
c_i and p_i can change throughout the time as if a customer doesn't have a credit product he is not part of the portfolio and every customer can change his product range from time to time. $t = (t_1, \ldots, t_k)$ is the vector of points in time the portfolio is observed. Figure 1 is an example for the portfolio composition of the customers c_1, c_2 and c_3 at $t_1, \ldots, t_4$.

A customer is defined by his products on the one hand, that means the number of his products overall, the different types of products, the number of different product types, the number of products within the types, how often these combinations change and the lifetime of his products. On the other hand, these products and therefore the customer is defined by his limit and his exposure. Let $(l_{i1}, \ldots, l_{im})$ be the vector containing the limits for each product p_{ij} of the customer c_i with

$$l_i = \sum_{j=1}^{m} l_{ij}$$

the total limit that customer c_i has. Thus l_i consists of all account limits, credit card limits and loan sums c_i can use.

The exposure of a product is the maximum of the actual utilized amount and the total limit. For a current account or a credit card, the utilized amount is the debit

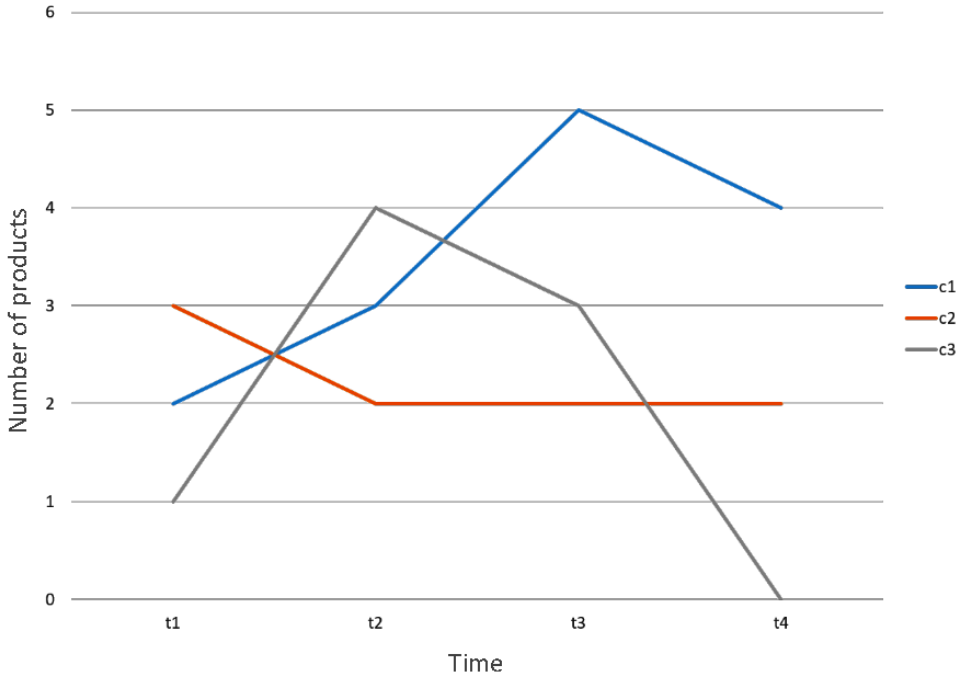

Figure 1: Example for the portfolio composition

balance. For a loan, the exposure is the loan sum minus the already repaid amount. Analogous to the notation for the limit is $(e_{i1}, \ldots, e_{im})$ the vector of the exposure for each product p_{ij} and

$$e_i = \sum_{j=1}^{m} e_{ij}$$

the total exposure of c_i. For defining the default scenario, one must be mindful of the fact that e_{ij} can be greater than l_{ij} for any product. The event of default describes the point in time from which a customer is not able to pay his debts. When is this moment? Based on the legal situation and the methods of practice, this event is subject to materiality thresholds to the time aspect as well as the level of debt. Thus the default relates to the following variables:

- c_i – as stated in the previous sentence
- p_i – a customer is defined by the products he owns
- l_i – the sum of the limits is the financial scope in which a customer can move
- e_i – compared to the limit, this variable provides information if the customer is within his scope or not
- t_h – the time component is substantial in many respects as the default event has a beginning and an end and the customer is observed over the time at different points t_h with $h \in \{1, \ldots, k\}$

Summarized this means that a customer is in default, if the following two conditions are fulfilled:

1. c_i is past due more than $\boldsymbol{d}$ days on a significant obligation.
2. The significant obligation exceeds the customer's total limit by $\boldsymbol{q}$ euros and $\boldsymbol{p}$ percent of the total limit.

Since the total limit is the permitted scope in which a debtor can move, the significant obligation is the exceeding amount $e_i - l_i > 0$. The begin of the default for c_i is defined as follows:

Definiton 2.1

i) $e_i - l_i > \boldsymbol{q}$

ii) $\frac{e_i - l_i}{l_i} > \boldsymbol{p}$

iii) i) and ii) are satisfied for $\boldsymbol{d}$ *consecutive days*

The date of day $\boldsymbol{d}$ is then defined as the beginning of the default. The end of the default is dependent from a lot of regulatory requirements and internal bank processes (dunning system, operations, payment plan, prolongation, etc.) and not relevant for the further discussions of this thesis except for the fact that the default period extends from the beginning to the end of the default.

During the modeling of the PD, the customers are regarded at the times $t_1, \ldots, t_k$. The default date gives a "natural" reference date for the choice of $\boldsymbol{t}=(t_1, \ldots, t_k)$ in the data history. Based on the forecast horizon the time lag of the data history is chosen accordingly for the estimation. As is usual, the probability of defaulting within one year is considered in this thesis.

Thus a customer is identified as a default in the data history, if he defaults within 365 days from the reference date. In Figure 2 for example, $c_1, \ldots, c_4$ are regarded at two reference dates.

- c_1 is not defaulted at all
- c_2 has its default event after the 365 days and is therefore not considered as default at reference date 1
- c_3 defaulted at reference date 2 and is consequently a default in the data history
- c_4 is obviously a default, as the customer defaulted between reference date 1 and 2

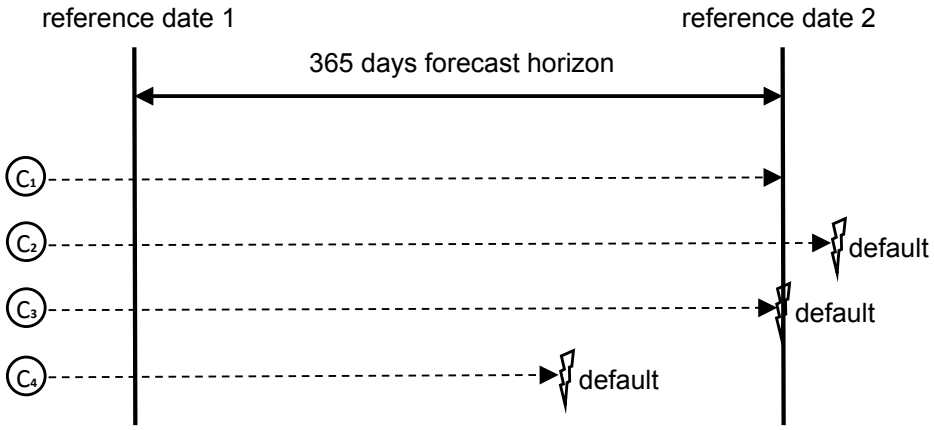

Figure 2: The forecasting horizon for the PD

2.2 Development of a scoring model

The determination of the PD is the basis for credit decisions and an appropriate pricing on the one hand and for a regular evaluation of the customer's behavior thus its rating on the other hand. The PD as a target figure forms the prerequisite for the sense of a rating model.

The development of an empirical, statistical model for the PD happens in three basic steps:

1. generation of the historical data base
2. selection of the variables
3. determination of the scoring function

In the following the individual steps are described.

2.3 Generation of the historical data base

An on empirical basis developed rating can only be as good as the underlying database.

For the data collection, the relevant segments and the level of consideration have to be defined. This thesis investigates a portfolio consisting of private customers. As described in Section 2.1, a customer owns credit products of different types. This view questions whether to model the PD for a customer or to build a model on product level. The aim is to evaluate the PD out of the customer's behavior. Because a customer is defined by his products, his behavior can be observed on these products. Building a model on product level for each type of products is well-suited for this purpose.

Time is a significant factor in the generation of the historical data base. The following questions have to be answered in this context:

What is the forecast horizon for the PD?
On which reference dates are the data observed?
How are the chronological layers defined?
How long is the development period?

The forecast horizon can be selected differently. If it is not specified, the model calculates the probability, that a customer ever defaults. Considering a mortgage loan with a maturity of 20 years, the necessary data might not be available since the model is based on historical data.

For determining the forecasted PD, one has to look as far into the future as the forecast horizon prescribes from a certain reference day.

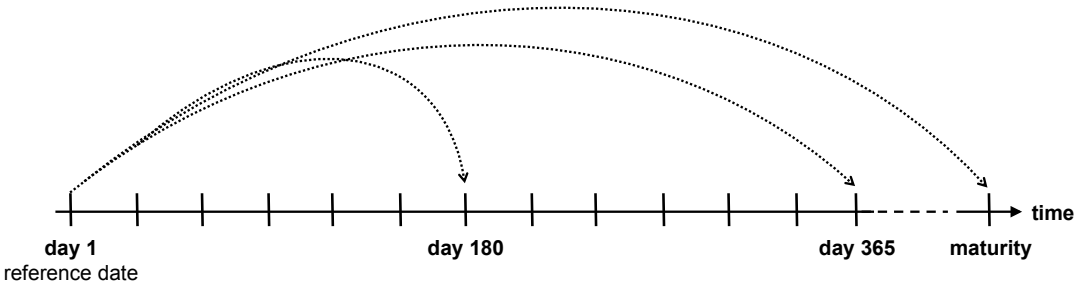

Figure 3: Necessary data history for the corresponding forecast horizon

In Figure 3 it is shown how the data history is dependent from the forecast horizon. Day 1 is the considered reference date in the past. Having a forecast horizon of 6 months (or 180 days) needs available data from 6 months before the beginning of the default event. A forecast horizon of the maturity dates needs, depending on the different maturities, available data from up to 20 years before the beginning of the default.

Within 20 years the processes in a bank might change. It is possible that these data weren't even recorded at that time. For a reliable and representative data base a forecast horizon of one year (i.e. 365 days) appears appropriate and is considered in the following analyses.

The choice of the reference dates is dependent from the availability of the data and what distance between the points in time makes sense for observing the desired effects. The various intervals have different advantages and disadvantages.

Insofar as monthly data are available, this interval seems as the most appropriate as monthly data contain sufficient genuine information.

interval	advantages	disadvantages
annually	- data sets only appear once within the forecast horizon - no need to consider chronological layers	- a lot of information isn't recognized - accounts that only exist for 11 months aren't in the database
quarterly	- contains more information than annually data - considers the economic cycle partially	- effects could be distorted (e.g.: holiday pays and bonuses) - chronological layers have to be defined
monthly	- corresponds to the interval of loan pay offs, salary payments, etc. - considers the economic cycle partially - sufficient number of data sets	- chronological layers have to be defined
daily	- information is considered completely - sufficient number of data sets	- there is no daily activity on every account (inconsistent information for one client) - chronological layers have to be defined

Table 1: Differences in the choice of the time horizon

As already shown in Figure 2, a customer is identified as a default in the data history, if he defaults within 365 days from the reference date. For determining defaults for the total set of data, the data set has to be considered for each reference date. In the total data set, all those cases are included which can be allocated to the

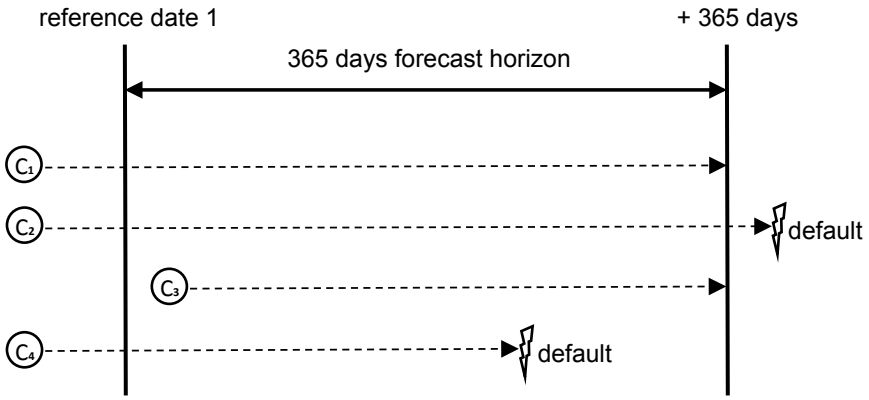

Figure 4: Chronological layers

relevant portfolio at the specific reference date. For the example shown in Figure 4, the following setting applies:

- c_1 is part of the portfolio at reference date 1 and not defaulted
- c_2 is part of the portfolio and also a "good"-case
- c_3 is not relevant, as the customer doesn't exist at reference date 1
- c_4 is relevant and a "bad"-case or default

The development period corresponds to the period of time within which the observations for modeling are collected. For the retail segment, a time series of at least one year and at most five years is provided. In terms of timeliness and to avoid process related changes within the data base, a period of one year is well suited.

Since the target value is the default, there is a minimum requirement for the number of defaults within the data base. The supervision provides at least 100 defaults in the modeling data base for the retail segment.

Another condition for the afterwards applied methods is the independence of the datasets. A customer may only be once in the modeling sample, otherwise the defaults may not be independent. To fulfill this condition, a random drawing is done, so that each customer is in the sample once over the time, with one product. This condition is another argument for modeling on product level and not on customer level.

For a reliable backtesting, the modeling sample is divided into a training (70%) and a test sample (30%) . The division is performed randomly. To ensure that the default rate in the training and the test sample is comparable, a stratification by the number of defaults is done. That means:

$$\textit{modeling sample} = \textit{training sample} + \textit{test sample}$$

2.3.1 Selection of the variables

At the beginning of the development of a rating model, a catalog of the criteria that need to be analyzed has to be specified (see Nösslinger and Thonbauer, 2004). For the analysis of the long list the previously determined modeling sample is used. With the help of single factor analysis (see Hosmer and Lemshow, 2000), the variables are tested for selectivity and significance. The result of the single factor analysis is a short list containing variables which are suitable for modeling. This variables are customer- and product-related ratios such as the number of dunnings a client got in the last six months or the number of months since the account opening.

In the first step of the variables selection, a complete list of all available variables is constituted. In conjunction with rating models, these variables usually contain information about the financial situation of the customer, turnover information of the considered account, dunning information and other product characterizations such as maturity, loan sum, etc.

The first check is for missing values and implausible values (e.g.: dummy values, values in different units, etc.). Missings or implausible values could occur due to technical reasons, for example if something goes wrong during the collection or the loading of the data. If it is possible the implausible values should be corrected.

To ensure an undistorted estimation, the missing values have to be handled neutral and be replaced by appropriate values:

a) for metrically scaled variables the missings can be replaced by a neutral value
b) for ordinally scaled variables the mode could be used or an extra category could be defined for the missing values

Nevertheless a threshold for a reasonable amount of missing values has to be defined. If a variable isn't determinable in 80% of all cases, the statistical valid handling of missing values can not be ensured and it has to be excluded from the data base.

In the next step of the univariate analyses the course of the variables is examined, i.e. the distribution of the defaults along the characteristics. For continuous variable a classification is done in order to get a comprehensible representation. Consider the integer variable "months since account opening" (MON_OPEN) for a sample containing 4000 data sets. A classification into percentile could look like this:

class	range
1	0-9
2	10-19
3	20-27
4	28-36
5	37-46
6	47-54
7	55-64
8	65-72
9	73-82
10	83-91
11	92-99
12	100-108
13	109-118
14	119-127
15	128-136
16	137-146
17	147-153
18	154-163
19	164-171
20	> 171

Table 2: Classification of a variable

For each category a default rate of the historical data is calculated and it is checked if the distribution reasonable concerning the processes. For example, if a customer has lots of dunnings it is expected that the default rate is higher than for little or no dunnings. The qualitative assessment of the frequency distribution is one of the minimum conditions for the variables of the shortlist.

One key task of rating models is to separate the good cases from the bad ones. Statistically this means that the model has to have a strong discriminatory power. In order to ensure a good selectivity of the multivariate scoring function, the individual variables have to demonstrate a certain degree of discriminatory power. A measure for this characteristic is the **Gini-coefficient** (GINI):

$$GINI = 2 * AUC - 1 \tag{2.1}$$

AUC is the "**Area-under-Curve**", a graphical measure for selectivity, derived from the ROC-curve. The **Receiver-Operating-Characteristic-curve** is a common form for presenting the discriminatory power of rating models and visualizes the dependence of the efficiency and the error rate. For each variable the frequency distributions in form of sensitivities (i.e. the right-positive-rate) and the false-positive-rate are determined. The cumulative frequencies of the "bad" cases (false-positive) are displayed on the Y-axis and the "good" cases (sensitivities) on the X-axis of the diagram. An example for the ROC-curve is shown in Figure 5.

Let us assume that each section of the ROC-curve represents one rating class, starting with the worst class in the origin.The gradient demonstrates the relation of the defaults to the non-defaults in each rating class. Since this relation is increasing normally, the ROC-curve has a concave course. If neighboring classes do not differ significantly concerning their default rate, this condition is violated. The ROC-curve for the perfect rating system would run vertically upwards to the point (0,100) and then straight to the right. If a system couldn't select between "good" and "bad" cases, its ROC-curve would correspond to the diagonal line.

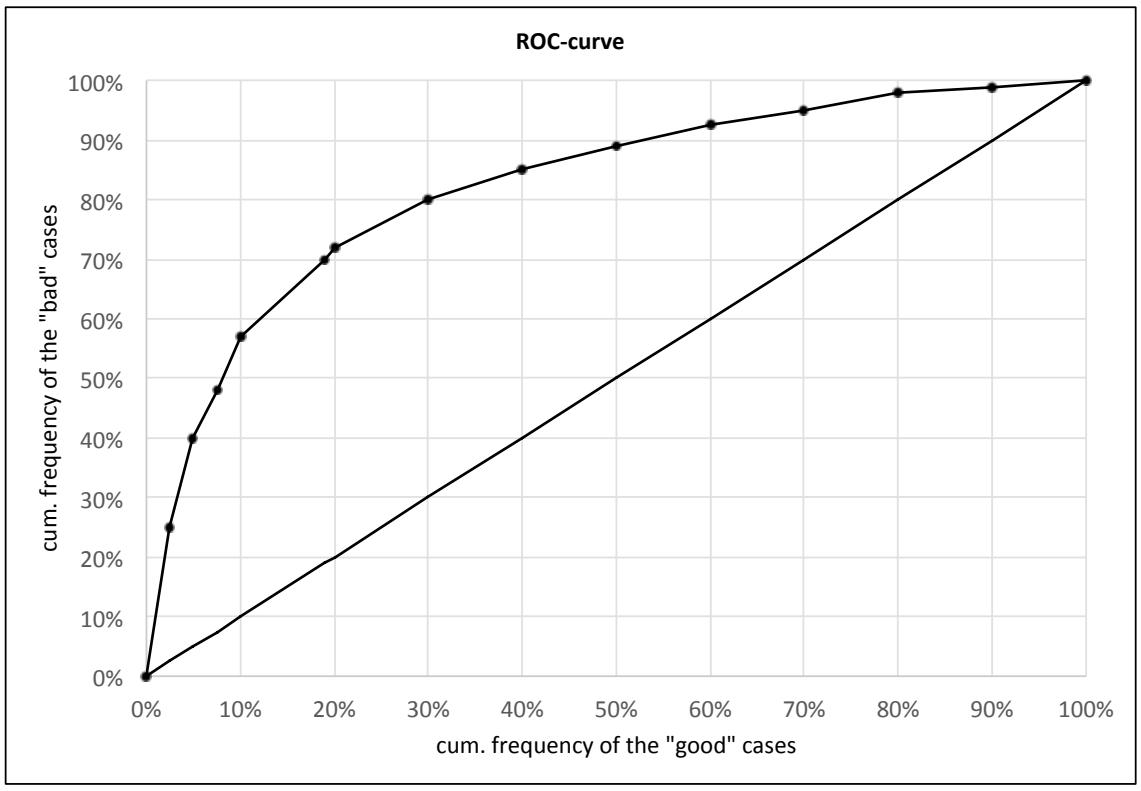

Figure 5: Example for the ROC-curve

In the univariate analysis, the ROC-curve is not determined for the different rating classes, but for the categories of a variable. Consider the variable "months since account opening" which is a positive integer. The categories could then be:

category	range
1	0-6
2	7-15
3	16-27
4	28-42
5	> 42

Table 3: Categories for the determination of the ROC-curve

The AUC is an aggregated key figure taking values from 0 to 1, which is derived directly from the ROC-curve. As the name says, this figure is the area under the ROC-curve and summarizes the discriminatory power of a rating system or a variable in one number. The AUC can be calculated with the trapezoidal method, which calculates the area as the sum of individual trapezoids under the ROC-curve.

$$AUC = \sum_{k=1}^{K} (ER_k - ER_{k-1}) \frac{HR_k + HR_{k-1}}{2} \tag{2.2}$$

with

$ER \ldots$ the error rate (values from Y-axis)
$HR \ldots$ the hit rate (values from X-axis)
$K \ldots$ number of categories or rating classes

For the end points applies:

$ER_0 = HR_0 = 0$
$ER_K = HR_K = 1$

An AUC with a value near to 1 refers to a very selective rating system, while a value of 0,5 equals a random experiment.

The univariate discriminatory power of a variable is no indication for its contribution for the selectivity of the multivariate rating model. Still a minimum separation efficiency must be given. For this, the GINI is calculated for the training sample on the one hand and for each year of the total time series on the other hand. For each variable the GINI is available as follows:

If one of the following criteria for the GINI is satisfied, the variable should be excluded from the shortlist:

- the training-GINI of a variable is $< 5\%$
- the difference between the total-GINI and the mean-GINI of all years is $> 10\%$

variable	2010	2011	2012	2013	2014	mean	total	training
MON_OPEN	19%	17%	21%	24%	18%	20%	21%	19%
TURNOVER	10%	11%	9%	12%	11%	11%	12%	12%
NR_DUNNINGS	-2%	4%	7%	-4%	6%	2%	6%	5%
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 4: GINI-coefficients for the selection of the variables

- the GINI changes the sign over the years

This would only leave the variable MON_OPEN from the given variables of the example above.

For the remaining variables, the linearity of the logit-PDs has to be verified. The basis for the verification is the logit model which will be explained more precisely in Section 4.1. Summarized, the logit model can be written as

$$\mathbb{P}_i = \mathbb{P}(y_i = 1) = F(\boldsymbol{\beta'} \cdot \boldsymbol{x}_i) = \frac{e^{\boldsymbol{\beta'} \cdot \boldsymbol{x}_i}}{1 + e^{\boldsymbol{\beta'} \cdot \boldsymbol{x}_i}}, \tag{2.3}$$

with

y_i ... the binary default variable of data set i
$\boldsymbol{x}_i$ ... the expression of data set i of the considered independent variable
$\boldsymbol{\beta}$ ... the parameter which captures the impact of a change in the characteristic on y_i
F ... the unknown distribution function

A linear relationship between the independent variable and the log odd is implied:

$$\text{log odd} = ln\left(\frac{\mathbb{P}_i}{1 - \mathbb{P}_i}\right) = \boldsymbol{\beta'} \cdot \boldsymbol{x}_i \tag{2.4}$$

The linearity assumption can be tested graphically by classifying the variable into equal groups as already done in Table 2. Then the historical default rate, i.e. the empirical log odd for each group is calculated in order to estimate a linear regression of the log odds on the mean values of the variable for each group. The result for the variable MON_OPEN is shown in Figure 6.

A non graphical way of testing the linearity assumption is the Box-Tidwell test which can be performed for each variable. In this case a logistic regression is performed with the binary default as dependent variable and the considered independent variable.

Additionally, an interaction term of the independent variable of the form

$$x \cdot ln(x)$$

is regarded in the logistic regression. In the next step the significance of the interaction term is verified. If the interaction is significant, the linearity assumption is not fulfilled and the variable needs to be transformed. If the interaction is not significant, it can be concluded that the variable behaves linear and can therefore be taken in the shortlist.

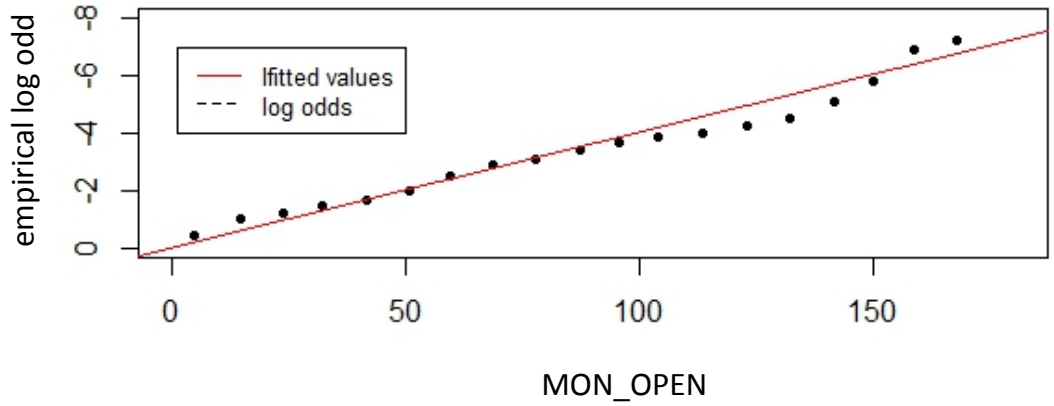

Figure 6: Relationship between log odd and MON_OPEN

Ultimately the expected dependence between the explanatory variables and the default probability has to be determined. The hypothesis could be:

1. an increase in the explanatory variable leads to an increase in the default probability
2. an increase in the explanatory variable leads to a decrease in the default probability

For the variable MON_OPEN, the second hypothesis would be formulated. This can be verified in Figure 6 as well. It can be seen that the behavior of the variable is as expected.

Summed up, the following minimum requirements have to be fulfilled by a variable of the shortlist:

- amount of missing values $\leq 20\%$
- plausible frequency distribution

- adequate GINI-coefficient
- linearity of the log odds
- working hypothesis satisfied

These criteria reduce the number of eligible variables significantly. The remaining ratios have a more or less marked similarity or correlation with each other. High correlation can lead to problems concerning the stability in the determination of the scoring function and the computing algorithm is not able to clearly determine the coefficients in the linear combinations of a variable. Therefore it is necessary to avoid high correlations between the variables in the model. Variable pairs with a correlation coefficient greater than 0.3 should rather not be combined in a model, as a rule of thumb. For the analysis of the correlation a hierarchical cluster analysis and Spearman's rank correlation are used.

The hierarchical cluster analysis combines variables into blocks. The blocks are formed so that the correlations within the cluster are very high and between the different clusters they are very low. As correlation measure, Spearman's rank correlation is suitable, as the determination is made along the ranks of the variable, not on the values. Besides, Spearman provides appropriate results for variables that aren't normally distributed and for small data samples. Therefore this method is applicable to not uniformly scaled data as well.

For the calculation the expressions of the variables need to be converted into rankings. The correlation coefficient results from the direct application of the linear correlation coefficient for metrical variables on the rank number (see Behr, 1999). For two variables let r_i be the rank of the observation i from X_i and s_i from Y_i. The difference of the ranks for observation i is

$$d_i = r_i - s_i.$$

The definition of the linear correlation coefficient r implies for the rank correlation coefficient r_s the following definition:

$$r_s = \frac{cov(r,s)}{std(r) \cdot std(s)} = \frac{\sum_{i=1}^{n}(r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{n}(r_i - \bar{r})^2}\sqrt{\sum_{i=1}^{n}(s_i - \bar{s})^2}} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{(n+1)(n-1)n} \tag{2.5}$$

with

$$-1 \;\leq\; r_s \;\leq\; 1.$$

With $\bar{r}$ and $\bar{s}$, the mean value of r_i resp. s_i, $i \in \{1, \ldots, n\}$. A value close to 0 means very low correlation between two variables.

2.3.2 Model construction

Different methods are used in modeling the PD. Basically there is a distinction between 3 different types of models in practice:

- heuristic models
- causal-analytical models
- empiric-statistical models

Heuristic models are rather qualitative systems like expert rating or fuzzy-logic-models. They can be applied to all rating segments and don't need an extensive data base.

Causal-analytical models like option-pricing-models and cash-flow-simulation-models are suitable for special lendings and listed companies as a data base. These models are not equally appropriate for all rating segments.

Empiric-statistical models need a sufficiently large data base in the development, especially concerning the defaults. Thus a considerably greater discriminatory power than heuristic models can be achieved. These models include methods like multivariate discriminant analysis, artificial neural networks, and regression models. The different approaches have various advantages and disadvantages. The suitability is closely related to the data requirements for the rating systems (qualitative and quantitative) and the different segments (e.g. retail or corporates segment).

In this thesis logistic regression models are considered and analyzed. They have the advantage that both quantitative and qualitative creditworthiness characteristics can be processed. Furthermore, the model results are mapped directly as a PD. This property facilitates the calibration in practice and allows to model binary dependent variables.

The defined shortlist in the variables selection is the starting basis for the logistic regression model. In general, a regression model contains 5 to 15 different explanatory variables. The estimated regression coefficients need to be statistically significant different from 0, i.e. a p-value $< 5\%$. Furthermore, the coefficients must have a sign appropriate to the working hypothesis.

The valuation of the different models is carried out with the test sample by the following criteria:

- discriminatory power (GINI-coefficient)
- distribution of the observations over the score values
- stability of the model (comparison of the results with the training sample)

A precise description with preconditions and the methodology is found in Section 4.1.

2.4 Regulatory requirements

The framework decided through the Basel Committee for Banking Supervisory (BCBS) "Basel II" and additional "Basel III" concerning the equity requirements form the basis for the regulation, the supervision and the risk management of a bank. Amongst others, these directives are found in the in Austria obligatorily applicable Capital Requirement Regulations (CRR) published by the European Commission.

The CRR regulates inter alia the standards for the default definition, the minimum requirements concerning the estimation of the PD, data requests and the conditions for the determination of the Risk Weighted Assets (RWAs) and consequently the allocation of equity. The legislative texts include numerous paragraphs and articles. This section summarizes some parts in conjunction with the estimation of the PD.

As already described in Section 2.1, a clear definition of the default event is necessary for modeling the probability in this context. Article 178 of Section 6 of the CRR describes the requirements to the default event. The following criteria have to be fulfilled to be defaulted:

- *the obligor is past due more than 90 days on any credit obligation, which is not secured by some specified property*
- *the underlying amount is material*
- *materiality of a credit obligation past due shall be assessed against a threshold, defined by the competent authorities*

The materiality refers to the thresholds $\boldsymbol{p}$ and $\boldsymbol{q}$ from Definition 2.1.

According to Article 4 of the CRR, the PD means *the probability of default of a counterparty over a one-year period* and institutions should estimate *PDs by obligor grade from long run averages of one-year default rates*[2]. As already elucidated in Section 2.3, a forecasting horizon of one year is appropriate and corresponds to the regulatory requirement.

One of a banks main goal of the PD estimation, is the allocation of capital for each outstanding claim. If the probability that a customer is going to default high, then more capital has to be allocated than for a customer who might not default with a strong probability. That means that a good forecast of the default risk is not only a supervisory regulation, but from a portfolio perspective it is much more the basis for the survival of a bank.

Basis for the equity requirement are the risk weighted assets (RWAs). According to Article 154 of the CRR, the RWAs for retail exposures should be calculated as follows:

$$RWA_{ij} = RW_j \cdot e_{ij}$$

[2] CRR: Section 6, Article 180

with

e_{ij} ... the exposure for product j from customer i
RW_j .. the risk weight of the claim

For the determination of the risk weight it has to be distinguished between the values of the PD. If the PD=1, i.e. for already defaulted exposures:

$$RW = max\{0;12{,}5 \cdot (LGD - EL_{BE})\}$$

EL_{BE} is the best estimate of the expected loss for the defaulted exposure (see CRR: article 181(1)). For not defaulted exposures or $0 < PD < 1$ the risk weight is depending on the probability that a customer will default, as well[3]:

$$RW = \left(LGD \cdot N\left(\frac{1}{\sqrt{1-R}} \cdot G(PD) + \sqrt{\frac{R}{1-R}}G(0,999)\right) - LGD \cdot PD\right) \cdot 12,5 \cdot 1,06$$

with

$N(x)$ ... the distribution function for a standard normal distributed random variable (i.e. the probability that a random variable with mean= 0 and variance= 1 is $\leq$ x)
$G(z)$ ... the inverse of the standard normal distribution function
R ... the correlation coefficient defined as

$$R = 0,03 \cdot \frac{1-e^{-35 \cdot PD}}{1-e^{-35}} + 0,16 \cdot \left(1 - \frac{1-e^{-35 \cdot PD}}{1-e^{-35}}\right)$$

As it can be seen, the PD makes a substantial contribution to the determination of the RWAs. The calculation is on the level of claims, that means for every product with exposure. The main recommendation of the Basel framework is, that the bank has to provide at least 8% of the RWAs as core capital.

Of course the supervisory regulations concerning default and the PD is much more extensive than the elaborated points of this section. The aim was it to receive an assessment of the regulatory importance of the PD estimation.

[3] CRR: Section 1 (ii), Article 154

3 Stress testing

The CRR doesn't only regulate the capital allocation and PD estimation on a microeconomic level. The crisis has taught us, that an economic collapse can come very quickly and even worse: without announcement. It is for this reason, that an annually supervisory stress test on banking institutes is prescribed by law.

3.1 The EBA stress test

In order to guarantee a strong common equity tier 1 (CET1 ratio) the European Banking Authority (EBA) performs a regular stress test with more than one hundred European banks. The test contains a baseline scenario which describes the expected economic developments and an adverse scenario that simulates the partial default of European government bonds. These scenarios should clarify the effects on capital requirements and the underlying estimated risk parameters such as the PD.

The stress test verifies if a bank is well enough equipped with equity. This implies that the CET1 ratio is the key indicator in this exercise. The ratio measures which proportion of the risk weighted assets needs to default until the liable equity of a credit institution is completely absorbed and consequently the risk of insolvency.

As already mentioned in the former section, the minimum requirement for the CET1 ratio is at 8%. In the stress test thresholds are defined for both scenarios. A bank passed the stress test[4] if the following limits were achieved (concerning the CET1 ratio):

- baseline scenario: $\boldsymbol{CET1\ ratio} \geq 8,0\%$
- adverse scenario: $\boldsymbol{CET1\ ratio} \geq 5,5\%$

The specified methodology of the European Banking Authority uses the following hierarchy.

As initial values the PDs resulting from the bank's internal models are considered. That means, the estimated probabilities which arise from the models specified in Section 2.3.2. It has to be determined on which level the starting values are required, i.e. if the stress test is performed on product level or customer level and how the database and consequently the starting values are defined regarding the segmentation, reference dates, etc. In Section 3.4 we amplified what has to be considered in particular concerning the starting values for estimating a stressed PD.

The second step of the hierarchy requires the application of the macroeconomic stress scenarios on the starting values. To make it clear: the estimation of a stressed PD. The objective is to find a model, that describes the impact of macro variables on probabilities of default.

[4] thresholds according to the EU-wide stress test in 2014

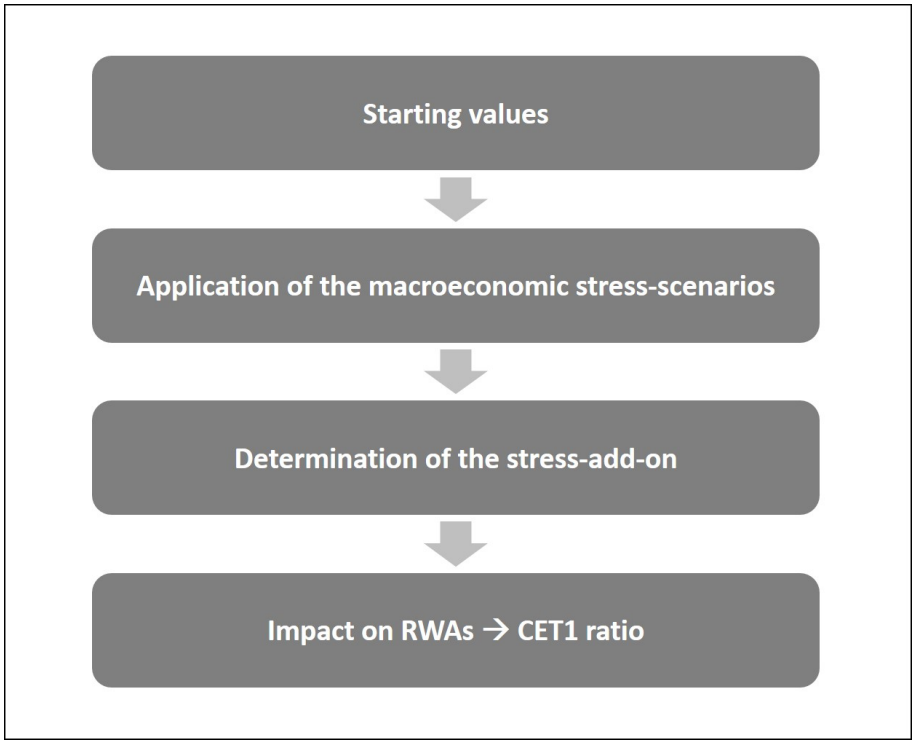

Figure 7: Methodological hierarchy of the EBA stress test

The determination of the stress-add-on means unequivocally the calculation of the difference between the predicted PD resulting from the banks internal models and the respective scenario (baseline and adverse). The difference should be added to the predicted PD in form of a margin. In this context it is assumed that an economic downturn leads to a higher PD.

Ultimately it is a matter of costs and how the stress scenarios influence the CET1-ratio. Initially the stressed PD is used as input parameter for the RWA calculation instead of the predicted PD from the internal models.

This hierarchy describes the procedure in the stress test roughly. This thesis doesn't deal with the last step of the hierarchy shown in Figure 7. The focus is on the estimation of the stressed PD. In the empirical part an example for a stressed PD model is shown.

3.2 Macroeconomic variables as risk factors

For predicting the creditworthiness or the probability of default of a borrower, certain risk factors are considered. Irrespective of which model is used for the estimation, the variables which are eligible as risk factors, have an explanatory function. This means that the probability of default should be dependent from the informa-

tion such a variable contains.

In paragraph 402 of the capital framework it is stated which risk factors should be used in a scoring model at minimum:

- Borrower risk characteristics
- Transaction risk characteristics

The borrower risk characteristics could be the *borrower type* and demographic information such as the *age*, the *occupation* or the *marital status*. Transaction risk characteristics include product information such as the *product type* or the *number of months since the account opening* and as well information like the *loan value*, the *account balance* or *limit increases*.

In the stress test it is assessed which influence macro economic variables have on the PD of a borrower. According to the Basel framework, the essential characteristic of the procedure of stress testing is the change of the risk factors for predicting the PD. Instead of borrower or transaction risk characteristics, macro economic factors represent the explanatory variables.

Since the choice of the variables for the short long list is greatly dependent on the design of the baseline and the adverse scenario, the following selection arises from the EU wide stress test in 2014:

- **Real Gross Domestic Product (GDP):** The GDP is the aggregated national demand of a country. It is evaluated at market prices and is related to the sales of companies. A low GDP means it is harder for companies to generate income through sales. It is a measure of the macroeconomic activity.
- **Consumer Price Index (CPI):** The CPI is the average price of all consumed goods. It must be noted that the number of consumed goods is not identical with the number of produced goods.
- **Unemployment Rate (UR):** The UR is closely related to the GDP and CPI. The harder it is for companies to generate income, the greater the possibility to lose the job and with little or without income, the number of consumed goods declines.
- **Residential Property Prices (RPP):** Increasing RPPs cause a higher indebtedness of the private households and weaken the purchasing power.
- **Equity Prices (EP):** Fluctuations on the equity market relate to the asset price index and the consumer behavior. The symptoms are not only concerns of consumers, but much more a change of the economic activity.

These variables were not only chosen because they are comparable for different countries with different economic environments, but also because they give rise to

particular concerns, if they develop into a certain direction.

One main advantage of using macroeconomic variables for modeling is, that the data history is completely available for a long time. Usually the variables are stated quarterly. Since the data history of the PD is available on a monthly basis and in order to get more data sets for the estimation, the time series of the macro economic variables can be disaggregated. This is a possibility of interpolating a high frequency series, where the average corresponds to the low frequency time series.

The procedure can be split into two different steps. The first step is the determination of a preliminary high frequency time series. The approach of *Denton-Cholette*, which will be used for the estimation, applies a single indicator as preliminary series. This could be a series consisting of only 1s.

The second step is to distribute the differences between the low frequency values of the preliminary series and the observed low frequency series among the high frequencies of the preliminary series. This is done by minimizing the squared absolute or relative deviations from the preliminary series.

3.3 Stress scenarios

In order to put the financial institutions into hypothetical stress, different scenarios are defined by the European Banking Authority. As already mentioned in Section 3.1, the following situations are simulated:

- The ***baseline scenario*** should constitute the expected economic development. It serves to compare the results of the "stressed" scenario. At the same time it is a good indicator for the development and the stability of the bank under the assumption that there are no economic shocks. Besides that it would be unrealistic to assume that the macroeconomic situation is stationary.
- The ***adverse scenario*** should simulate a "stressed" situation and is derived from systemic risks assessed by the European Systemic Risk Board (ESRB). Originating from different sources of risk, the financial and economic shocks are translated into the change in different key figures.

According to the European stress test of 2014 the following potential risks and damages were identified by the ESRB:

The table and a detailed explanation can be read in the publication about the design of the macroeconomic adverse scenario of the 2014 stress test, available on the EBA web page.

According to those risk indicators, the following rates were defined for the above mentioned variables:

Source of risk	Financial and economic shocks
Increase in global bond yields amplified by an abrupt reversal in risk assessment, including towards emerging-market-economies (EMEs), and pockets of market liquidity	• worldwide financial market shocks • demand shocks in EMEs • EU countries: foreign demand shocks via a decline in world trade • currency depreciation and funding stress affecting Central and Eastern European economies
Further deterioration of credit quality in countries with feeble demand, with weak fundamentals and still vulnerable banking sectors	• EU country−specific aggregate demand shocks (via fixed capital formation and private consumption) • EU country−specific aggregate supply shocks (via shock on user cost of capital, nominal wages) • EU country-specific house price shocks
Stalling policy reforms jeopardising confidence in the sustainability of public finances	• EU country specific sovereign bond spread shocks
Lack of necessary bank balance sheet repair to maintain affordable market funding	• EU-wide shock to short-term interbank interest rates • EU country−specific shocks to borrowing costs for households (via shocks to household nominal wealth and user cost of capital)

Table 5: ESRB mapping of financial stability risks to shocks

SCENARIOS	Baseline growth in %			Adverse growth in %		
	2014	2015	2016	2014	2015	2016
real GDP	1,5	1,8	1,7	-0,2	-1,5	-0,1
CPI	1,8	1,8	1,9	1,4	1,3	1,2
UR	4,8	4,7	4,7	5	5,5	6,1
RPP inflation	2,9	3,2	3,8	-4,2	-2,5	2,3
EP	5,4	-0,2	0		3,7	-0,9

Figure 8: Scenarios defined for 2014s stress test

These scenarios were forecasts for the future and available in this form for every European Country that took part in the stress test.

3.4 The stressed PD

The goal of the stress test is, among other things, to estimate a stressed PD, which operates as an input factor for the RWA calculation. The model should evaluate, how the customer characterized PD behaves under the assumptions made for the macroeconomic variables in Figure 8.

The approach of using increased risk parameters, in this case the PD, for the calculation of the required equity is called *Uniform Stress Test*. There are several ways of performing this kind of stress test. In this thesis one way is presented.

In order to make a forecast in consideration of the future stress scenarios, the information of the past is used. The macroeconomic variables can be regarded over their historical course of time as shown in Figure 9.

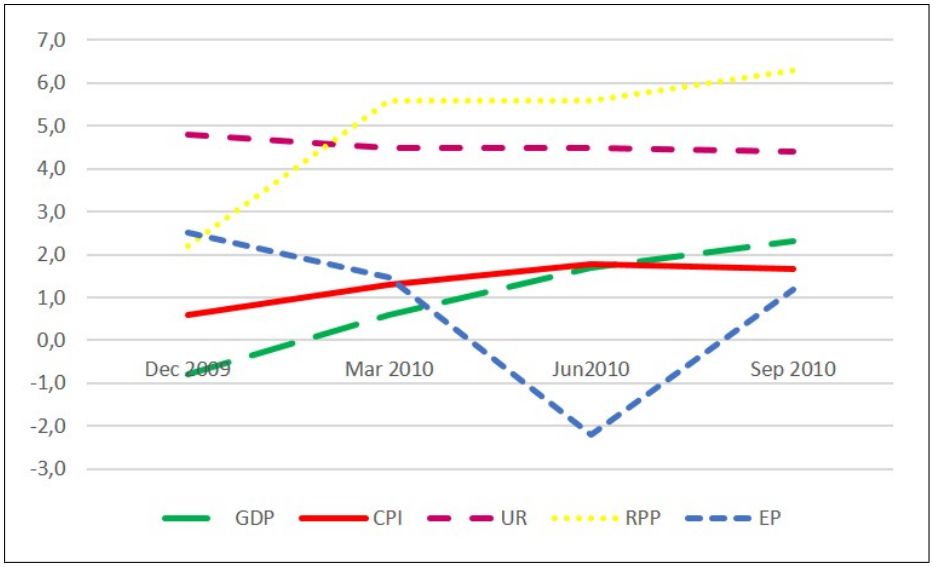

Figure 9: Part of the time series of the macroeconomic variables

In accordance with the hierarchy shown in Figure 7, the starting values are estimated PDs for each customer. As a customer c_i can own products p_{ij} from the type loan (L), credit card (CC) or current account (CA), the starting values are PDs estimated on product level for each product of a customer. For simplification it is assumed that every customer owns maximally one product of each type. This results in the following data base with one data set for every customer at every point in time a PD was calculated:

Jan 2008	c_1	$\text{PD}_{\boldsymbol{L}_1}$	$\text{PD}_{\boldsymbol{CC}_1}$	$\text{PD}_{\boldsymbol{CA}_1}$
Feb 2008	c_1	$\text{PD}_{\boldsymbol{L}_1}$	$\text{PD}_{\boldsymbol{CC}_1}$	$\text{PD}_{\boldsymbol{CA}_1}$
$\vdots$	c_1	$\text{PD}_{\boldsymbol{L}_1}$	$\text{PD}_{\boldsymbol{CC}_1}$	$\text{PD}_{\boldsymbol{CA}_1}$
Oct 2013	c_1	$\text{PD}_{\boldsymbol{L}_1}$	$\text{PD}_{\boldsymbol{CC}_1}$	$\text{PD}_{\boldsymbol{CA}_1}$
Mar 2009	c_2	$\text{PD}_{\boldsymbol{L}_2}$	$\text{PD}_{\boldsymbol{CC}_2}$	$\text{PD}_{\boldsymbol{CA}_2}$
Apr 2009	c_2	$\text{PD}_{\boldsymbol{L}_2}$	$\text{PD}_{\boldsymbol{CC}_2}$	$\text{PD}_{\boldsymbol{CA}_2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Dec 2013	c_n	$\text{PD}_{\boldsymbol{L}_n}$	$\text{PD}_{\boldsymbol{CC}_n}$	$\text{PD}_{\boldsymbol{CA}_n}$

Table 6: Example for the data base of customers

In order to consider the PD in dependency from the explanatory variables in the form of the macro-economic key figures, the temporal component has to be taken into account. To accomplish this, the mean PD of every product type for each specific point in time is calculated.

$$\overline{PD}_{\boldsymbol{L}_t} = \frac{1}{n}\sum_{i=1}^{n} PD_{\boldsymbol{L}_i},\ t = 1, \ldots, T \tag{3.1}$$

$$\overline{PD}_{\boldsymbol{CC}_t} = \frac{1}{n}\sum_{i=1}^{n} PD_{\boldsymbol{CC}_i},\ t = 1, \ldots, T \tag{3.2}$$

$$\overline{PD}_{\boldsymbol{CA}_t} = \frac{1}{n}\sum_{i=1}^{n} PD_{\boldsymbol{CA}_i},\ t = 1, \ldots, T \tag{3.3}$$

$t \in \{1, \ldots, T\}$ is the specific point in time and $i \in \{1, \ldots, n\}$ a customer. Together with the macroeconomic variables, the following historical data base of starting values would be given.

t	Date	GDP	CPI	UR	RPP	EP	$\overline{PD}_{\boldsymbol{L}_t}$	$\overline{PD}_{\boldsymbol{CC}_t}$	$\overline{PD}_{\boldsymbol{CA}_t}$
1	Jan 2008	0,6	3,7	4,1	0,3	2,5	1,09	2,21	0,87
2	Feb 2008	1,7	3,2	4,5	5,6	6,8	1,14	1,19	1,11
3	Mar 2008	2,3	3,8	3,7	4,2	1,3	1,25	1,75	0,91
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 7: Example for the historical data base of starting values

In the next step a multivariate regression model can be used to estimate the stressed PD (the methodology and the preconditions are discussed in the next chapter). Thereby a coefficient $\boldsymbol{\beta}$ for every explanatory variable on each dependent variable is estimated.

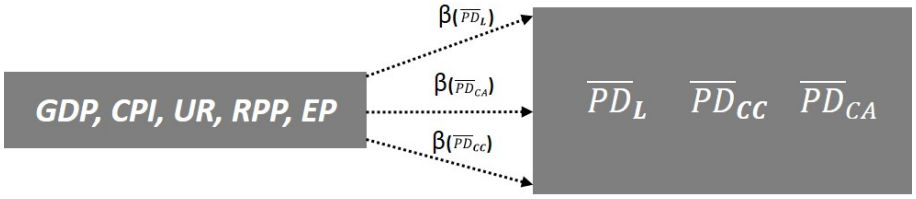

Figure 10: Coefficients are estimated for every dependent variable

The coefficient describes the effect of the macroeconomic variable on the dependent PD, whereby the effect is specifically for the particular component. The output of the multivariate regression for one component (i.e. the PD for one product type) is an equation of the following form.

$$y_i = \beta_0 + \beta_{i,1}x_1 + \beta_{i,2}x_2 + \cdots + \beta_{i,n}x_n \tag{3.4}$$

with

y_i . . . the respond variable, in this case the PD for product type $i \in \{1, 2, 3\}$
x_j . . . the jth explanatory variable, e.g. the GDP
$\beta_{i,j}$ . . the parameter which captures the effect of x_j on y_i
$\beta_{i,0}$ . . the intercept resulting from the regression

Equation (3.4) makes it possible to forecast a stressed PD with the specified scenarios by simply inserting the adverse scenario values (marked with ^{adv} in the equation). If we assume, that the regression provided estimates for every macroeconomic variable, the stressed PD would be the result of:

$$PD_{\boldsymbol{L}}^{stress} = \beta_0 + \beta_1 \boldsymbol{GDP}^{adv} + \beta_2 \boldsymbol{CPI}^{adv} + \beta_3 \boldsymbol{UR}^{adv} + \beta_4 \boldsymbol{RPP}^{adv} + \beta_5 \boldsymbol{EP}^{adv} + \epsilon$$

ϵ is the error of the estimation. The evaluation of the PD for the baseline scenario and the other product types works analogous. Through the comparison of $PD_{\boldsymbol{L}}^{stress}$ and $PD_{\boldsymbol{L}}^{base}$ an add-on for the PD of each customer could be derived.

4 Regression models

Financial institutions in the evaluation, whether a customer is able to pay his debts or not, the grocer round the corner when ordering his goods just like everyone in his or her everyday life is reliant on recognizing interrelationships and much more to assess them.

If we buy bread at the bakery, we know that the amount payable (y) results from the product of the price of one loaf and the number of loaves (x) purchased. This accords to the linear function

$$y = \beta_0 + \beta_1 x,$$

where y is the amount payable, β_1 is the price of one loaf and x is the number of loaves we bought. The intercept β_0 would in this case be 0, because we wouldn't have to pay anything for bread if $x = 0$. If we buy more bread, the amount payable increases as well.

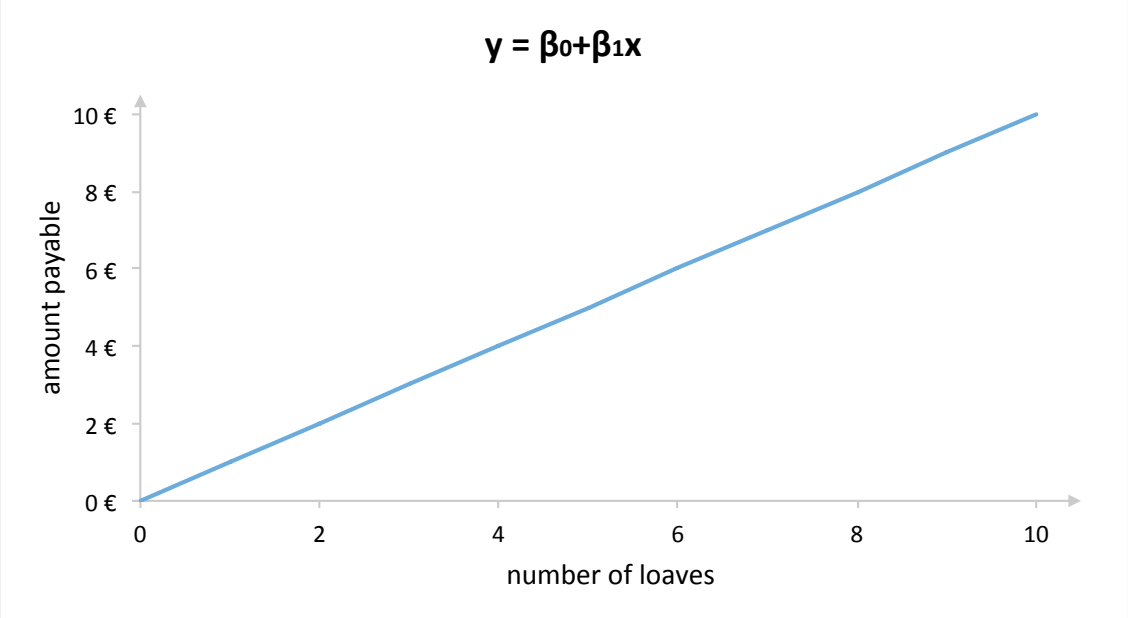

Figure 11: Exact relation between the amount payable and the number of loaves

This example shows, that exact relations can be described in the form of a function very easily. A lot of interrelationships aren't exact but apply approximately. If we examine for example how the amount of income depends on the number of working hours, a positive relation is expected indeed, but this relation can't be valid exact as the income depends on numerous factors, not just the number of working hours.

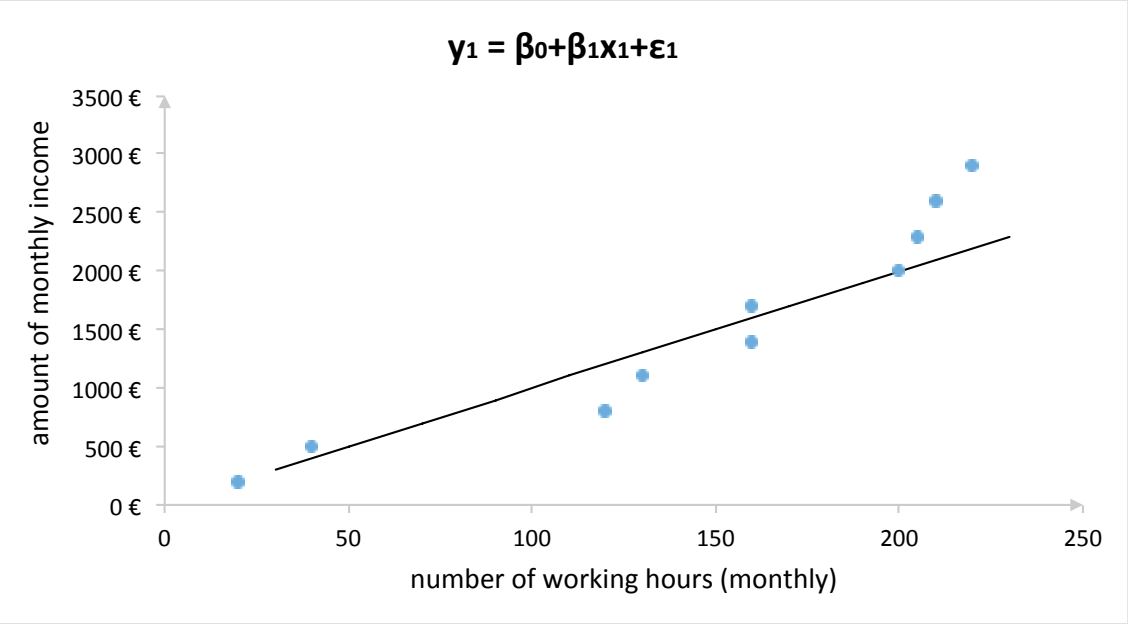

Figure 12: Relation between the amount of income and the number of working hours

As Figure 12 shows, normally people have a higher income with more working hours. But to describe this approximate relation a certain deviation or error has to be taken into account. In the equation

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

this error is described through the disruptive term ϵ_i, where i is the ith observation of the data sample. Since the relation between the amount of income and the number of working hours can't be described in a single function for all observations, each of the n considered person would need a separate function.

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2 \\ \vdots &= \vdots \quad\quad \vdots \quad\quad \vdots \\ y_n &= \beta_0 + \beta_1 x_n + \epsilon_n \end{aligned}$$

This approach is not only confusing, but also inexpedient as each of the n different functions would again describe the relation for just one observation. In this case the *regression models* provide remedy. There are several ways of describing relations between variables, since there are several types how variables can relate to each other (linear, exponential, ...). This chapter will focus on logistic regression and multivariate linear regression.

The basis for every regression is

- the dependent or response variable (y - according to the example above)
- the explanatory variables, also called *regressors* (x in the example)

The objective is to find a function of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \epsilon \tag{4.1}$$

with the smallest possible ϵ (in this case the indexes don't refer to the number of observations, but to the number of explanatory variables). Thereby the coefficients $\beta_1, \ldots, \beta_m$ and an intercept β_0 are estimated to describe the relation. The output contains those coefficients, that minimize the disruptive term ϵ. Different types of regressions use different methods for minimizing the error term.

In conjunction with PD modeling and stress testing, regression models are used for forecasting. If there are additional observations for $\boldsymbol{x} = (x_1, \ldots, x_k)$ without knowing y, the values can be imputed into the regression function to predict y.

4.1 Multiple logistic regression

4.1.1 Binary response

What differentiates the logistic regression model from usual linear regression model the most, is the type of the response variable y. While the dependent value in the linear regression is continuous, the logistic regression model requires a binary response. That leads to the following definition for y_i (i denotes the ith observation of the sample with $i \in \{1, \ldots, n\}$)

$$y_i = \begin{cases} 1, \ default \\ 2, \ non - default \end{cases}$$

Let the vector

$$\boldsymbol{x_i}' = (x_{i,1}, \ldots, x_{i,k})$$

be the collection of independent, explanatory variables (e.g. product risk characteristics). The key figure of every regression problem is the conditional expected value

$$\mathbb{E}(y_i | \boldsymbol{x_i}),$$

i.e. the mean of the outcome variable y_i given the independent variables $x_{i,1}, \ldots, x_{i,k}$. In linear regression it is assumed, that this mean can be written as

$$\mathbb{E}(y_i | \boldsymbol{x_i}) = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k},$$

which is linear in $\boldsymbol{x_i}$. This allows any value for $\mathbb{E}(y_i | \boldsymbol{x_i}) \in (-\infty, +\infty)$. For binary variables applies:

$$\boldsymbol{x_i}' = (1, x_{i,1}, \ldots, x_{i,k})$$

This requires a cumulative distribution, such as the logistic distribution. A transformation function F is needed, which ensures that

$$F(\boldsymbol{\beta}'\boldsymbol{x_i}) \in [0,1]$$

with $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_k)$ and $\boldsymbol{\beta}'\boldsymbol{x_i} = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k}$. The *logit* transformation is based on the distribution function of the logistic distribution:

$$\mathbb{P}(y_i = 1) = \Lambda(\boldsymbol{\beta}'\boldsymbol{x_i}) = \frac{exp(\boldsymbol{\beta}'\boldsymbol{x_i})}{1 + exp(\boldsymbol{\beta}'\boldsymbol{x_i})} \tag{4.2}$$

Λ is the distribution function of the standard-logistic distribution with mean 0 and variance $\frac{\pi^2}{3}$. Thus the transformation function is Λ, which implies:

$$\mathbb{P}(y_i = 1|\boldsymbol{x_i}) = F(\boldsymbol{\beta}'\boldsymbol{x_i})$$

and

$$\mathbb{P}(y_i = 0|\boldsymbol{x_i}) = 1 - F(\boldsymbol{\beta}'\boldsymbol{x_i})$$

With an independent data sample, the joint probability is

$$\mathbb{P}(y_1, y_2, \ldots, y_n|\boldsymbol{X} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})') = \prod_{y_i=0}[1 - F(\boldsymbol{\beta}'\boldsymbol{x_i})] \prod_{y_i=1}[F(\boldsymbol{\beta}'\boldsymbol{x_i})]$$

For simplification of the notation:

$$\frac{exp(\boldsymbol{\beta}'\boldsymbol{x_i})}{1 + exp(\boldsymbol{\beta}'\boldsymbol{x_i})} =: \pi(\boldsymbol{x_i}) \tag{4.3}$$

This means that the transformation function F is equivalent to $\pi(\boldsymbol{x_i})$.

4.1.2 The disruptive term

One assumption for the logistic regression model is aimed at the disruptive term. Let $\mathbb{E}(y_i|\boldsymbol{x_i}) + \epsilon$ be the observation of the outcome variable with the error ϵ, which can take two possible values:

1. $\epsilon = 1 - \pi(\boldsymbol{x_i})$ if $y_i = 1$
2. $\epsilon = -\pi(\boldsymbol{x_i})$ if $y_i = 0$

Possibility 1 has probability $\pi(\boldsymbol{x_i})$ and the second possibility has probability $1 - \pi(\boldsymbol{x_i})$. This implies, that ϵ has a distribution with mean 0 and variance $\pi(\boldsymbol{x_i}) \cdot [1 - \pi(\boldsymbol{x_i})]$, and thus is described by the binomial distribution.

4.1.3 Fitting the model

Let's consider a sample of n independent observations of both, the explanatory variables $\boldsymbol{x_i}$ and the binary response variable y_i. The variables are defined as above. Fitting the logistic regression model of Equation 4.3 to the given set of data means estimating the values of the unknown parameters

$$\beta_0, \beta_1, \ldots, \beta_m.$$

Linear regression uses the method of *least squares* (LS) for estimating parameters. The result are β_j's that minimize the sum of squared deviations of the observed values from the predicted values resulting from the model.

Estimators with a number of statistical properties are received. The estimators loose those properties if LS is applied to a model with binary response. Therefore we assume a more general approach, that leads to the LS function on the one hand and to the specific method used for logistic regression models on the other hand, the *maximum likelihood* (ML) estimation.

Very roughly, the ML estimation provides values for $\beta_0, \ldots, \beta_m$, that maximize the probability of obtaining the observed set of data. Therefore the so called *likelihood function* is constructed, which depicts the probability of the observed data. The likelihood function is a function from the unknown parameters and the *maximum likelihood estimators* of the parameters are those values, that maximize the ML function.

Equation (4.3) provides the conditional probability for $y_i = 1$ given $\boldsymbol{x_i}$, i.e. $\mathbb{P}(y_i = 1|\boldsymbol{x_i})$. Thus $1 - \pi(\boldsymbol{x_i})$ is analogously $\mathbb{P}(y_i = 0|\boldsymbol{x_i})$. Therefore:

- the contribution to the likelihood function is $\pi(\boldsymbol{x_i})$ for those observations $(y_i, \boldsymbol{x_i})$, where $y_i = 1$
- the contribution to the likelihood function is $1 - \pi(\boldsymbol{x_i})$ for those observations $(y_i, \boldsymbol{x_i})$, where $y_i = 0$

Thus the contribution to the likelihood function for the observation $(y_i, \boldsymbol{x_i})$ can be summarized as

$$\pi(\boldsymbol{x_i})^{y_i} \cdot [1 - \pi(\boldsymbol{x_i})]^{1-y_i} \tag{4.4}$$

Now the independence of the observations comes into play, which allows the outlining of the likelihood function as follows:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\boldsymbol{x_i})^{y_i} [1 - \pi(\boldsymbol{x_i})]^{1-y_i} \tag{4.5}$$

This function is maximized in dependence of $\boldsymbol{\beta}$. If l is differentiable, the maximum can be obtained by forming the first derivative after $\boldsymbol{\beta}$ and equate it to 0. Since this

procedure can get very complex for density functions with intricate exponents, the *logarithmized likelihood function* is used.

$$L(\boldsymbol{\beta}) = \ln l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i \ln \pi(\boldsymbol{x_i}) + (1 - y_i) \ln [1 - \pi(\boldsymbol{x_i})]\} \tag{4.6}$$

In order to find the $\boldsymbol{\beta}$ that maximizes $L(\boldsymbol{\beta})$ the function is differentiated with respect to $\beta_0, \ldots, \beta_m$ and equated to 0. The first order conditions are:

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n} \left[\frac{y_i}{\pi_i} \cdot \frac{d\pi_i}{d(\boldsymbol{\beta'} \boldsymbol{x_i})} + \frac{(1 - y_i)}{\pi_i} \cdot \left(- \frac{d\pi_i}{d(\boldsymbol{\beta'} \boldsymbol{x_i})} \right) \right] \boldsymbol{x_i}' \stackrel{!}{=} 0 \tag{4.7}$$

The values of $\beta_0, \ldots, \beta_m$, thus the maximum likelihood estimators are then given by the solution to (4.7) and are referred to with $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta_m}$. Those values are the so called *coefficients* or the *intercept* of the logistic regression.

Summarized the model can be written as

$$ln\left(\frac{\mathbb{P}(y_i = 1)}{1 - \mathbb{P}(y_i = 1)} \right) = \beta_0 + \beta_1 \boldsymbol{x_{i,1}} + \cdots + \beta_m \boldsymbol{x_{i,m}}$$

4.1.4 Interpretation of the model

Interpreting a logistic regression model leads to the question: *Is the outcome described better by a model including those variables, than a model not including those variables?* In order to answer this question, the outcome has to be compared with the observed value for the model including the questioned variables and the model that doesn't include the questioned variables.

The interpretation of the estimators for a logistic regression model is not as easy as for a linear model. Since the function is not linear, no direct claims can be made about the relation between the dependent and the explanatory variables (see: example from above). It is for this reason, that the coefficients aren't considered directly for interpretation, but the so called *odds* are analyzed. The odds describe the relation between the probability of occurrence ($y_i = 1$) to the probability that it does not occur ($y_i = 0$).

$$odds(y_i = 1) = \frac{\mathbb{P}(y_i = 1)}{1 - \mathbb{P}(y_i = 0)} \tag{4.8}$$

For the logistic regression, the following correspondence applies:

$$Logit = ln(odds)$$

That means, the logits of a logistic regression model are the logarithmized odds. The odds ratio (OR) is the relation between the odds, i.e. how strongly the presence or absence of an explanatory variable is in conjunction with the presence/absence of

another explanatory variable. Let's assume that $\boldsymbol{x_i}$ is coded as either 0 or 1. Then it applies:

$$OR = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)[1-\pi(0)]}$$

For the logistic regression model is then obtained:

outcome variable y_i	**explanatory variable**	
	$\boldsymbol{x_i} = 1$	$\boldsymbol{x_i} = 0$
$y_i = 1$	$\pi(1) = \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}$	$\pi(0) = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}$
$y_i = 0$	$1-\pi(1) = \frac{1}{1+\exp(\beta_0+\beta_1)}$	$1-\pi(0) = \frac{1}{1+\exp(\beta_0}$
total	**1**	**1**

Figure 13: Values of the logistic regression

It follows for the OR:

$$\begin{aligned} OR &= \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right) / \left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right) / \left(\frac{1}{1+e^{\beta_0}}\right)} \\ &= \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} \\ &= e^{(\beta_0+\beta_1)-\beta_0} \\ &= e^{\beta_1} \end{aligned}$$

From that follows that the relationship between the regression coefficient and the odds ratio is

$$OR = e^{\beta_1}$$

In other words, the odds ratio approximates how much more likely it is for the outcome to be present among those with $x = 1$ to those with $x = 0$. For example, if x denotes whether the borrower is a new client ($x = 1$) or not ($x = 0$), then $\hat{OR} = 2$ estimates, that it is twice as likely to default as a new client, then it is as a regular customer. This simple relationship is one reason, why logistic regression is such an efficient analytical tool.

4.1.5 The significance of the coefficients

After fitting the model, it has to be tested on significance. The Wald test statistic

(W) is used to assess the significance of each coefficient on its own. This univariate statistic is:

$$W_j = \left(\frac{\hat{\beta}_j}{\hat{SE}(\hat{\beta}_j)} \right)^2$$

for the jth explanatory variable with the univariate standard error $\hat{SE}(\hat{\beta}_j)$ of $\hat{\beta}_j$. Under the null hypothesis that the coefficients $\beta_0, \ldots, \beta_m$ of the model are equal to 0, the distribution of the statistic W is χ^2 with $n-1$ degrees of freedom. This leads to the p-value:

$$\mathbb{P}[\chi^2(m) > W] =: p - value$$

The null hypothesis, that the coefficient equals zero is rejected in the case of

$$p - value < \alpha$$

for a specified level $\alpha > 0$. Normally α is set equal to $0,05$.

4.1.6 Preconditions

The logistic regression has less preconditions than the linear regression. Still there are some facts that have to be considered for applying a logistic regression model. Most of the issues were discussed in the previous sections.

- The inclusion of all relevant variables and at the same time the exclusion of the irrelevant variables are important for a good fit. Otherwise the model could be under fitted or over fitted.
- The disruptive terms need to be independent.
- The amount of missing values should be small.
- Linearity and additivity of the explanatory variables with the logit.
- Multicollinearity between the explanatory variables should be avoided, otherwise the standard errors of the coefficients get too big.
- A large sample to improve the significance and the discriminatory power of the model.
- Independence of the data sets for the preconditions of the maximum likelihood estimation.
- No outliers, which will be discussed in the next section.

4.2 Outliers and robustness

Estimations can be strongly influenced by just a few data sets. In this case, the result of the estimation is biased as it doesn't fit the largest part of the data. When such a problem occurs, it has to do with outliers.

In the data of financial institutions, which are used for estimating PDs, phenomena appear sometimes, which lead to the occurrence of outliers. The cause often lies in the careless data collection (e.g. manual entries, process work-arounds, ...). However, it is possible that some observations really are that different from the greatest part of the observations. Considering the variable *months since account opening* it is possible, that for some few customers, who opened their accounts decades ago, the value of the variable is very high. In this case the fitting of the model has to be adjusted.

For the identification of outliers, a clean definition is necessary. In general it can be said, that an outlier is an observation that strongly deviates from the other observations. This can be on univariate, as well as on multivariate level.

Outliers on univariate level

In the context of PD modeling, the outlier detection on univariate level is performed for each of the explanatory variables. In the following a data set of n observations for the variable x is considered.

$$\begin{matrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{matrix}$$

In general, outliers can be detected graphically or computationally. One possibility of doing this, is the application of the frequency distribution. The edges of the distribution should not be classified in this case. A weakness of this method is, that the decision, whether an observation is an extreme value or not is discretionary.

A very common way for a graphical detection of outliers is the box plot, which illustrates the homogeneity as well as the range of the considered variable. In Figure 14 an example is given.

For the characterization of the homogeneity, the first and third quartile is used. The 50% of the observations, that lie between the first and the third quartile are represented by a box. The interquartile range (IQR) is illustrated by the lines at the edges of the box. The outliers are defined as the points right from the upper whisker which corresponds to the *third quartile* + 1.5· *IQR* (see Tukey, 1977).

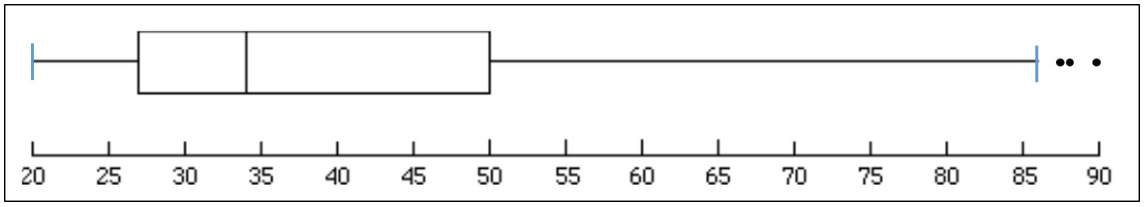

Figure 14: Box plot of the variable x

One way of detecting outliers is with the help of test procedures. One precondition for the implementation of such tests is set to the underlying distribution.

Usually it is required that the data follow at least approximately a normal distribution. One of the most common tests of this kind is the outlier test by Grubbs.

Outliers on multivariate level

For multivariate analyses like the multiple logistic regression, the problem of outliers gets much more complex.

1. A potential outlier doesn't only distort the location and dispersion parameters, but also the correlation between the variables can be affected.

2. While a univariate outlier is located at the edges of the distribution, the multivariate outliers position is not restricted in its location.

The consideration of the frequency distribution alone is not sufficient, as in this case outliers refer to a whole data set, not only a point.

The basis for the multivariate outlier detection is a data set of n observations for m different variables $x_1, \ldots, x_m$.

$$\begin{array}{cccc} x_{1,1}, & x_{1,2}, & \cdots & x_{1,m} \\ x_{2,1}, & x_{2,2}, & \cdots & x_{2,m} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n,1}, & x_{n,2}, & \cdots & x_{n,m} \end{array}$$

The probably most obvious way of detecting outliers is to calculate the Euclidean distance for each observation and compare it to the central point of the data.

For simplification and in order to allow a graphical representation, consider two-dimensional observations $(x_{i,1}, x_{i,2}) =: \boldsymbol{x_i}$ with the central point $\overline{\boldsymbol{x}} := (\overline{x}_1, \overline{x}_2)$. Then, the Euclidean distance $d(\cdot, \cdot)$ of x_i to the central point $\overline{\boldsymbol{x}}$ is defined as:

$$d(\boldsymbol{x_i}, \overline{\boldsymbol{x}}) = \sqrt{(x_{i,1} - \overline{x}_1)^2 + (x_{i,2} - \overline{x}_2)^2} \tag{4.9}$$

An example should demonstrate, that the Euclidean distance is not reasonable in this case. Consider the two-dimensional data set, each consisting of 22 observations. The central point of the sample is $(4,8; 3,5)$.

i	$x_{i,1}$	$x_{i,2}$	$d(x_i; \overline{x})$
1	2,3	2,2	2,4
2	3,1	4,5	1,7
3	2,7	2,6	1,9
4	3	2,6	1,6
5	3,5	2,8	1,1
6	4	3	0,6
7	4,2	3,2	0,3
8	4,5	3,3	0,1
9	4,5	2,9	0,5
10	4,6	3,5	0,2
11	4,7	2,6	0,8
12	4,8	3,7	0,5
13	4,9	3,1	0,6
14	5	4	0,8
15	5,3	3,5	0,9
16	5,6	3,2	1,2
17	5,9	4,2	1,7
18	6,2	3,9	1,8
19	6,5	4,7	2,5
20	6,8	4,9	2,8
21	7	4,6	2,8
22	5,8	3,8	1,4

The average Euclidean distance lies at $1,3$. This suggests that observation 2 with an Euclidean distance of $1,7$ is located nearby the central point of the sample. As it can be seen in Figure 15, this precise observation is the only one which does not lie in the general direction of the scatter.

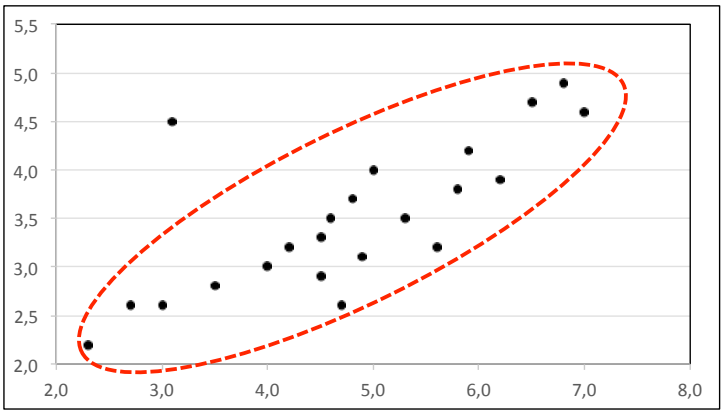

Figure 15: Example for the approach of the Euclidean distance

This example should demonstrate, that the detection of multivariate outliers is not only dependent from the distance to one "point" of the sample. Other tools than the Euclidean distance are necessary. One common measure for identifying multivariate outliers is the so called *Mahalanobis distance*. This quantity is a distance measure between points in a multidimensional vector space. This quantity measures the similarity of the observations to the sample mean with regard to the dependencies of the m different variables, i.e. the covariances.

The classical Mahalanobis distance is given by

$$MD_i = \sqrt{(\boldsymbol{x_i} - \overline{\boldsymbol{x}})' \Sigma^{-1} (\boldsymbol{x_i} - \overline{\boldsymbol{x}})}, \ \forall i = 1, \ldots, n \tag{4.10}$$

with

$\boldsymbol{x_i}$	...	the tuple $(x_{i,1}, x_{i,2})$
i	...	the index of the ith observation of the sample
n	...	the number of observations in the sample
$\overline{\boldsymbol{x}}$	...	the mean value $(\overline{x}_1, \overline{x}_2)$
Σ^{-1}	...	the inverse covariance matrix of the independent variables

For demonstrating the idea of the Mahanlanobis distance, consider a bivariate normally distributed sample of 500 observations (x_1, x_2) with $\boldsymbol{x_1} = (x_{1,1}, \ldots, x_{n,1})$ and $\boldsymbol{x_2} = (x_{1,2}, \ldots, x_{n,2})$.

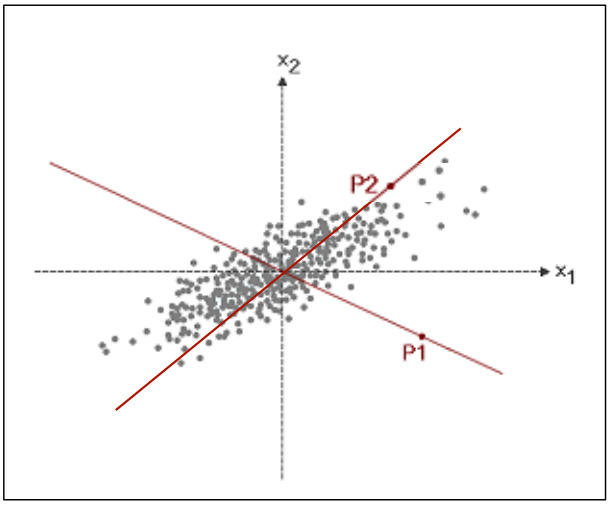

Figure 16: Example of a bivariate normally distributed sample

The points $P_1 = (x_{i_1,1}, x_{i_1,2})$ and $P_2 = (x_{i_2,1}, x_{i_2,2})$ in Figure 16 are at different distances from the central point of the sample. This is shown clearly by the intersection lines. The difference is even more notable through the distribution along the intersection lines P_1M and P_2M.

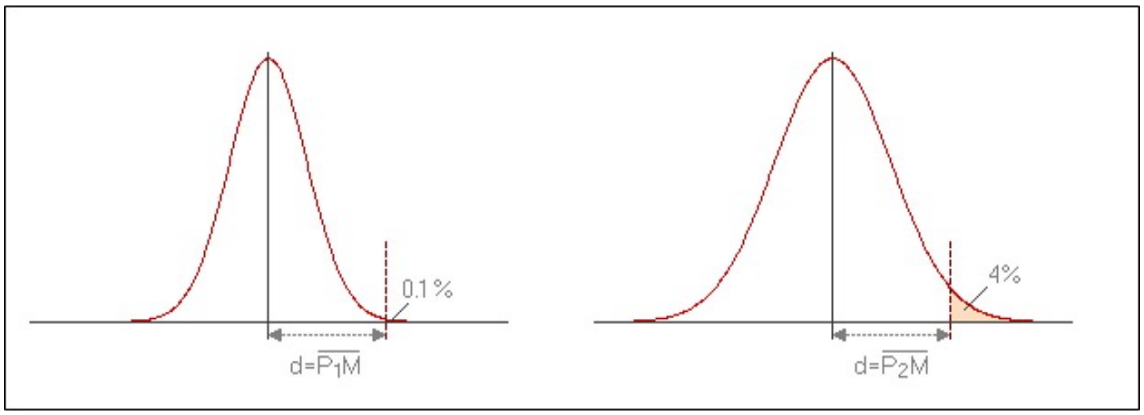

Figure 17: Distribution along the intersection lines

The probability of an observation at position P_1 equals $0,1\%$ and at position P_2 it is at $0,4\%$. This confirms that the points don't have the same distance to the origin with regard to the dependencies of the variables, i.e. the distribution of the sample.

The Euclidean distance corresponds to drawing a circle around the origin. According to this the points would have the same distance, although the Mahalanobis distances in Figure 17 differ from each other. The correct approach is to draw curves with the same probability, i.e. ellipses so that the points at curves have the same probability of occurrence. This method considers the multivariate standard deviation and the ellipses with constant probabilities correspond to the constant Mahalanobis distance.

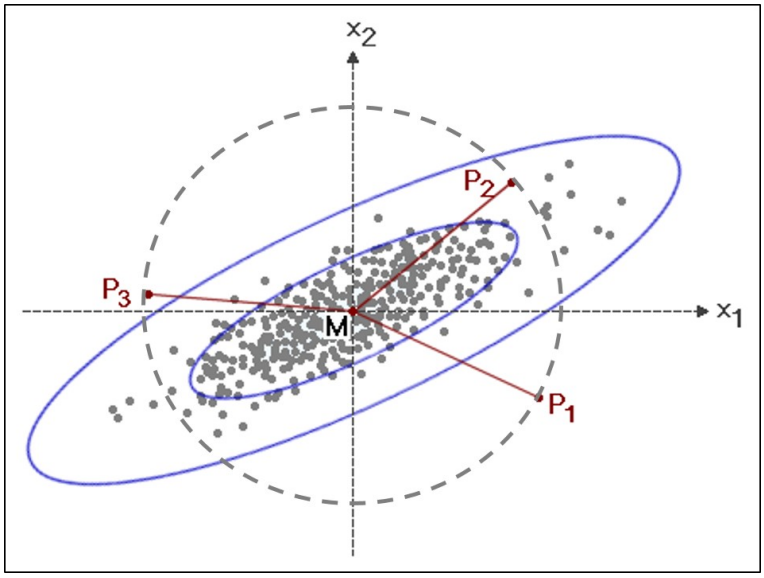

Figure 18: Graphical depiction of the Mahalanobis distance

If the sample mean and the sample variance are used as estimators of $\overline{\boldsymbol{x}}$ and Σ^{-1} for calculating the Mahalanobis distance, the values could be biased due to outliers. To avoid this problem more robust estimators are necessary like the *minimum covariance determinant* estimators (MCD) (see Rousseeuw and Van Driessen, 1999).

In this context $\overline{x}_{MCD}$ is the arithmetic mean of those h observations that have the smallest possible determinant of the classical covariance matrix. Σ^{-1}_{MCD} results from the sample covariance of the same h observations.

This rises the question to what extent contaminated data sets or outliers are allowed so that the estimator still provides reliable information about the underlying variable. In other words, what is the breakdown value? As a measure for robustness, the breakdown value of a location estimator determines the smallest amount $\frac{m}{n}$ of data sets, whose replacement leads to an unlimited change of the estimation. n is

the number of all observations and m is the number of replaced observations.

$$\epsilon^*(\overline{x}) := \min_m \left\{ \frac{m}{n}; \sup_{\tilde{x}_i} \|\overline{x}(\tilde{x}_i) - \overline{x}(x_i)\| = \infty \right\}, \tag{4.11}$$

where $\overline{x}(x_i)$ denotes the mean value of the whole data sample including all outliers and $\overline{x}(\tilde{x}_i)$ is the mean value of the respective comparative sample (see Hubert and Debruyne, 2010).

The breakdown value of Σ^{-1} is defined as the smallest amount $\frac{m}{n}$ of data sets that with one of the following characteristics:

- For the largest eigenvalue: $\lambda_1(\Sigma) \to \infty$
- For the smallest eigenvalue: $\lambda_p(\Sigma) \to 0$ (in this case p is the number of considered variables $\hat{=}$ m in Equation (4.10))

For the breakdown values of MCD estimates it applies

$$\epsilon^*(\overline{x})_{MCD} = \epsilon^*(\Sigma)_{MCD} \approx \frac{n-h}{n}$$

with a maximum value of $\epsilon^* = 50\%$ for $h \approx \frac{n}{2}$.

As the quadratic Mahalanobis distances MD_i^2 are χ_p^2 distributed under multivariate normally distributed data, a limit for the number of expected outliers can be found with $\sqrt{\chi^2_{p;1-\alpha}}$ as threshold. This threshold identifies all data sets with $MD_i > \sqrt{\chi^2_{p;1-\alpha}}$ as outliers.

4.3 Robust logistic regression

The main target of a regression is to find a curve that fits into a given cloud of points. The most intuitive and common approach is to search for the curve with the minimum distance to the data points. This method is called *least-square* and minimizes the sum of the squares of errors for the choice of the parameters of the regression function. Larger deviations have a stronger weight in the calculation, which leads to distortions if the underlying data contain outliers.

The effects reach from changes in the significance or the sign of individual parameters up to deterioration of the model accuracy. Either way, to guarantee a reasonable risk assessment and valid results, robust methods have to be applied.

In Section 4.1 it was described how parameters were estimated with the help of the maximum likelihood method. This method is not steeled against outliers as the breakdown value equals $\frac{1}{n}$ which doesn't leave much room for deviations. Robust alternatives have been constructed based on the ML-estimation, which are elucidated in this section.

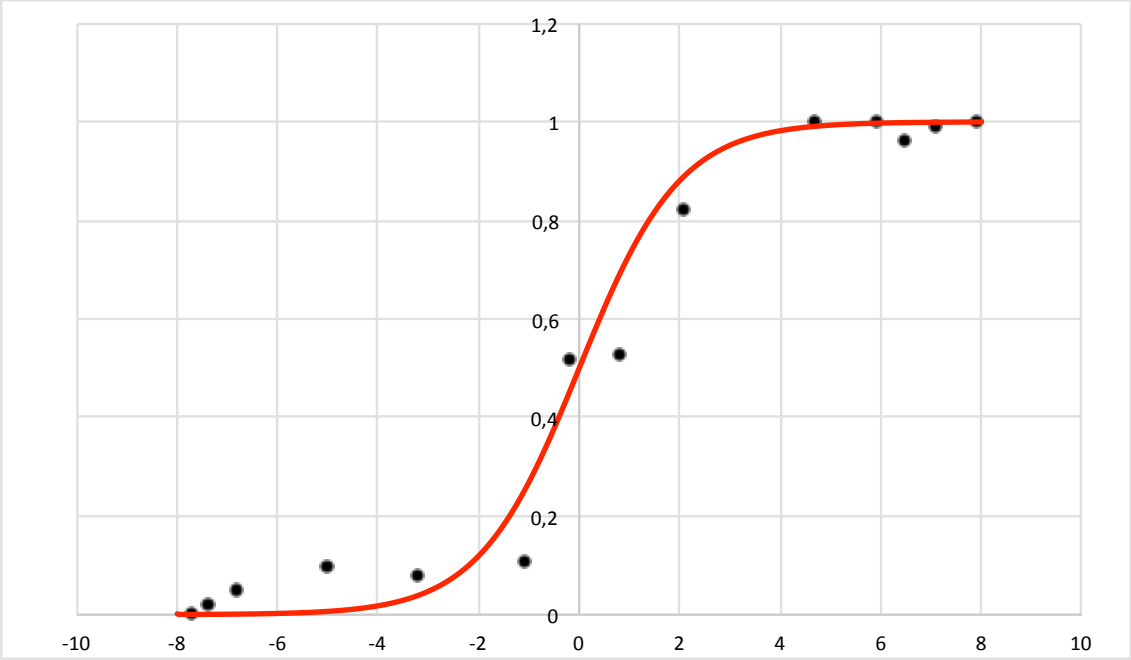

Figure 19: Fitted curve with residuals

Let Y_i with $1 \leq i \leq n$ be independent Bernoulli variables and $X_1, \ldots, X_n$ the p-dimensional explanatory variables that fulfill:

$$\mathbb{P}(Y_i = 1 | X_i = \boldsymbol{x_i}) = F(\alpha + \boldsymbol{\beta' x_i})$$

with the cumulative distribution function F. The robustness is analyzed on the basis of logistic regression, therefore it applies

$$F(u) = \frac{1}{1 + e^{-u}}$$

The ML-estimator can be written as

$$\hat{\gamma}_n^{ML} = \underset{\gamma}{\operatorname{argmax}} \log L(\gamma; X_n) = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} d(z_i'\gamma; y_i) \tag{4.12}$$

with

γ	...	the parameters $(\alpha, \beta')'$
z_i	...	$(1, x_i')'$ for all $1 \leq i \leq n$
n	...	the number of observations in the sample
X_n	...	the sample $\{(x_1, y_1), \ldots, (x_n, y_n)\}$
$\log L(\gamma; X_n)$	...	the log-likelihood function in γ
$d(z_i'\gamma; y_i)$	...	the deviance function given by

$$d(z_i'\gamma; y_i) = -y_i \log F(z_i'\gamma) - (1 - y_i) \log\{1 - F(z_i'\gamma)\}$$

In the ML-estimation the function $d(z_i'\gamma; y_i)$ is minimized in dependence of γ. The idea of a more robust estimator is based on a generalization of Equation (4.12). The

concept of M-type estimators, to which the generalization of the ML estimation refers, uses the approach of replacing the deviance function by a function φ, which is not that sensitive against outliers.

$$\hat{\gamma}_n = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} \varphi(z_i'\gamma; y_i) \tag{4.13}$$

φ satisfies $\varphi(s;0) = \varphi(-s;1)$ for any score s, with $s_i = z_i'\gamma$. For simplification, the univariate function $\phi(s) = \varphi(s;0)$ is considered, which describes the impact of a specific score s on the value of the target function in the Equation (4.13) for an observation that corresponds to a null y. For $\phi(s)$ it applies:

- $\phi(s)$ is non-decreasing
- $\lim\limits_{s\to\infty} \phi(s) = 0$

The task is to minimize $\sum\limits_{i=1}^{n} \varphi(z_i'\gamma; y_i)$ in dependence of γ or equivalently to solve

$$\frac{1}{n} \sum_{i=1}^{n} \Psi(z_i'\gamma; y_i) z_i = 0 \tag{4.14}$$

with

$$\Psi(s;0) = \tfrac{\partial\varphi(s;0)}{\partial s} \text{ and } \Psi(s;1) = -\Psi(-s;0)$$

For $\psi(s) = \Psi(s;0)$ follows $\psi(s) = \varphi'(s)$. The solution of (4.14) is the so called M-estimator.

Considering the deviance function of the log-likelihood function, the ML estimator apparently belongs to the class of M-estimators with

$$\varphi_{ML} = -\ln(1 - F(s))$$

By introducing a bounded function ρ and a bias correction term $C(\cdot;\cdot)$, Bianco and Yohai (1996) constructed a robust and consistent estimator (BY estimator), which belongs to the M-estimators as well:

$$\hat{\gamma}_n^{BY} = \underset{\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} \{\rho(d(z_i'\gamma; y_i)) + C(z_i'\gamma; y_i)\} \tag{4.15}$$

with

$$C(s,y) = G(F(s)) + G(1 - F(s)) \text{ and } G(t) = \textstyle\int_0^t \rho'(-\ln u)\mathrm{d}u$$

and the univariate function

$$\phi_{BY}(s) = \rho(-\ln(1 - F(s))) + G(F(s)) + G(1 - F(s)) - G(1) \tag{4.16}$$

(4.16) shows that the result depends on the selection of ρ. According to Bianco and Yohai (1996), an appropriate choice would be

$$\rho(t) = \begin{cases} t - \frac{t^2}{2c} & t \leq c \\ \frac{c}{2} & otherwise \end{cases} \tag{4.17}$$

with the tuning parameter c. The benefit of using robust estimators such as the BY estimator is that with ϕ_{BY}, large but bounded values are gained for very large scores > 0 while ϕ_{ML} probably leads to unlimited high values for such scores.

As the construction, also the conditions of existence for M-estimators can be derived from the ones for the ML-estimator, which leads to the following Proposition:

Proposition 4.1

1. *There is overlap of the data points.*
2. *ψ is increasing on $(-\infty, \infty)$ or $\exists\ k > 0$ with $\psi = \begin{cases} increasing\ on\ (-\infty, k] \\ decreasing\ on\ [k, \infty) \end{cases}$*
3. $\lim\limits_{s \to \infty} \frac{\psi(st)}{\psi(-s)} = \infty, \forall t > 0$

If these conditions are fulfilled, $\hat{\gamma}_n$ exists with finite norm.

Condition 1 can be traced back to Albert and Anderson (1984), who proved that the ML estimator exists if the space of the explanatory variables can't be separated into groups with $y_i = 0$ and $y_i = 1$.

With ρ as defined in Equation (4.17), condition 3 doesn't hold for $\hat{\gamma}_n^{BY}$ as the derivative $\rho'(t)$ vanishes for large t. As an alternative, for which the condition holds, a function with a derivative of the following form is recommended:

$$\rho'(t) = \begin{cases} e^{-\sqrt{d}} & t \leq d \\ e^{-\sqrt{t}} & otherwise \end{cases} \tag{4.18}$$

Proof.
In order to prove the existence of $\hat{\gamma}_n$ it is shown that the function

$$S: \ \gamma \mapsto \frac{1}{n} \sum_{i=1}^{n} \varphi(\gamma' z_i; y_i)$$

has a minimum inside a sphere with radius $< \infty$. Define

$$\lambda^* = \max_{\xi \in S^{p-1}} \lambda(\xi)$$

with

$$\lambda(\xi) = inf\{\lambda \geq 0 | \frac{\mathrm{d}K_\xi(\tilde{\lambda}}{\mathrm{d}\lambda} > 0, \forall \tilde{\lambda} \geq \lambda\} \text{ and } K_\xi : \lambda \mapsto \frac{1}{n} \sum_{i=1}^{n} \varphi(\lambda \xi' z_i; y_i)$$

and the surface

$$S = \{\xi \in \mathbb{R}^p | g(\xi) = 0\}, \text{ where } g : (u, \lambda) \mapsto \frac{\mathrm{d}K_u(\lambda)}{\mathrm{d}\lambda}$$

The function $\xi \mapsto \lambda(\xi)$ is continuous, λ^* is well defined. According to the definition of $\lambda(\xi)$ it follows that the minimum of the function S is always inside the compact set $\{\gamma : \|\gamma\| \leq \lambda^*\}$. The continuity of S guarantees that its minimum will always be reached for a $\hat{\gamma}_n$ with $\|\hat{\gamma}_n\| < \lambda^*$.

□

Croux and Haesbroeck developed an algorithm for the logistic regression under application of the BY-estimator. They split the problem of optimizing (4.15) by defining the parameter vector as $\gamma = \frac{\xi}{\sigma}$ with $\|\xi\| = 1$ and $\sigma = \frac{1}{\|\gamma\|} \geq 0$. This leads to:

$$(\hat{\sigma}, \hat{\xi}) = \underset{(\sigma,\xi) \in \mathbb{R}^+ \times S^{p-1}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \varphi(\frac{z_i' \xi}{\sigma}; y_i) \tag{4.19}$$

The estimation is performed by alternately optimizing (4.19) over σ and ξ. For the optimization over ξ, (4.19) can be written as function from ξ with a secondary condition:

$$min\ f(\xi) = \frac{1}{n} \sum_{i=1}^{n} \varphi(\frac{z_i' \xi}{\sigma}; y_i) \text{ with } g(\xi) = \xi' \xi - 1 = 0 \tag{4.20}$$

f can be approximated by $f(\hat{\xi}_0 + h) \approx f(\hat{\xi}_0 + \mathrm{grad} f(\hat{\xi}_0)' h$ close to the initial solution $\hat{\xi}_0$ for (4.20). With $\mathrm{grad} g(\hat{\xi}_0) = 2\hat{\xi}_0$ and $h = -\mathrm{grad} f(\hat{\xi}_0) + |\hat{\xi}_0' \mathrm{grad} f(\hat{\xi}_0)| \hat{\xi}_0$ the updated estimate $\hat{\xi}_1$ equals $\hat{\xi}_0 + \frac{\epsilon h}{\|h\|}$ with $\epsilon > 0$.

Overall, the following steps are performed in order to find the BY-estimators.

1. Selection of a subset with exclusion by means of the robust Mahalanobis distance.
2. Determination of the starting values $\gamma_0^{intercept}, \gamma_0^{(1)}, \ldots, \gamma_0^{(p)}$ for the coefficients by applying logistic regression on the "robust" subset.
3. The initial values $\sigma_0 = \frac{1}{\sqrt{\sum_{i=0}^{p} (\gamma_0^{(i)})^2}}$ and $\xi_0^{(i)} = \gamma_0^{(1)} \cdot \sigma_0$ for $i = 0, \ldots, p$ ($i = 0$ is for the intercept) are calculated.
4. Calculation of the initial value of the objective function $\hat{\gamma}_n^{BY}$ according to 4.15 with $score_0 = \frac{z_i' \xi_0}{\sigma_0}$.
5. Optimization of the objective function over σ_j.

6. Setting of $\gamma_j^{(i)} = \frac{\xi_{j-1}^{(i)}}{\sigma_j}$, $score_j = \frac{z_i'}{\sigma_j}$ and the objective function with $score_{j-1} = \frac{z_i' \xi_{j-1}}{\sigma_j}$.

7. Determination of $h = -\text{grad}f(\xi_{j-1}) + |\xi_{j-1}' \text{grad}f(\xi_{j-1})|\xi_{j-1}$, ξ_j and $score_j$.

8. Update of the average of the objective function with $score_j$.

9. Comparison of the mean values. If the update is greater than the average of the objective function with $score_{j-1}$, convergence is achieved, i.e. a local minimum has been reached at $(\sigma_j, \hat{\xi}_{j-1})$. If this inequality is not fulfilled, the procedure is repeated from step 5 to step 8 with the updated values.

10. If steps 5-8 were repeated for a predetermined maximum number of iterations (e.g. 1000) and convergence has failed, no estimators could be found.

4.4 Comparison of the approaches

The fitted models can be compared with the help of a coefficient of determination. McKelveys and Zavoinas pseudo R^2 is a measure for the explanatory power of a model fit.

$$R^2 = \frac{var(\hat{y})}{var(\hat{y}) + var(e)} \in [0, 1], \tag{4.21}$$

with the error e. The greater R^2, the better is the explanatory power of the model. It can be said, that the model with the larger R^2 is "better".

An additional key figure for comparing models of the Probability of default consider the discriminatory power: the Gini coefficient.

For this the Gini coefficient is calculated, as described in Section 2.3.1 with the difference, that the ROC curve is not determined for the expressions of a variable but for the vector of the fitted values.

$$fitted\ values = \begin{pmatrix} \frac{1}{1+e^{-(\beta_0+\beta_1 \cdot x_{1_1}+\cdots+\beta_m \cdot x_{m_1})}} \\ \frac{1}{1+e^{-(\beta_0+\beta_1 \cdot x_{1_2}+\cdots+\beta_m \cdot x_{m_2})}} \\ \vdots \\ \frac{1}{1+e^{-(\beta_0+\beta_1 \cdot x_{1_n}+\cdots+\beta_m \cdot x_{m_n})}} \end{pmatrix}$$

The model with the greater Gini coefficient has a better discriminatory power and thus can differentiate better between "good" and "bad" customers.

4.5 Multivariate regression

In contrast to the univariate regression, the multivariate regression bears *partial correlations* between the response variables $\boldsymbol{Y}$ given $\boldsymbol{X}$. These are equivalent to the correlations between the random errors $\boldsymbol{E}$ and can be seen in their covariance

matrix. It measures to what extent the prediction error $E^{(j)}$ of the jth model can be predicted from the error $E^{(k)}$ from the kth model through $\boldsymbol{X}$ and vice versa.

For estimating the impact of macroeconomic variables on different PDs a multivariate regression model is needed. The multivariate regression differs from the univariate, in that the respond is multidimensional, so that not only one variable y is considered, but q variables.

$$\boldsymbol{Y_i} = (Y_i^{(1)}, Y_i^{(2)}, \ldots, Y_i^{(q)}),\ 1 \leq i \leq n \tag{4.22}$$

where n is the number of observations in the sample. The explanatory p variables are given by $\boldsymbol{X_i} = (X_i^{(1)}, X_i^{(2)}, \ldots, X_i^{(p)})$, with $1 \leq i \leq n$.

The multivariate regression describes the relation or the impact of $\boldsymbol{X}$ on each of the responding variables. With the methods described in the previous sections, $p \cdot q$ parameters $\boldsymbol{B}$ are estimated at the same time, which leads to the following models:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E} \tag{4.23}$$

With $\boldsymbol{Y} \in \mathbb{R}^{n \times q}$, $\boldsymbol{X} \in \mathbb{R}^{n \times (p+1)}$, $\boldsymbol{B} \in \mathbb{R}^{(p+1) \times q}$ and $\boldsymbol{E} \in \mathbb{R}^{n \times q}$. The single models are obtained by regarding the jth column of $\boldsymbol{Y}$, $\boldsymbol{B}$ and the error $\boldsymbol{E}$ with $j \in \{0, \ldots, q\}$:

$$Y_i^{(j)} = \boldsymbol{X}\boldsymbol{B}^{(j)} + \boldsymbol{E}^{(j)},\ i = 1, \ldots, n$$

5 Empirical estimation

The generous provision of data by BAWAG P.S.K. made it possible to apply the described methods to a portfolio of private customers of an Austrian bank. In order to do justice to data protection and the banking secrecy the provided data were anonymized in advance and the original variable names were substituted by neutral designations (like *var1, var2, ...*).

5.1 Procedure

The aim was to try out the methods on real-life data and compare results for the stressed PD. This was realized in the following manner.

As described in Section 2.1, three types of products (loans, credit cards and current accounts) are provided in the available data. The mentioned analyses where implemented for all of the three types, but to avoid repetition the results of the PD estimation are only shown for loans.

The first step is the development of a PD model on product level by means of the following procedure:

1. Preparation of the available data, including the definition of the regarded period of time and the quantitative description.
2. Definition of the long list of variables.
3. Univariate analyses and variables selection.
4. Multivariate analyses and model fitting, both "classical" and robust.
5. Comparison of PD_{clas} and PD_{rob}.

The next step is the definition of the macro variables for the stress test and the preparation of the historical data. This also includes the calculation of the average PDs for the available points in time of the data history.

In the last step the stress test model is estimated for both, the PD_{clas} and the PD_{rob} as an input for the response. Finally the results can be compared.

5.2 Data preparation

As time horizon for the modeling sample a period of one year was chosen. In order to comply with the forecasting horizon of 365 days, the data were selected from the year 2014.

In total the raw data contain 1.233.015 data sets, i.e. loans, at 12 different point in times. The loans are considered at every month-end of the year and therefore contains duplicates for most of the data sets. In order to counteract the autocorrelation, each loan was drawn once from the sample, giving priority to the defaults.

Since it is possible that a customer has more than one loan, for each customer, the more current data set was chosen. After that the sample was unique per customer with 116.403 data sets. This sample forms the basis for the univariate and multivariate analyses and is distributed as follows.

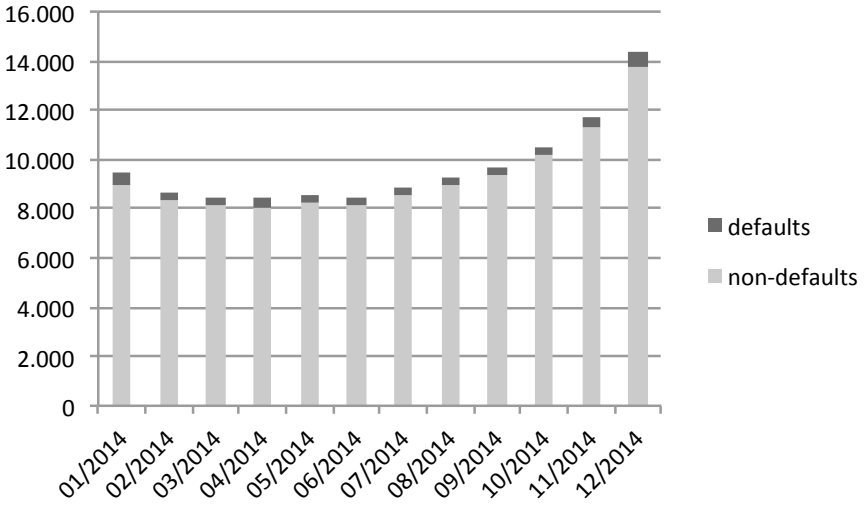

Figure 20: Distribution of the analysis-sample

The long list originally contains 23 variables, with product related information like dunnings, redemptions or arrears. The following issues were analyzed for those variables:

- missing values
- default values
- GINI coefficient
- distribution along the course of the variable
- linearity in the log odds
- working hypothesis
- outliers

The data preparation and the univariate analyses were performed in the program IBM SPSS Modeler 16.0[5] and in R[6]. First, all variables were checked for missing values and default values in order to exclude affected variables from further analyses.

[5] Data mining tool by IBM Corp., URL http://www-01.ibm.com/software/analytics/spss/products/modeler

[6] R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

	missing values	default values
var1	0%	0%
var2	0%	0%
var3	0%	0%
var4	0%	21%
var5	0%	0%
var6	0%	0%
var7	0%	0%
var8	0%	0%
var9	0%	0%
var10	0%	0%
var11	0%	0%
var12	30%	-
var13	0%	0%
var14	0%	27%
var15	0%	0%
var16	0%	0%
var17	0%	0%
var18	0%	0%
var19	0%	0%
var20	0%	0%
var21	0%	0%

Figure 21: Missing values and default values

As it can be seen in Figure 21, var12 has an amount of $> 20\%$ of missing values and is therefore excluded from the long list. var4 and var14 aren't filled meaningful for more than 20% and therefore excluded as well. In total, 18 variables are left for the analysis of the discriminatory power.

The GINI coefficient was calculated for the whole sample. Since the database contains data from one year, the comparison of the GINI coefficients over the time is not conclusive. The GINI coefficient was calculated as described in Section 2.3.1. The verification was, if the absolute value of the GINI $< 0,05$, this was an indication for the exclusion of the variable.

In this context the variables var3, var10, var11 and var18 were excluded. For the leftover variables, the distribution of the default rate along the course of each variable was analyzed. For this, the variable was classified into vingtiles. For each class the default rate was calculated. As an example, Figure 23 shows the distribution of the default rate of var1.

	total GINI	abs(GINI) < 5%
var1	-11%	no
var2	-5%	no
var3	**3%**	**yes**
var5	-5%	no
var6	-5%	no
var7	**-4%**	**yes**
var8	-5%	no
var9	-5%	no
var10	**-4%**	**yes**
var11	**-4%**	**yes**
var13	-56%	no
var15	32%	no
var16	15%	no
var17	5%	no
var18	**1%**	**yes**
var19	-8%	no
var20	-6%	no
var21	-6%	no

Figure 22: Total GINI coefficient

Figure 23: Distribution of the default rate along the course of var1

It can be seen that the trend of the default rate decreases with a higher value of var1 which corresponds to the working hypothesis of the variable. Still the course of the variable shows some peaks and troughs, for example in the 3rd, the 6th, the 13th and the 16th and 17th vingtile. This could mean that the variable has to be transformed in order to fulfill the precondition of linearity in the log odds.

The linearity assumption was examined graphically and with the Box-Tidwell test (see Menard, 2002).For the graphical test var1 was classified into vingtiles and plotted against the log odds.

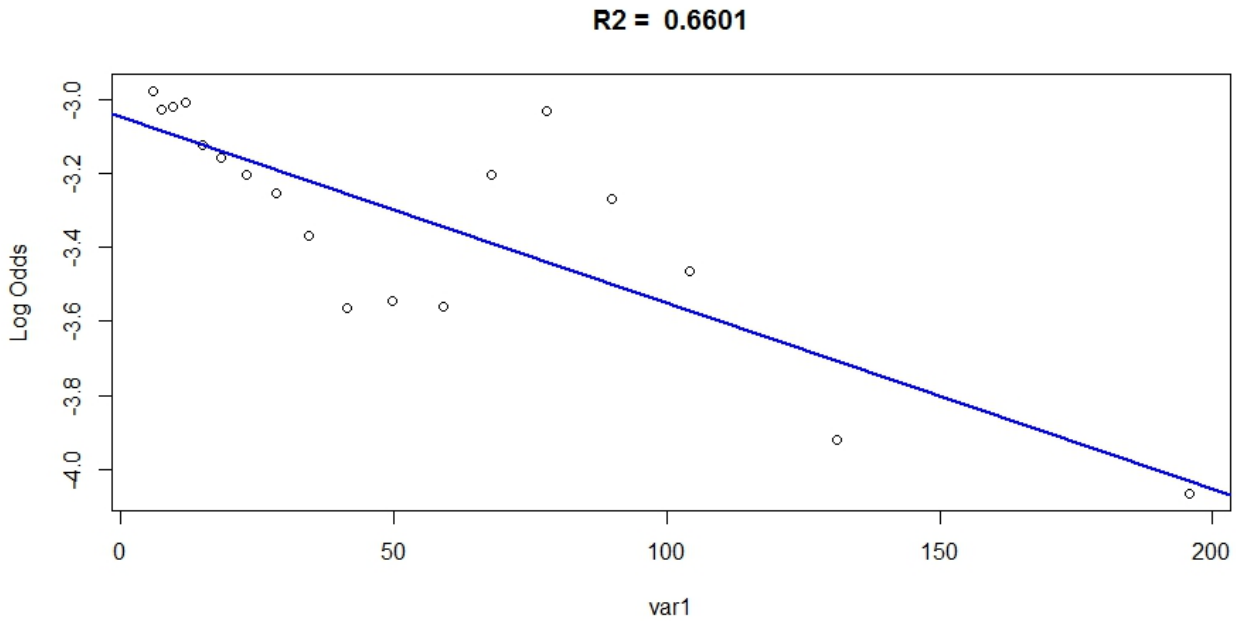

Figure 24: Vingtiles of var1 against log odds

As already assumed in Figure 23, the peak and the trough distorts the course of var1 and probably as well the linearity. The presumption is confirmed by the Box-Tidwell test. Performing the Box-Tidwell test with the approach of Bianco-Yohai leads to significant results as well.

```
> box_tidwell_test_var1<-glm(y~x1+x1_ln_x1,family="binomial")
> summary(box_tidwell_test_var1)

Call:
glm(formula = y ~ x1 + x1_ln_x1, family = "binomial")

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.3071  -0.2993  -0.2790  -0.2532   2.8229

Coefficients:
                        Estimate       Std. Error      z value
                Pr(>|z|)
(Intercept)     -2.960097       0.040365        -73.334         < 2e-16 ***
x1              -0.015629       0.004513        -3.463          0.000534 ***
x1_ln_x1         0.004974       0.002052        2.424           0.015361 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 37378  on 116402  degrees of freedom
Residual deviance: 37230  on 116400  degrees of freedom
AIC: 37236

Number of Fisher Scoring iterations: 6
```

```
> box_tidwell_test_rob_var1
$convergence
[1] TRUE

$objective
[1] 0.4358217

$coefficients
 Intercept                      x1
             x1_ln_x1
-2.5574792    -0.0637659     0.0261462

$cov
            [,1]         [,2]         [,3]
[1,]  9.635119e-04 -7.872713e-05  3.159717e-05
[2,] -7.872713e-05  9.195936e-06 -3.791916e-06
[3,]  3.159717e-05 -3.791916e-06  1.575616e-06

$sterror
[1] 0.031041    0.003032480  0.001255235

$iter
[1] 2

> p-value
   x1_ln_x1
3.491362e-96
```

In order to fulfill the linearity assumption, the variable needs to be transformed. As suitable transformation, the following function was determined:

$$var1_{trafo} = 0.5 \cdot ln(var1^{0.5})$$

The Box-Tidwell test, both with the Maximum-Likelihood and the Bianco-Yohai approach confirm the linearity assumption for the transformation of var1.

```
> summary(box_tidwell_test_var1_trafo)

Call:
glm(formula = y ~ x1 + x1_ln_x1, family = "binomial")

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.3160  -0.2991  -0.2727  -0.2517   2.7508

Coefficients:
                Estimate Std.  Error           z value         Pr(>|z|)
(Intercept)     -2.8304          0.3524          -8.031          9.65e-16 ***
x1               -1.8017         0.1587          -11.356          < 2e-16 ***
x1_ln_x1         -1.5118         2.4405          -0.619          0.536
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 37378  on 116402  degrees of freedom
Residual deviance: 37226  on 116400  degrees of freedom
AIC: 37232

Number of Fisher Scoring iterations: 6
```

```
> box_tidwell_test_rob_var1_trafo
$convergence
[1] TRUE

$objective
[1] 0.4357155

$coefficients
Intercept       x1              x1_ln_x1
-2.148093       -1.705603       3.191111

$cov
          [,1]      [,2]      [,3]
[1,] 0.10708780 0.01104141 0.7293526
[2,] 0.01104141 0.02241358 0.1237327
[3,] 0.72935260 0.12373274 5.0884407

$sterror
[1] 0.3272427 0.1497117 2.2557572

$iter
[1] 17

> significance
 x1_ln_x1
0.1571733
```

Also graphically the linearity is confirmed.

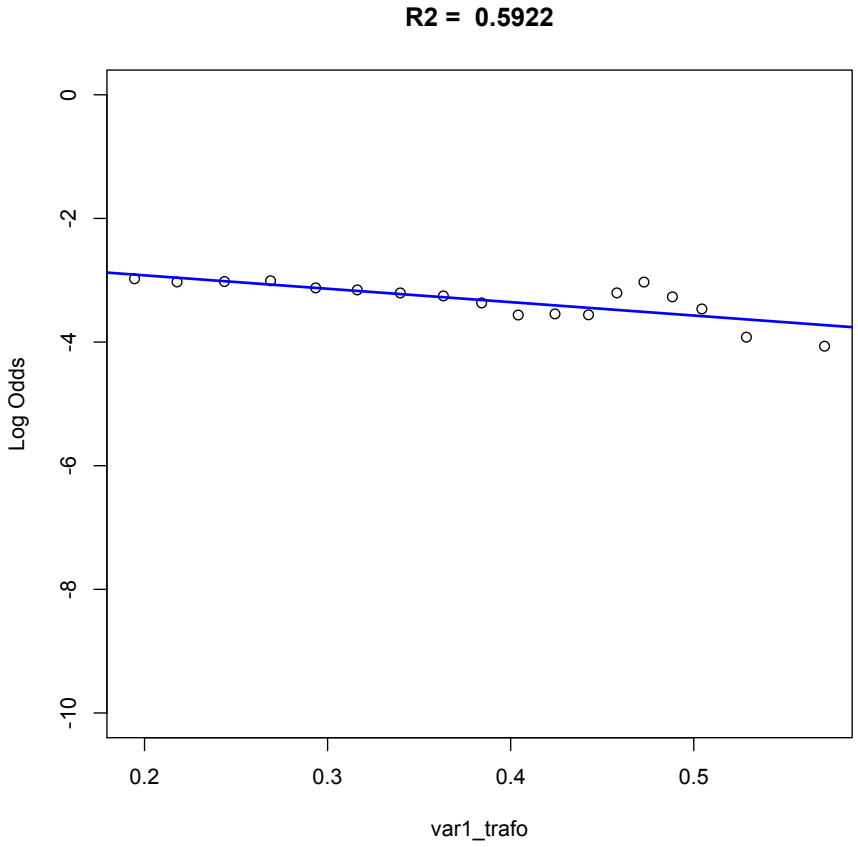

Figure 25: Vingtiles of var1$_{trafo}$ against log odds

This test of the linearity assumption and probably a transformation was performed for the other variables as well. Table 26 shows the final selection of the variables and the corresponding transformation.

In a next step the correlation between the transformed variables, remaining in the final selection, was examined. As it can be seen in the correlations matrix, there is no correlation between the variables, except for var20 and var21. Those two variables should not be together in one model, which leads to the following combinations as first tries for a model:

- *var1, var2, var8, var13, var15, var19, var20*
- *var1, var2, var8, var13, var15, var19, var21*

	working hypothesis	hypothesis fulfilled	linearity assumption	transformation
var1	the bigger, the better	yes	not fulfilled	0.5*ln(var1^0.5)
var2	the bigger, the better	yes	not fulfilled	sin(0.5*var2^5)
var5	the bigger, the better	no	n.r.	n.r.
var6	the bigger, the better	no	n.r.	n.r.
var8	the bigger, the better	yes	not fulfilled	sin(-0.5*var8^3)
var9	the bigger, the better	no	n.r.	n.r.
var13	the smaller, the better	yes	not fulfilled	sin(-8*var13)
var15	the smaller, the better	yes	not fulfilled	sin(2*var15)
var16	the smaller, the better	no	n.r.	n.r.
var17	the smaller, the better	no	n.r.	n.r.
var19	the bigger, the better	yes	not fulfilled	var19^2
var20	the bigger, the better	yes	not fulfilled	sin(-var20)
var21	the bigger, the better	yes	not fulfilled	sin(-var21)

Figure 26: Final variable selection with transformations

	var1	var2	var8	var13	var15	var19	var20	var21
var1	100%	-1%	0%	11%	-18%	4%	0%	0%
var2	-1%	100%	0%	0%	0%	0%	-1%	0%
var8	0%	0%	100%	0%	1%	0%	0%	0%
var13	11%	0%	0%	100%	-14%	2%	0%	0%
var15	-18%	0%	1%	-14%	100%	-1%	0%	0%
var19	4%	0%	0%	2%	-1%	100%	0%	0%
var20	0%	-1%	0%	0%	0%	0%	100%	61%
var21	0%	0%	0%	0%	0%	0%	61%	100%

Figure 27: Spearmans rank correlation

Since the variables influence the PD multivariate, the data set is tested for multivariate outliers. With the approach of the Mahalanobis distance multivariate outliers were determined and excluded from the sample which reduces the sample to 115.106 data sets.

Again, the linearity assumption was proofed for the reduced data sample without the multivariate outliers with the following result:

	working hypothesis	hypothesis fulfilled	linearity assumption	transformation
var1	the bigger, the better	yes	not fulfilled	ln(var1)
var2	the bigger, the better	yes	not fulfilled	sin(var2^2)
var8	the bigger, the better	yes	not fulfilled	sin(var8)
var13	the smaller, the better	yes	not fulfilled	wasn't found
var15	the smaller, the better	yes	not fulfilled	sin(var15^3)
var19	the bigger, the better	yes	not fulfilled	ln(var19^+1)
var20	the bigger, the better	yes	not fulfilled	(-var20)^5

Figure 28: Final variable selection with transformations after excluding multivariate outliers

Compared to Figure 26 the transformations are different. This might lead to totally different models. Also the correlations changed as var1 and var15 aren't correlated anymore. Instead the correlation of var1 and var20 increased.

	var1	var2	var8	var15	var19	var20
var1	100%	0%	1%	-1%	4%	20%
var2	0%	100%	0%	0%	0%	5%
var8	1%	0%	100%	1%	0%	1%
var15	-1%	0%	1%	100%	-1%	0%
var19	4%	0%	0%	0%	100%	0%
var20	-20%	0%	0%	1%	5%	100%

Figure 29: Spearmans rank correlations after excluding multivariate outliers

5.3 The PD-model

In a first step all the final variables included into the model with the expectation, that not all the variables will be significant.

As assumed not all variables are significant. For both, the Maximum Likelihood (ML) and the Bianco Yohai (BY) approach var2, var8, var19 and var20 aren't significant in this combination. The next step was to exclude the less significant variable from the model and perform the regression again. In total the logistic regression was performed for about 30 different combinations of variables. The result of the first model from Figure 30 was confirmed in every try, as var1, var13 and var15 were always significant.

	COEFF_ML	ERROR_ML	t-VAL_ML	SIG_ML	COEFF_BY	ERROR_BY	t-VAL_BY	SIG_BY
Intercept	0,1022	0,0027	38,2860	0,0000	-2,3039	0,0817	-28,2037	0,0000
var1	-0,1081	0,0067	-16,1300	0,0000	-1,5685	0,1975	-7,9432	0,0000
var2	-0,0004	0,0009	-0,4660	0,6410	-0,0043	0,0273	-0,1588	0,8738
var8	0,0008	0,0009	0,9010	0,3670	0,0341	0,0269	1,2670	0,2052
var13	0,0415	0,0011	37,5900	0,0000	0,8219	0,0261	31,4953	0,0000
var15	0,0104	0,0011	9,2920	0,0000	0,2497	0,0332	7,5106	0,0000
var19	0,0000	0,0000	-1,2930	0,1960	-0,0021	0,0150	-0,1378	0,8904
var20	-0,0012	0,0009	-1,2330	0,2170	-0,0375	0,0272	-1,3774	0,1684

Figure 30: Coefficients of the first model

	COEFF_ML	ERROR_ML	t-VAL_ML	SIG_ML	COEFF_BY	ERROR_BY	t-VAL_BY	SIG_BY
Intercept	0,1021	0,0027	38,2680	0,0000	-1,8233	0,0693	-26,3260	0,0000
var1	-0,1082	0,0067	-16,1480	0,0000	-3,0019	0,1904	-15,7635	0,0000
var13	0,0415	0,0011	37,5700	0,0000	0,8514	0,0256	33,2292	0,0000
var15	0,0104	0,0011	9,2860	0,0000	0,2087	0,0344	6,0636	0,0000

Figure 31: Coefficients of the final model for loans

It is conspicuous, that although both approaches select the same variables as significant, the BY-estimates have a much bigger influence but also bigger standard errors. Still the ranking of the variables concerning the influence, is the same except the intercept.

In order to compare the models, McKelvey and Zavoinas Pseudo R^2 was calculated as well as the Gini coefficient.

	ML	BY
PseudoR2	0,0002	0,0953
Gini	0,3829	0,3758
PseudoR2_Test	0,0002	0,0942
Gini_Test	0,3527	0,3527

Figure 32: Comparison of the models for loans

Both performance indicators were calculated for the training sample (70% of the the whole sample - 81.569 data sets) and the test sample (30% of the the whole sample - 34.834 data sets) separately. Figure 32 shows that both indicators are very similar for the training and the test sample. This is an indication for the stability of the model since the partition into training and test sample was done randomly.

Since the Gini coefficient considers the rank of a variable and the related default rates, the results are comparable for the ML and the BY approach. Matters are quite different when it comes to the Pseudo R^2. As McKelvey and Zavoina use the variance, which is a measure that is very sensitive to outliers, the R^2 of the

ML model is significantly smaller than the one of the BY model. The results are comparable for the other products.

	COEFF_ML	ERROR_ML	t-VAL_ML	SIG_ML	COEFF_BY	ERROR_BY	t-VAL_BY	SIG_BY
Intercept	0,0020	0,0004	4,8630	0,0000	-6,3610	0,2256	-28,1968	0,0000
var1	0,0000	0,0000	-5,3450	0,0000	-0,0072	0,0016	-4,5402	0,0000
var4	-0,0025	0,0003	-7,6610	0,0000	-1,0323	0,2328	-4,4350	0,0000

Figure 33: Coefficients of the final model for credit cards

	ML	BY
PseudoR2	0,0000	0,1152
Gini	0,3090	0,3095
PseudoR2_test	0,0000	0,1168
Gini_test	0,2428	0,2388

Figure 34: Comparison of the models for credit cards

	COEFF_ML	ERROR_ML	t-VAL_ML	SIG_ML	COEFF_BY	ERROR_BY	t-VAL_BY	SIG_BY
Intercept	0,0412	0,0025	16,6110	0,0000	-3,8276	0,0563	-67,9997	0,0000
var2	-0,0188	0,0011	-16,3420	0,0000	-0,3390	0,0235	-14,4368	0,0000
var4	0,0564	0,0007	79,3380	0,0000	1,3596	0,0209	64,9949	0,0000
var5	0,0072	0,0014	5,1630	0,0000	-0,1622	0,0293	-5,5384	0,0000

Figure 35: Coefficients of the final model for current accounts

	ML	BY
PseudoR2	0,0008	0,2720
Gini	0,5494	0,5335
PseudoR2_Test	0,0008	0,2718
Gini_Test	0,5443	0,5482

Figure 36: Comparison of the models for current accounts

5.4 The stress test model

The stress test model is based on the idea of including macro economic factors for the prediction of the probability of default. The data base consists of the following historical Austrian macro economic factors:

- Real Gross Domestic Product (GDP)
- Consumer Price Index (CPI)
- Unemployment Rate (UR)
- Residential Property Prices (RPP)
- Equity Prices (EP)

The information is available on a quarterly basis and is considered from the beginning of 2006, which makes 36 data points in total.

date	GDP	CPI	UR	RPP	EP
31.03.06	3,7	1,43	5,2	4,6	822
30.06.06	3,5	2,03	4,8	4,5	793
30.09.06	3,8	1,80	4,5	5,1	784
31.12.06	4,2	1,50	4,4	2,2	864
31.03.07	4,4	1,77	4,4	4,1	931
30.06.07	4,1	1,87	4,6	5,3	985
30.09.07	3,2	1,93	4,6	5,3	884
31.12.07	2,9	3,20	4,1	4,0	847
31.03.08	2,9	3,23	3,9	1,8	704
30.06.08	2,3	3,70	3,5	-1,0	762
30.09.08	0,6	3,70	3,7	0,3	596
31.12.08	-2,2	2,27	4,1	3,3	312
31.03.09	-4,7	1,07	4,4	4,5	284
30.06.09	-5,0	0,10	4,8	5,7	356
30.09.09	-3,2	-0,07	5,1	3,2	438
31.12.09	-0,8	0,60	4,8	2,2	450
31.03.10	0,6	1,30	4,5	5,6	456
30.06.10	1,7	1,77	4,5	5,6	446
30.09.10	2,3	1,67	4,4	6,3	451
31.12.10	2,9	2,00	4,2	7,3	500
31.03.11	4,0	2,97	4,3	4,6	532
30.06.11	3,9	3,70	4,1	1,3	516
30.09.11	2,4	3,80	3,9	6,0	422
31.12.11	1,3	3,70	4,2	5,0	363
31.03.12	0,8	2,70	4,1	10,7	402
30.06.12	0,5	2,23	4,4	15,6	377
30.09.12	0,7	2,40	4,5	11,9	382
31.12.12	0,7	2,90	4,6	11,5	426
31.03.13	0,3	2,60	4,9	4,8	450
30.06.13	0,1	2,23	4,7	5,0	437
30.09.13	0,3	1,97	5,0	4,7	451
31.12.13	0,7	1,67	5,0	4,1	485
31.03.14	1,0	1,47	4,9	4,7	481
30.06.14	0,9	1,41	5,1	5,2	463
30.09.14	1,1	1,49	4,7	4,9	471
31.12.14	1,0	1,32	5,0	5,4	490

Figure 37: Historical macro economic factors

In order to get more data sets for the regression model a disaggregation of the time series is performed from quarterly data to monthly data. This increases the number of observations to 108 data sets.

date	GDP	CPI	UR	RPP	EP
31.01.06	3,8	1,29	5,3	4,7	827
28.02.06	3,7	1,40	5,2	4,6	823
31.03.06	3,6	1,60	5,1	4,5	815
30.04.06	3,5	1,92	4,9	4,3	804
31.05.06	3,5	2,08	4,8	4,4	793
30.06.06	3,5	2,09	4,7	4,8	783
31.07.06	3,7	1,94	4,6	5,5	773
31.08.06	3,8	1,80	4,5	5,4	779
30.09.06	3,9	1,66	4,4	4,4	800
31.10.06	4,1	1,51	4,4	2,6	836
30.11.06	4,2	1,47	4,4	1,9	866
31.12.06	4,3	1,52	4,4	2,1	890
31.01.07	4,4	1,67	4,4	3,3	906
28.02.07	4,4	1,78	4,4	4,2	929
31.03.07	4,4	1,86	4,4	4,8	957
30.04.07	4,3	1,90	4,5	5,1	992
31.05.07	4,1	1,89	4,6	5,3	995
30.06.07	3,8	1,82	4,7	5,5	968
31.07.07	3,5	1,70	4,7	5,5	911
31.08.07	3,2	1,84	4,6	5,4	876
30.09.07	3,0	2,25	4,5	5,0	865
31.10.07	2,9	2,91	4,2	4,6	877
30.11.07	2,9	3,29	4,1	4,0	858
31.12.07	2,9	3,40	4,0	3,4	806
31.01.08	3,0	3,22	4,0	2,8	724
29.02.08	2,9	3,18	3,9	1,9	688
31.03.08	2,8	3,29	3,8	0,8	700
30.04.08	2,6	3,53	3,6	-0,5	758
31.05.08	2,3	3,72	3,5	-1,2	776
30.06.08	1,9	3,85	3,5	-1,2	752
31.07.08	1,4	3,93	3,6	-0,6	687
31.08.08	0,7	3,78	3,7	0,2	602
30.09.08	-0,2	3,39	3,8	1,3	499
31.10.08	-1,2	2,77	4,0	2,5	377
30.11.08	-2,2	2,24	4,1	3,4	298
31.12.08	-3,2	1,80	4,2	3,9	262
31.01.09	-4,1	1,43	4,3	4,1	269
28.02.09	-4,8	1,07	4,4	4,4	282
31.03.09	-5,2	0,70	4,5	5,0	301
30.04.09	-5,3	0,33	4,7	5,8	327
31.05.09	-5,1	0,07	4,8	5,9	355
30.06.09	-4,7	-0,10	4,9	5,4	386
31.07.09	-4,0	-0,16	5,1	4,1	420
31.08.09	-3,2	-0,11	5,1	3,1	442
30.09.09	-2,4	0,06	5,1	2,3	452
31.10.09	-1,5	0,34	4,9	1,7	449
30.11.09	-0,7	0,61	4,8	1,9	449
31.12.09	-0,2	0,85	4,7	3,0	451
31.01.10	0,2	1,08	4,6	4,8	456
28.02.10	0,6	1,30	4,5	5,9	457
31.03.10	1,0	1,51	4,5	6,2	455
30.04.10	1,4	1,71	4,5	5,7	449
31.05.10	1,7	1,80	4,5	5,5	445
30.06.10	2,0	1,80	4,5	5,6	443

date	GDP	CPI	UR	RPP	EP
31.07.10	2,1	1,69	4,5	5,9	442
31.08.10	2,3	1,65	4,4	6,3	449
30.09.10	2,5	1,68	4,3	6,8	462
31.10.10	2,6	1,78	4,2	7,4	483
30.11.10	2,9	1,97	4,2	7,5	501
31.12.10	3,2	2,26	4,2	7,0	516
31.01.11	3,7	2,64	4,3	6,0	527
28.02.11	4,1	2,98	4,3	4,7	533
31.03.11	4,2	3,29	4,3	3,1	536
30.04.11	4,2	3,56	4,2	1,1	534
31.05.11	4,0	3,73	4,1	0,7	520
30.06.11	3,5	3,81	4,0	2,1	494
31.07.11	2,9	3,79	3,9	5,1	455
31.08.11	2,4	3,79	3,9	6,5	421
30.09.11	1,9	3,82	4,0	6,4	391
31.10.11	1,6	3,87	4,1	4,8	365
30.11.11	1,3	3,76	4,2	4,5	357
31.12.11	1,1	3,47	4,2	5,7	367
31.01.12	0,9	3,01	4,1	8,3	395
29.02.12	0,8	2,66	4,1	10,7	407
31.03.12	0,7	2,42	4,1	13,1	404
30.04.12	0,5	2,28	4,3	15,4	386
31.05.12	0,5	2,21	4,4	16,1	375
30.06.12	0,5	2,20	4,5	15,3	370
31.07.12	0,6	2,25	4,5	12,8	373
31.08.12	0,7	2,38	4,5	11,5	380
30.09.12	0,8	2,57	4,5	11,4	393
31.10.12	0,8	2,83	4,5	12,5	412
30.11.12	0,7	2,95	4,6	12,0	427
31.12.12	0,6	2,92	4,7	10,0	439
31.01.13	0,4	2,75	4,9	6,4	449
28.02.13	0,3	2,59	4,9	4,3	452
31.03.13	0,2	2,46	4,9	3,7	449
30.04.13	0,1	2,33	4,7	4,6	440
31.05.13	0,1	2,22	4,7	5,1	435
30.06.13	0,1	2,13	4,7	5,3	436
31.07.13	0,2	2,06	4,9	5,0	441
31.08.13	0,3	1,97	5,0	4,7	450
30.09.13	0,4	1,87	5,1	4,4	462
31.10.13	0,6	1,76	5,0	4,1	478
30.11.13	0,7	1,66	5,0	4,0	487
31.12.13	0,8	1,58	4,9	4,1	490
31.01.14	1,0	1,52	4,9	4,4	487
28.02.14	1,0	1,47	4,9	4,7	482
31.03.14	1,0	1,43	5,0	5,0	475
30.04.14	0,9	1,40	5,1	5,2	466
31.05.14	0,9	1,40	5,1	5,3	462
30.06.14	0,9	1,43	5,0	5,2	461
31.07.14	1,1	1,50	4,8	4,9	465
31.08.14	1,1	1,51	4,6	4,8	471
30.09.14	1,1	1,47	4,7	4,9	477
31.10.14	1,0	1,37	4,9	5,2	485
30.11.14	1,0	1,31	5,0	5,4	491
31.12.14	1,0	1,28	5,1	5,5	494

Figure 38: Disaggregated time series of macro economic factors

In Figure 38 the high frequency time series of all macro economic factors can be seen. The interpolation of the values was performed with the approach of Denton-Cholette, as described in Chapter 3.2. These time series will represent the independent variables in the stress test model.

The dependent variables for the stress test model equate the estimated PDs resulting from the PD models. In order to get PDs for every month, the average values of the input variables of each model were considered, leading to the following input:

date	var1 (loans)	var13 (loans)	var15 (loans)	var1 (cards)	var4 (cards)	var2 (accounts)	var4 (accounts)	var5 (accounts)
31.01.06	64,2	96,5	98,1	88,3	10,2	84,5	30,5	60,9
28.02.06	47,5	95,9	100,0	88,2	8,0	74,4	29,2	43,2
31.03.06	47,3	95,2	95,8	83,4	10,8	82,1	33,4	48,1
30.04.06	38,4	99,9	101,6	82,6	10,1	85,4	36,0	42,8
31.05.06	46,7	100,4	102,4	85,9	10,6	80,7	33,0	55,9
30.06.06	44,9	93,2	95,9	87,7	9,4	80,2	32,0	45,1
31.07.06	50,8	98,7	99,3	91,8	10,3	89,0	36,1	62,4
31.08.06	47,3	89,7	97,0	96,2	7,6	72,1	33,8	49,3
30.09.06	50,1	91,1	100,9	82,2	9,1	70,3	33,3	55,3
31.10.06	47,2	88,1	96,2	97,9	10,1	88,1	35,6	60,6
30.11.06	51,4	98,9	102,5	98,2	10,6	76,7	32,3	64,1
31.12.06	36,4	91,4	95,5	86,1	8,5	81,7	37,0	60,3
31.01.07	50,4	98,6	101,2	94,0	10,4	71,8	33,3	57,7
28.02.07	51,5	98,7	101,3	82,7	9,4	87,0	32,2	51,8
31.03.07	49,2	93,8	99,0	82,9	8,5	71,1	33,0	57,6
30.04.07	36,8	91,5	97,7	92,0	9,4	86,8	30,6	60,7
...	...	...	...	...	...	...	...	...
30.06.14	48,3	96,6	100,1	93,2	9,2	84,2	30,9	48,3
31.07.14	47,6	96,4	99,7	92,2	8,8	85,2	30,5	49,4
31.08.14	45,2	96,7	99,9	92,3	9,2	86,3	30,5	43,8
30.09.14	43,5	96,9	100,2	92,5	9,5	84,5	29,8	43,9
31.10.14	41,0	97,0	99,8	91,1	9,4	82,4	30,7	53,5
30.11.14	38,7	97,1	100,0	90,0	9,6	76,4	30,5	40,6
31.12.14	33,9	97,2	100,2	85,7	9,3	71,8	31,5	94,1

Figure 39: Section of the time series of input values for the PD models

To get the final dependent values for the estimation, the variables in Figure 39 were transformed in accordance with the transformations mentioned above. Inserting into the following equation leads to the estimated PDs for each model.

$$PD = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m)}} \tag{5.1}$$

date	PD - ML (loans)	PD - BY (loans)	PD - ML (cards)	PD - BY (cards)	PD - ML (accounts)	PD - BY (accounts)
31.01.06	52,377%	9,025%	50,114%	0,931%	50,636%	1,463%
28.02.06	50,978%	3,120%	50,111%	0,885%	50,176%	0,888%
31.03.06	50,161%	1,631%	50,115%	0,908%	51,124%	2,411%
30.04.06	50,241%	1,767%	50,114%	0,893%	51,418%	2,837%
31.05.06	52,224%	8,286%	50,114%	0,921%	51,105%	2,213%
30.06.06	52,108%	7,610%	50,113%	0,913%	50,829%	2,235%
31.07.06	52,578%	10,634%	50,114%	0,956%	51,370%	2,987%
31.08.06	50,321%	1,851%	50,111%	0,929%	51,243%	2,410%
30.09.06	51,438%	4,485%	50,113%	0,871%	51,205%	2,302%
...	...	...	...	...	...	...
30.06.14	51,835%	6,107%	50,113%	0,945%	50,729%	1,625%
31.07.14	52,667%	11,416%	50,112%	0,929%	50,677%	1,434%
31.08.14	50,921%	2,990%	50,113%	0,937%	50,593%	1,505%
30.09.14	51,052%	3,333%	50,113%	0,945%	50,395%	1,252%
31.10.14	52,008%	7,052%	50,113%	0,935%	50,575%	1,696%
30.11.14	52,493%	10,200%	50,113%	0,930%	50,540%	1,682%
31.12.14	52,657%	11,618%	50,113%	0,896%	50,859%	1,902%

Figure 40: Section of the time series of final PD's for the stress test model

Figure 40 shows a part of the PD's for both, the ML estimation and the BY estimation for each product. It is conspicuous, that the PDs of the models applying the ML estimation vary widely from the BY-PDs. These differences can be lead back to the results in Section 5.3. As indicated in the comparison of each model, the Pseudo-R^2, thus the percentage of variance explained by the model is significantly lower for the models with the Maximum Likelihood approach.

For estimating stressed PDs, a linear regression was performed with the macro economic factors as independent variables and the respective PD as dependent variable. For every of the six dependent variables, different constellations of predictors were tried as input for the regression. The following models were the most significant results.

In every model only one macro variable could be identified as significant for the prediction of the respective PD. The adjusted R^2 was again very low for each model but comparable for the ML-PD and the BY-PD. The coefficients of the variables are, as in the logistic regression models, more influential for the robust PDs than for the non-robust ones.

```
> summary(lm(pd_loans_ML~cpi))

Call:
lm(formula = pd_loans_ML ~ cpi)

Residuals:
Min       1Q    Median       3Q      Max
-0.0148374 -0.0061279  0.0004362  0.0068582  0.0133560

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.5170809  0.0017181 300.963   <2e-16 ***
cpi       -0.0013469  0.0007424  -1.814   0.0725 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.007579 on 106 degrees of freedom
Multiple R-squared:  0.03011, Adjusted R-squared:  0.02096
```

```
> summary(lm(pd_loans_BY~cpi))

Call:
lm(formula = pd_loans_BY ~ cpi)

Residuals:
Min       1Q   Median       3Q      Max
-0.047023 -0.023430 -0.005828  0.023789  0.061582

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.063820   0.006571   9.712  2.5e-16 ***
cpi       -0.005388   0.002840  -1.897   0.0605 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02899 on 106 degrees of freedom
Multiple R-squared:  0.03284, Adjusted R-squared:  0.02372
```

Figure 41: Stress test models for the PD of the loans

```
> summary(lm(pd_cards_ML~ur))

Call:
lm(formula = pd_cards_ML ~ ur)

Residuals:
Min       1Q    Median       3Q      Max
-4.966e-05 -1.274e-05  5.069e-06  1.386e-05  3.134e-05

Coefficients:
Estimate Std. Error   t value Pr(>|t|)
(Intercept) 5.011e-01  2.057e-05 24359.975  < 2e-16 ***
ur        1.264e-05  4.555e-06    2.775  0.00653 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.946e-05 on 106 degrees of freedom
Multiple R-squared:  0.06771, Adjusted R-squared:  0.05891
```

```
> summary(lm(pd_cards_BY~cpi))

Call:
lm(formula = pd_cards_BY ~ cpi)

Residuals:
Min       1Q    Median       3Q      Max
-1.087e-03 -3.391e-04 -7.480e-06  2.510e-04  9.320e-04

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.454e-03  9.671e-05  97.759  < 2e-16 ***
cpi       -1.386e-04  4.179e-05  -3.317  0.00125 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0004266 on 106 degrees of freedom
Multiple R-squared:  0.09402, Adjusted R-squared:  0.08547
```

Figure 42: Stress test models for the PD of the credit cards

```
> summary(lm(pd_accounts_ML~rpp))

Call:
lm(formula = pd_accounts_ML ~ rpp)

Residuals:
Min       1Q    Median       3Q      Max
-0.0076051 -0.0027202  0.0006048  0.0026570  0.0058887

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.102e-01  6.039e-04 844.761   <2e-16 ***
rpp         -1.747e-04  9.978e-05  -1.751   0.0828 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.003302 on 106 degrees of freedom
Multiple R-squared:  0.02811, Adjusted R-squared:  0.01894
```

```
> summary(lm(pd_accounts_BY~rpp))

Call:
lm(formula = pd_accounts_BY ~ rpp)

Residuals:
Min       1Q    Median       3Q      Max
-0.0140122 -0.0057101 -0.0000711  0.0050343  0.0130104

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0246770  0.0012091  20.410   <2e-16 ***
rpp         -0.0003858  0.0001998  -1.932   0.0561 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.00661 on 106 degrees of freedom
Multiple R-squared:  0.034, Adjusted R-squared:  0.02489
```

Figure 43: Stress test models for the PD of the current accounts

Since the scenarios for the respective years are very similar, the PDs are as well. The results vary the same as for the initial models.

Finally the models were applied to the scenarios indicated in Figure 8, resulting in the following stressed PDs:

loans	PD - ML (loans)			PD - BY (loans)		
	2014	***2015***	***2016***	***2014***	***2015***	***2016***
baseline scenario	51,533%	51,533%	51,533%	5,681%	5,681%	5,680%
adverse scenario	51,534%	51,534%	51,534%	5,684%	5,684%	5,685%

cards	PD - ML (cards)			PD - BY (cards)		
	2014	***2015***	***2016***	***2014***	***2015***	***2016***
baseline scenario	50,117%	50,117%	50,117%	0,927%	0,927%	0,927%
adverse scenario	50,117%	50,117%	50,117%	0,927%	0,927%	0,927%

accounts	PD - ML (accounts)			PD - BY (accounts)		
	2014	***2015***	***2016***	***2014***	***2015***	***2016***
baseline scenario	50,921%	50,920%	50,920%	2,248%	2,248%	2,246%
adverse scenario	50,927%	50,926%	50,921%	2,263%	2,260%	2,249%

Figure 44: Final stressed PDs for the different scenarios

6 Conclusion

Section 5.2 shows how modeling is dependent from the underlying data. Although the preconditions of logistic regression are by far not as restrictive as they are for other methods, the eligible variables and therefore possible models were reduced to very small selection. Especially the linearity assumption is challenging in conjunction with the used data sample.

The difference of the reduced sample without multivariate outliers is not significant concerning the selection of the variables. Noticeable are the different transformations for fulfilling the linearity assumption and the changing correlations. Since the Mahalanobis distance is a multivariate measure for outliers, it is not surprising that this effects the correlations on a multivariate level.

As the data preparation, also the modeling itself confirms that the different methods don't differ in the selection of the variables within the model but the contribution of each variable to the model. The ranking of the contributions is comparable for the the logistic regression with the approach of Maximum Likelihood and the approach of Bianco Yohai, but with Bianco Yohai a much bigger contribution was achieved as the estimated coefficients have bigger values.

When comparing the models with the help of the Gini coefficient, no significant difference was visible, as the Gini coefficient considers the ranking of a variable and the respective defaults. Regarding the Pseudo-R^2 the differences in the used methods become obvious. Since the variables make a bigger contribution when using BY-estimation, the explained variance is bigger as well.

Still both models could be improved even more by optimizing the data sample (e.g. by expanding the period of time) or by constructing more variables.

The differences of the two approaches became really obvious when regarding the fitted values of the times series for the stress test models, as the values PDs vary enormously for the Maximum Likelihood estimators and the Bianco Yohai estimators. This difference is due to the impact of the variables to the estimation of the coefficients. This effect finds itself again in the result of the stress test models, as the adjusted R^2 is bigger for the model with the BY-PD as dependent variable.

Also the stress test models could be improved by decomposing the macro variables, using time lags or other economic indicators for predicting the PD.

Appendix

Time series of input values for the PD models (compare Figure 39):

date	var1 (loans)	var13 (loans)	var15 (loans)	var1 (cards)	var4 (cards)	var2 (accounts)	var4 (accounts)	var5 (accounts)
31.01.06	64,2	96,5	98,1	88,3	10,2	84,5	30,5	60,9
28.02.06	47,5	95,9	100,0	88,2	8,0	74,4	29,2	43,2
31.03.06	47,3	95,2	95,8	83,4	10,8	82,1	33,4	48,1
30.04.06	38,4	99,9	101,6	82,6	10,1	85,4	36,0	42,8
31.05.06	46,7	100,4	102,4	85,9	10,6	80,7	33,0	55,9
30.06.06	44,9	93,2	95,9	87,7	9,4	80,2	32,0	45,1
31.07.06	50,8	98,7	99,3	91,8	10,3	89,0	36,1	62,4
31.08.06	47,3	89,7	97,0	96,2	7,6	72,1	33,8	49,3
30.09.06	50,1	91,1	100,9	82,2	9,1	70,3	33,3	55,3
31.10.06	47,2	88,1	96,2	97,9	10,1	88,1	35,6	60,6
30.11.06	51,4	98,9	102,5	98,2	10,6	76,7	32,3	64,1
31.12.06	36,4	91,4	95,5	86,1	8,5	81,7	37,0	60,3
31.01.07	50,4	98,6	101,2	94,0	10,4	71,8	33,3	57,7
28.02.07	51,5	98,7	101,3	82,7	9,4	87,0	32,2	51,8
31.03.07	49,2	93,8	99,0	82,9	8,5	71,1	33,0	57,6
30.04.07	36,8	91,5	97,7	92,0	9,4	86,8	30,6	60,7
31.05.07	48,4	91,6	98,9	83,4	10,6	74,3	33,4	50,7
30.06.07	51,1	92,5	99,4	81,5	9,5	79,5	31,6	41,0
31.07.07	50,2	93,5	96,2	95,7	10,1	80,0	34,4	56,5
31.08.07	43,3	95,0	95,2	97,7	10,2	73,1	31,7	51,0
30.09.07	35,8	98,2	100,6	85,9	9,2	79,1	35,2	60,2
31.10.07	48,0	96,2	97,9	81,3	11,0	80,1	29,9	45,3
30.11.07	44,9	88,6	98,7	95,6	9,4	79,4	35,3	48,6
31.12.07	47,5	98,2	100,3	89,1	6,6	70,7	34,7	64,4
31.01.08	36,6	87,8	96,2	95,5	8,7	82,6	31,0	46,0
29.02.08	40,5	97,3	97,9	96,5	6,7	81,9	35,3	57,2
31.03.08	41,0	91,4	96,7	80,4	8,8	74,2	32,0	63,2
30.04.08	37,5	98,3	101,8	90,8	7,0	81,4	31,9	58,6
31.05.08	46,8	95,2	101,5	95,6	6,1	84,0	32,7	53,0
30.06.08	38,6	90,7	96,8	90,6	7,2	75,5	30,5	43,4
31.07.08	41,8	89,4	101,2	86,9	6,2	81,2	32,5	61,9
31.08.08	38,9	89,5	95,3	83,5	6,6	76,9	33,3	64,2
30.09.08	49,6	96,8	102,4	80,9	8,9	72,0	29,9	52,2
31.10.08	37,0	95,2	96,1	93,9	9,5	78,1	34,2	63,8
30.11.08	46,2	95,4	100,3	91,5	7,3	78,2	33,9	41,7
31.12.08	41,6	87,1	95,8	90,2	6,5	80,2	36,2	53,9
31.01.09	39,3	87,7	99,8	86,3	7,8	71,7	34,9	61,5
28.02.09	36,7	98,1	100,4	85,4	9,2	74,6	33,2	55,4
31.03.09	41,5	89,0	95,0	89,5	10,5	72,1	33,7	46,2
30.04.09	50,8	93,2	97,8	97,7	9,7	74,2	30,2	52,9
31.05.09	38,2	91,9	98,7	98,3	10,8	77,5	35,9	52,7
30.06.09	48,1	92,9	95,5	95,8	6,2	84,5	33,0	62,6
31.07.09	49,0	95,6	99,7	96,1	10,0	72,9	30,0	54,9
31.08.09	41,5	97,0	97,2	97,1	7,8	81,9	34,2	62,4
30.09.09	40,3	92,9	95,3	94,9	9,9	88,6	30,2	51,1
31.10.09	47,4	98,8	101,8	89,5	8,3	85,0	30,3	47,1
30.11.09	48,3	96,3	99,8	94,1	10,0	84,0	33,3	64,2
31.12.09	51,1	92,7	99,7	85,5	10,3	87,9	32,3	61,3
31.01.10	34,8	97,0	97,2	88,7	9,9	71,2	31,2	60,9
28.02.10	39,3	90,0	101,5	85,6	8,6	85,4	33,9	53,7
31.03.10	42,0	92,6	101,6	89,7	10,4	74,2	37,0	54,8
30.04.10	39,0	96,5	98,6	92,6	9,0	83,2	32,7	54,4
31.05.10	42,6	96,2	96,6	93,6	9,2	80,9	35,0	50,0
30.06.10	50,0	95,0	102,7	93,8	6,3	80,5	35,6	56,8

date	var1 (loans)	var13 (loans)	var15 (loans)	var1 (cards)	var4 (cards)	var2 (accounts)	var4 (accounts)	var5 (accounts)
31.07.10	45,2	96,1	100,2	87,3	9,2	83,9	33,1	48,5
31.08.10	40,2	90,0	98,6	89,7	7,3	85,7	30,4	57,5
30.09.10	42,6	96,6	99,9	97,1	7,5	77,2	29,5	61,2
31.10.10	43,6	97,8	99,1	96,7	9,9	81,7	34,0	63,1
30.11.10	38,9	88,7	101,3	95,0	6,8	71,6	36,2	52,7
31.12.10	50,0	93,9	97,4	80,2	7,8	75,5	33,9	51,4
31.01.11	49,5	100,3	101,9	81,9	9,5	88,1	36,5	64,2
28.02.11	47,0	92,1	102,7	95,5	9,2	85,8	35,8	57,2
31.03.11	50,5	94,6	99,5	88,0	8,4	80,7	33,5	46,3
30.04.11	51,2	88,9	98,6	91,9	8,9	70,9	35,1	44,5
31.05.11	51,2	90,5	101,7	90,5	7,7	73,3	34,6	56,8
30.06.11	54,0	95,6	98,5	93,0	7,5	70,2	36,9	41,5
31.07.11	47,3	92,1	99,0	95,9	10,0	83,3	33,0	42,1
31.08.11	34,2	94,3	99,7	80,2	9,3	84,1	33,5	41,3
30.09.11	43,7	88,7	98,7	89,5	10,8	71,1	35,9	63,7
31.10.11	46,0	89,2	99,7	82,0	11,0	88,4	30,0	52,8
30.11.11	38,1	88,0	101,6	85,9	8,6	81,5	31,2	56,6
31.12.11	48,4	91,2	100,9	80,9	8,2	87,2	32,1	60,5
31.01.12	46,4	96,1	100,4	94,2	10,1	75,5	30,8	46,4
29.02.12	42,5	93,8	101,6	93,3	10,4	71,3	30,0	62,9
31.03.12	36,5	94,2	100,2	82,0	9,4	74,9	30,0	48,6
30.04.12	36,7	93,7	96,9	92,4	10,3	74,7	33,2	48,0
31.05.12	39,3	92,7	99,8	95,5	10,2	88,3	33,3	56,1
30.06.12	39,0	100,0	100,6	94,8	7,4	79,4	32,1	46,3
31.07.12	46,8	91,2	100,1	98,3	7,5	77,7	30,5	49,5
31.08.12	39,6	88,5	98,5	80,7	9,8	72,0	29,7	45,7
30.09.12	43,1	87,5	95,1	96,6	6,4	71,8	30,2	42,7
31.10.12	48,2	96,8	99,2	96,9	10,7	75,1	29,5	61,3
30.11.12	45,7	95,2	99,0	86,7	7,2	79,5	35,8	59,2
31.12.12	45,3	92,4	97,5	92,5	6,4	79,7	30,3	59,2
31.01.13	49,7	93,1	102,2	86,5	10,0	70,6	30,4	44,7
28.02.13	39,1	90,6	95,7	98,3	10,3	70,7	31,7	46,9
31.03.13	36,4	93,6	100,5	85,0	10,1	76,8	34,1	42,7
30.04.13	39,8	90,8	96,8	93,9	9,8	74,6	29,6	58,3
31.05.13	47,0	88,2	100,7	88,4	10,9	71,6	30,1	49,1
30.06.13	52,0	93,7	98,0	86,8	10,7	74,0	30,0	62,7
31.07.13	46,6	87,3	96,3	81,5	6,5	88,1	35,2	57,5
31.08.13	34,2	96,6	98,9	93,8	6,9	74,1	33,3	59,3
30.09.13	47,2	97,2	102,6	81,8	6,6	74,6	33,7	57,0
31.10.13	36,8	97,4	102,2	88,4	7,7	71,4	36,1	50,8
30.11.13	35,0	92,7	99,5	90,8	8,8	88,5	35,4	56,1
31.12.13	43,5	93,6	100,2	90,0	9,7	78,2	34,4	41,6
31.01.14	53,0	95,0	101,3	89,4	9,5	71,1	35,5	48,8
28.02.14	51,6	96,0	100,6	91,5	9,5	87,6	32,7	44,0
31.03.14	50,5	96,2	100,8	91,3	9,2	86,6	33,8	58,2
30.04.14	50,0	96,2	100,1	92,0	9,5	87,7	31,2	64,9
31.05.14	49,1	96,4	100,0	92,2	9,0	86,1	31,5	55,8
30.06.14	48,3	96,6	100,1	93,2	9,2	84,2	30,9	48,3
31.07.14	47,6	96,4	99,7	92,2	8,8	85,2	30,5	49,4
31.08.14	45,2	96,7	99,9	92,3	9,2	86,3	30,5	43,8
30.09.14	43,5	96,9	100,2	92,5	9,5	84,5	29,8	43,9
31.10.14	41,0	97,0	99,8	91,1	9,4	82,4	30,7	53,5
30.11.14	38,7	97,1	100,0	90,0	9,6	76,4	30,5	40,6
31.12.14	33,9	97,2	100,2	85,7	9,3	71,8	31,5	94,1

Time series of final PD's for the stress test model (compare Figure 40):

date	PD - ML (loans)	PD - BY (loans)	PD - ML (cards)	PD - BY (cards)	PD - ML (accounts)	PD - BY (accounts)
31.01.06	52,377%	9,025%	50,114%	0,931%	50,636%	1,463%
28.02.06	50,978%	3,120%	50,111%	0,885%	50,176%	0,888%
31.03.06	50,161%	1,631%	50,115%	0,908%	51,124%	2,411%
30.04.06	50,241%	1,767%	50,114%	0,893%	51,418%	2,837%
31.05.06	52,224%	8,286%	50,114%	0,921%	51,105%	2,213%
30.06.06	52,108%	7,610%	50,113%	0,913%	50,829%	2,235%
31.07.06	52,578%	10,634%	50,114%	0,956%	51,370%	2,987%
31.08.06	50,321%	1,851%	50,111%	0,929%	51,243%	2,410%
30.09.06	51,438%	4,485%	50,113%	0,871%	51,205%	2,302%
31.10.06	50,473%	2,098%	50,114%	0,995%	51,298%	2,938%
30.11.06	51,718%	5,584%	50,115%	1,007%	50,875%	2,339%
31.12.06	50,499%	2,185%	50,112%	0,882%	51,390%	3,382%
31.01.07	51,639%	5,254%	50,114%	0,974%	51,024%	2,678%
28.02.07	51,982%	6,835%	50,113%	0,880%	50,833%	2,272%
31.03.07	51,595%	5,050%	50,112%	0,864%	50,990%	2,577%
30.04.07	51,956%	6,837%	50,113%	0,940%	50,650%	1,542%
31.05.07	52,411%	9,429%	50,114%	0,905%	51,066%	2,595%
30.06.07	52,569%	10,560%	50,113%	0,874%	50,788%	2,053%
31.07.07	51,098%	3,438%	50,114%	0,979%	51,202%	2,749%
31.08.07	51,565%	5,019%	50,114%	0,994%	50,807%	2,116%
30.09.07	51,253%	3,971%	50,113%	0,896%	51,251%	3,007%
31.10.07	51,622%	5,174%	50,115%	0,898%	50,333%	1,404%
30.11.07	52,583%	10,766%	50,113%	0,964%	51,359%	2,753%
31.12.07	51,065%	3,347%	50,109%	0,856%	51,173%	3,095%
31.01.08	52,256%	8,634%	50,112%	0,949%	50,615%	1,844%
29.02.08	52,198%	8,158%	50,109%	0,906%	51,214%	3,094%
31.03.08	50,897%	2,971%	50,112%	0,854%	50,884%	2,175%
30.04.08	50,482%	2,150%	50,109%	0,878%	50,803%	2,201%
31.05.08	50,388%	1,959%	50,108%	0,883%	50,910%	2,421%
30.06.08	51,329%	4,202%	50,110%	0,881%	50,677%	1,482%
31.07.08	51,989%	6,977%	50,108%	0,833%	51,057%	2,041%
31.08.08	51,865%	6,385%	50,109%	0,825%	51,010%	2,664%
30.09.08	50,158%	1,622%	50,112%	0,859%	50,361%	1,440%
31.10.08	50,231%	1,759%	50,113%	0,956%	51,106%	2,871%
30.11.08	51,027%	3,254%	50,110%	0,891%	51,162%	2,597%
31.12.08	51,579%	5,091%	50,109%	0,863%	51,325%	3,265%
31.01.09	52,584%	10,884%	50,111%	0,870%	51,358%	2,672%
28.02.09	52,412%	9,655%	50,113%	0,893%	51,176%	2,245%
31.03.09	50,345%	1,909%	50,114%	0,944%	51,059%	2,785%
30.04.09	52,262%	8,427%	50,113%	0,986%	50,437%	1,542%
31.05.09	52,093%	7,547%	50,115%	1,010%	51,250%	3,336%
30.06.09	50,238%	1,733%	50,108%	0,888%	51,058%	2,280%
31.07.09	52,654%	11,280%	50,114%	0,979%	50,541%	1,268%
31.08.09	51,447%	4,578%	50,111%	0,938%	51,220%	2,563%
30.09.09	50,229%	1,744%	50,114%	0,969%	50,433%	1,516%
31.10.09	52,108%	7,579%	50,111%	0,900%	50,459%	1,545%
30.11.09	52,119%	7,577%	50,114%	0,966%	50,986%	2,614%
31.12.09	51,576%	4,969%	50,114%	0,914%	51,008%	1,968%
31.01.10	51,399%	4,465%	50,114%	0,928%	50,839%	1,758%
28.02.10	51,608%	5,226%	50,112%	0,881%	51,086%	2,682%
31.03.10	51,925%	6,646%	50,114%	0,943%	51,499%	3,173%
30.04.10	52,318%	8,936%	50,112%	0,936%	51,025%	2,221%
31.05.10	51,047%	3,340%	50,113%	0,947%	51,280%	2,819%
30.06.10	51,770%	5,826%	50,108%	0,877%	51,277%	3,103%

date	PD - ML (loans)	PD - BY (loans)	PD - ML (cards)	PD - BY (cards)	PD - ML (accounts)	PD - BY (accounts)
31.07.10	50,566%	2,255%	50,113%	0,906%	51,119%	2,252%
31.08.10	52,495%	10,168%	50,110%	0,879%	50,494%	1,618%
30.09.10	51,494%	4,724%	50,110%	0,931%	50,344%	1,047%
31.10.10	52,011%	7,024%	50,114%	0,983%	51,123%	2,693%
30.11.10	51,718%	5,695%	50,109%	0,901%	51,296%	3,471%
31.12.10	51,843%	6,137%	50,111%	0,832%	51,076%	2,818%
31.01.11	52,153%	7,820%	50,113%	0,876%	51,275%	3,435%
28.02.11	50,298%	1,822%	50,113%	0,959%	51,244%	3,208%
31.03.11	51,172%	3,616%	50,112%	0,893%	51,015%	2,684%
30.04.11	50,682%	2,444%	50,112%	0,930%	51,235%	3,085%
31.05.11	50,209%	1,685%	50,111%	0,893%	51,209%	2,873%
30.06.11	52,531%	10,230%	50,110%	0,903%	51,423%	3,394%
31.07.11	50,684%	2,463%	50,114%	0,979%	51,088%	2,227%
31.08.11	51,432%	4,572%	50,113%	0,863%	51,065%	2,567%
30.09.11	52,122%	7,640%	50,115%	0,949%	51,287%	3,360%
31.10.11	51,967%	6,771%	50,115%	0,902%	50,341%	1,423%
30.11.11	50,770%	2,705%	50,112%	0,884%	50,746%	1,822%
31.12.11	50,630%	2,368%	50,111%	0,845%	50,912%	2,068%
31.01.12	50,604%	2,321%	50,114%	0,969%	50,598%	1,806%
29.02.12	50,811%	2,771%	50,114%	0,969%	50,470%	1,401%
31.03.12	52,082%	7,534%	50,113%	0,875%	50,536%	1,266%
30.04.12	50,546%	2,263%	50,114%	0,961%	51,093%	2,429%
31.05.12	51,611%	5,204%	50,114%	0,981%	51,113%	2,328%
30.06.12	50,473%	2,120%	50,110%	0,913%	50,840%	2,276%
31.07.12	51,118%	3,492%	50,110%	0,938%	50,678%	1,442%
31.08.12	52,455%	9,886%	50,114%	0,874%	50,277%	1,330%
30.09.12	50,829%	2,806%	50,108%	0,899%	50,643%	1,354%
31.10.12	50,693%	2,477%	50,115%	0,999%	50,339%	1,030%
30.11.12	50,691%	2,483%	50,110%	0,859%	51,243%	3,305%
31.12.12	52,300%	8,753%	50,108%	0,872%	50,447%	1,564%
31.01.13	51,346%	4,188%	50,114%	0,915%	50,521%	1,613%
28.02.13	50,370%	1,959%	50,114%	1,001%	50,793%	2,132%
31.03.13	50,683%	2,521%	50,114%	0,908%	51,274%	2,466%
30.04.13	51,775%	5,937%	50,114%	0,962%	50,212%	1,252%
31.05.13	50,435%	2,027%	50,115%	0,943%	50,600%	1,298%
30.06.13	50,543%	2,188%	50,115%	0,929%	50,468%	1,324%
31.07.13	50,569%	2,267%	50,108%	0,808%	51,181%	3,103%
31.08.13	52,152%	7,955%	50,109%	0,895%	51,015%	2,652%
30.09.13	52,210%	8,186%	50,109%	0,813%	51,090%	2,686%
31.10.13	51,282%	4,072%	50,111%	0,880%	51,318%	3,344%
30.11.13	51,216%	3,848%	50,112%	0,919%	51,317%	2,833%
31.12.13	50,627%	2,376%	50,113%	0,933%	51,214%	2,742%
31.01.14	51,139%	3,536%	50,113%	0,924%	51,410%	2,812%
28.02.14	50,460%	2,054%	50,113%	0,938%	50,987%	2,267%
31.03.14	51,436%	4,474%	50,113%	0,932%	51,035%	2,747%
30.04.14	51,635%	5,215%	50,113%	0,943%	50,656%	1,923%
31.05.14	52,506%	10,129%	50,112%	0,933%	50,890%	1,771%
30.06.14	51,835%	6,107%	50,113%	0,945%	50,729%	1,625%
31.07.14	52,667%	11,416%	50,112%	0,929%	50,677%	1,434%
31.08.14	50,921%	2,990%	50,113%	0,937%	50,593%	1,505%
30.09.14	51,052%	3,333%	50,113%	0,945%	50,395%	1,252%
31.10.14	52,008%	7,052%	50,113%	0,935%	50,575%	1,696%
30.11.14	52,493%	10,200%	50,113%	0,930%	50,540%	1,682%
31.12.14	52,657%	11,618%	50,113%	0,896%	50,859%	1,902%

The following codes where implemented for all of the three product types, but to avoid repetition the results are only shown for loans.

R Code for Section 5.3

```
#Graphical depiction of the linearity of the logodds
logodds<-function(x,y,class,title) {
 nr_class<-floor(length(x)/class)
 x_sort<-sort(x)
 help<-x_sort[min(length(x_sort),nr_class)]
 x_sort<-x_sort[x_sort>help]
 while (length(x_sort)>0) {
  help<-c(help,x_sort[min(length(x_sort),nr_class)])
  x_sort<-x_sort[x_sort>help[length(help)]]
 }
 help[length(help)]<-max(x)
 logodd<-numeric(length(help))
 x_class<-numeric(length(help))
 logodd[1]<-mean(y[x<=help[1]])
 logodd[1]<-logoddg(logodd[1]/(1-logodd[1]))
 x_class[1]<-mean(x[x<=help[1]])
 for (i in 2:length(help)) {
  logodd[i]<-mean(y[x>help[i-1]&x<=help[i]])
  logodd[i]<-log(logodd[i]/(1-logodd[i]))
  if (logodd[i]=="-Inf"){logodd[i]<-0}
  x_class[i]<-mean(x[x>help[i-1]&x<=help[i]])
 }
 lm1<-lm(logodd~x_class)
 plot(x_class,logodd,ylab="Log Odds",title=title,main=paste("R2 = ",round(summary(lm1)$r.squared,digits=4)),ylim=c(-10,0))
 abline(lm1,col="blue",lwd=2)
}

logodds<-(var1,def,20,title="var1")



#Box-Tidwell test for one of the variables
x<-var1
x_ln_x<-x*log10(x)
box_tidwell_test_var1<-glm(def~x+x_ln_x,family="binomial")
summary(box_tidwell_test_var1)

#Transformation of the variables
var1_trafo<-sin(0.5*log10(var1^0.5))
var2_trafo<-sin(0.5*var2^5)
var8_trafo<-sin(-0.5*var8^3)
var13_trafo<-sin(-8*var13)
var15_trafo<-sin(-var15)
var19_trafo<-var19
var20_trafo<-sin(-var20)
var21_trafo<-sin(-var21)
def_trafo<-def
```

```
#Mahalanobis distance
m1_trafo<-cbind(var1_trafo,var2_trafo,var8_trafo,var15_trafo,var19_trafo,var20_trafo)

m1.mcd=covMcd(m1)
m1.md=sqrt(mahalanobis(m1,m1.mcd$center,m1.mcd$cov))
data1=cbind(def_trafo,var1_trafo,var2_trafo,var8_trafo,var15_trafo,var19_trafo,var20_trafo,m1.md)
write.table(data1, file = "MAH_m1.csv",row.names=FALSE, na="",col.names=TRUE, sep=",")



#ML-model
m1_trafo<-cbind(var1_trafo,var2_trafo,var8_trafo,var15_trafo,var19_trafo,var20_trafo)

gl_m1<-glm(def_trafo~m1_trafo)
summary(gl_m1)


#BY-model
m1_trafo<-cbind(var1_trafo,var2_trafo,var8_trafo,var15_trafo,var19_trafo,var20_trafo)

by_m1<-BYlogreg(m1_trafo, def_trafo)
by_m1
```

R Code for Section 5.4

```
pd_kred_gl<-c()
pd_kk_gl<-c()
pd_kto_gl<-c()
pd_kred_by<-c()
pd_kk_by<-c()
pd_kto_by<-c()
for(i in 1:length(gdp))
{
pd_kred_gl[i]<-1/(1+exp(-(int_gl_kred+coeff1_gl_kred*v1_kred[i]+coeff13_gl_kred*v13_kred[i]+coeff15_gl_kred*v15_kred[i])))
pd_kred_by[i]<-1/(1+exp(-(int_by_kred+coeff1_by_kred*v1_kred[i]+coeff13_by_kred*v13_kred[i]+coeff15_by_kred*v15_kred[i])))
pd_kk_gl[i]<-1/(1+exp(-(int_gl_kk+coeff1_gl_kk*v1_kk[i]+coeff4_gl_kk*v4_kk[i])))
pd_kk_by[i]<-1/(1+exp(-(int_by_kk+coeff1_by_kk*v1_kk[i]+coeff4_by_kk*v4_kk[i])))
pd_kto_gl[i]<-1/(1+exp(-(int_gl_kto+coeff2_gl_kto*v2_kto[i]+coeff4_gl_kto*v4_kto[i]+coeff5_gl_kto*v5_kto[i])))
pd_kto_by[i]<-1/(1+exp(-(int_by_kto+coeff2_by_kto*v2_kto[i]+coeff4_by_kto*v4_kto[i]+coeff5_by_kto*v5_kto[i])))
}



#Stress test models
mod_gl1<-lm(cbind(pd_kred_gl,pd_kk_gl,pd_kto_gl)~gdp+cpi+ur+rpp+ep)
mod_by1<-lm(cbind(pd_kred_by,pd_kk_by,pd_kto_by)~gdp+cpi+ur+rpp+ep)

summary(mod_gl1)
summary(mod_by1)
```

References

[Aut14] European Banking Authority. Main features of the 2014 eu-wide stress test. Technical report, European Banking Authority, 2014.

[Aut16] European Banking Authority. *EBA announces key features of the 2014 EU-wide Stress Test*, 2014 (accessed April 20, 2016).

[Beh99] A. Behr. *SAS für Ökonomen*. Oldenbourg Verlag München Wien, 1999.

[CH03] C. Croux and G. Haesbroeck. Implementing the Bianco and Yohai Estimator for Logistic Regression. In *Computational Statistics & Data Analysis*, volume 44, pages 273 – 295. Elsevier, 2003.

[ER06] B. Engelmann and R. Rauhmeier. *The Basel II Risk Parameters*. Springer Science & Business Media, Berlin Heidelberg, 2006.

[Fil04] P. Filzmoser. A Multivariate Outlier Detection Method. In *Proceedings of the Seventh International Conference on Computer Data Analysis and Modeling*, volume 1, pages 18 – 22. Belarusian State University, 2004.

[FT09] P. Filzmoser and V. Todorov. *Multivariate Robust Statistics - Methods and Computation*. Südwestdeutscher Verlag, Saarbr, 2009.

[HD10] M. Hubert and M. Debruyne. Minimum covariance determinant. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):36 – 43, 2010.

[HL00] D. W. Hosmer and S. Lemshow. *Applied Logistic Regression*. John Wiley & Sons, New York, 2000.

[Lan09] D. Lando. *Credit Risk Modeling - Theory and Applications*. Princeton University Press, Kassel, 2009.

[Men02] S. Menard. *Applied Logistic Regression Analysis* -. SAGE, London, 2 rev ed. edition, 2002.

[NT04] B. Nösslinger and G. Thonabauer. Ratingmodelle und Validierung. In *Leitfadenreihe zum Kreditrisiko*. Oesterreichische Nationalbank (OeNB) and Finanzmarktaufsicht (FMA), Wien, 2004.

[Rex16] A. Rexer. *Wer unbemerkt von der Bankenrettung profitierte*, 2014 (accessed April 9, 2016).

[RVD99] P. Rousseeuw and K. Van Driessen. A fast Algorithm for the Minimum Covariance Determinant Estimator. In *Techonometrics*, volume 41, pages 212 – 223. Taylor & Francis, Ltd., 1999.

[Tuk77] J. Tukey. *Exploratory Data Analysis* -. Addison-Wesley Publishing Company, Reading, 01. aufl. edition, 1977.

[VF16] K. Varmuza and P. Filzmoser. *Introduction to Multivariate Statistical Analysis in Chemometrics* -. CRC Press, Boca Raton, Fla, 2016.