UNIVERSITY OF TWENTE

MASTER THESIS

Non-Linearity Issues in Probability of Default Modelling

Author:
Lucas KLINKERS

Supervisor:
Dr. Berend ROORDA

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in

Industrial Engineering and Management

October 24, 2017

University of Twente

Abstract

Faculty of Behavioural, Management and Social sciences
Industrial Engineering and Management

Master of Science

Non-Linearity Issues in Probability of Default Modelling

by Lucas KLINKERS

Almost all the financial institutions that provide credit, estimate the client's probability of default, and the most widely used method in the industry is logistic regression. This process has very convenient characteristics, but a potential flaw exists in the restricting assumption of log-odds linearity. The purpose of this paper is to investigate the accuracy of predicting the probability of default with logistic regression and whether the linearity assumption is violated when multiple risk drivers are included in the model. Violation of the linearity assumption will cause a deviation between predicted PDs and observed PDs. Correcting for this deviation will increase the prediction accuracy of the PD model and therefore the regulatory capital calculation of the Rabobank will more accurately reflect the risks.

We suggest making an adjustment to the transformation of client score to PD. This adjustment allows us to identify whether the linearity assumption is violated and estimates the size of the correction that is needed. The great benefit is that the ranking performance based on the creditworthiness of the clients remains the same. The correction is applied before the transformation to probability, so only the absolute value of the PD is affected to improve the prediction accuracy. The average PD prediction error improved from 16% to 4% by correcting the log-odds. The calculated PDs for each clients will therefore represent the corresponding risks, which is essential for efficient capital allocation and RAROC measures.

Acknowledgements

This thesis has been written in order to obtain the Master's degree Financial Engineering & Management at the University of Twente. Most of the work has been done at the Risk Management department of the Rabobank.

The completion of this work would not have been possible without the help of my mentors at the Rabobank, Amrita Juhi & Viktor Tchistiakov. I sincerely enjoyed my time as an intern at their department and appreciate all the time they took to guide me through the process. From day one, I received all the data and resources I needed to shape my research. This freedom and their feedback led to the opportunity to present my thesis at the Eurobanking 2017 conference in Slovenia, which was a fantastic experience.

I would also like to thank my supervisors from the University of Twente, Berend Roorda & Reinoud Joosten. Their supervision and feedback had great impact on my work and their ideas provided me with perspectives from other angles.

Lucas Klinkers

Contents

Abstract iii

Acknowledgements v

1 Introduction 1
1.1 Background . . . 1
1.1.1 Rabobank Group . . . 1
1.1.2 Risk Management . . . 1
1.1.3 Credit Risk . . . 2
1.1.4 PD model . . . 2
Non-linearity . . . 2
1.2 Research proposal . . . 3
1.2.1 Problem statement . . . 3
1.2.2 Research goal . . . 3
1.2.3 Research Questions . . . 4
Sub-questions . . . 4
1.2.4 Outline . . . 4

2 Literature review 5
2.1 Non-Linearity of Scorecard log-odds . . . 5
2.1.1 Does non-linearity matter in credit risk modelling? . . . 8
2.2 Non-linear Logistic regression . . . 8
2.2.1 Generalized Partial Linear Model . . . 8
2.2.2 Principal component analysis . . . 8
2.3 Missing Values . . . 9
2.3.1 Missing value treatment and classifier accuracy . . . 9
Case Deletion . . . 9
Mean Imputation or Median Imputation . . . 9
KNN Imputation . . . 9
2.4 Identifying non-linearity . . . 10
2.4.1 Linearity assumption logistic regression . . . 10
2.4.2 Significance of regression parameters . . . 12
AIC and SBIC statistic . . . 13
2.4.3 Gini coefficient . . . 13
2.5 Impact on regulatory capital . . . 14

3 Data Analysis 17
3.1 Sample Data . . . 17
3.1.1 Causes of non-linearity in the data . . . 17
Difference in distributions between classes . . . 17
Individual risk drivers . . . 18
Correlation between risk drivers . . . 19
3.1.2 Actual vs. predicted log-odds . . . 20

4 Results **23**
4.1 Adjustment to PD transformation . . . 23
4.1.1 Identification of non-linearity . . . 24
4.1.2 Correction non-linearity . . . 24
Effect on performance . . . 24
AIC and SBIC test results . . . 26
4.2 Alternative methods . . . 27
4.2.1 Adjustment to score transformation . . . 27
Penalty for minimum risk driver . . . 27
Squared transformation of every risk driver . . . 28
4.2.2 Data transformation . . . 29
Principal component analysis . . . 29
4.3 Impact on regulatory capital calculation . . . 30

5 Missing values and non-linearity **31**
5.1 Approaches to missing value analysis . . . 31
5.1.1 Missing value bias . . . 31
Parameter estimation . . . 32
Unexpected losses . . . 33
Non-linearity and missing values . . . 34

6 Conclusion **35**

7 Discussion **37**
7.1 Limitations . . . 37
7.2 Suggestions for further research . . . 37

A Properties of logistic regression parameters **39**
A.1 Asymptotic Normality . . . 39

B Normally generated data **41**
B.1 Non-linearity in normally generated data . . . 41

Bibliography **43**

List of Figures

1.1 Risk Management Organogram (Rabobank Group, 2014) 1

2.1 Actual and inferred log-odds, logistic regression (Mcdonald, Sturgess, and Smith, 2012) 5
2.2 Transforming distribution of predictions (Mcdonald, Sturgess, and Smith, 2012) 6
2.3 Transforming inferred log-odds (Mcdonald, Sturgess, and Smith, 2012) 7
2.4 Logistic regression 10
2.5 Transformed explanatory variable 11
2.6 Effect of combing correctly transformed risk drivers 12
2.7 Income equality (Taylor, 1970) 13
2.8 Ranking Performance 14

3.1 Class distributions 18
3.2 Log-odds vs. risk drivers 18
3.3 Example of effect of correlation on MLE conversion 19
3.4 Comparison actual vs. inferred log-odds 20

4.1 Comparison actual vs. inferred log-odds 24
4.2 Actual vs. inferred log-odds squared model 29

5.1 Effect of method on indexed unexpected losses 33

A.1 Q-Q plot of the intercept versus $N(0,1)$ 39
A.2 Q-Q plot of the β parameter residuals versus standard normal distribution 40

B.1 Regression including and excluding correlated variables 41

List of Tables

1.1 Thesis outline 4

3.1 Mortgage statistics 17
3.2 Mortgage Correlation 19
3.3 PD prediction performance 21

4.1 Regression output if linearity holds 23
4.2 Regression output 24
4.3 Bucket performance 25
4.4 Indexed differences for PD, LGD & EAD 25
4.5 AIC and SBIC test results 26
4.6 Regression output of penalty model 27
4.7 Regression output 28
4.8 Regression output of squared model 28
4.9 PCA non-linearity identification 30

5.1 Effect of method on average indexed parameter 32
5.2 Imputing mean effect on non-linearity 34

Chapter 1

Introduction

Logistic regression is the most widely used method in the financial industry for estimating the probability of default. The interpretability of the model and the excellent performance in discriminating between creditworthy and uncreditworthy clients led to the widespread use in financial institutions. The problem with logistic regression is that a potential flaw exists in the restricting assumption of log-odds linearity. If the linearity assumption is violated, the accuracy of the model can be improved and the regulatory capital calculation of the Rabobank will more accurately reflect the risks.

1.1 Background

1.1.1 Rabobank Group

Rabobank Group is a cooperative financial services provider which offers retail banking, wholesale banking, private banking, leasing and real estate services. The group comprises of 103 local Rabobanks with over 475 branches within The Netherlands and international offices in forty countries. These local banks provide services to over 7 million customers and the international branches add another 1.2 million (Rabobank Group, 2014).

1.1.2 Risk Management

The Risk Management department of the Rabobank consists of three teams, each designated to a specialized function. The Credit Risk team is in charge of the Credit Portfolio through setting policies and limits, carrying out performance analyses, model construction and senior management reporting. The Balance Sheet Risk team is focused on asset & liability management, liquidity, funding, market risk and internal interest rates. The third team, Non-Financial Risk, manages the operational risk including operational continuity, IT risks and group insurance.

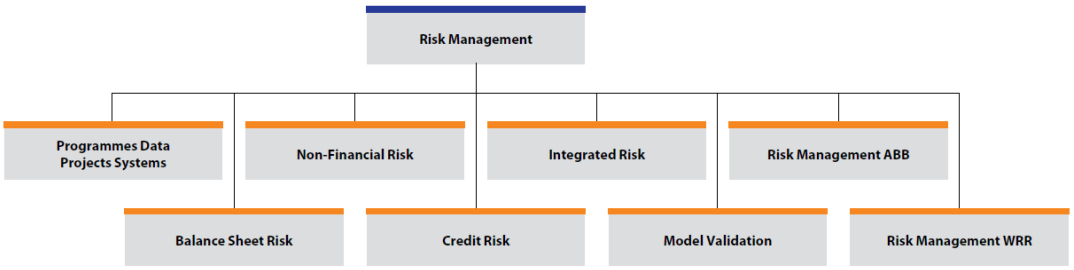


FIGURE 1.1: Risk Management Organogram (Rabobank Group, 2014)

1.1.3 Credit Risk

The definition used by Rabobank for Credit Risk is: "*The risk that a borrower/counterparty is unable to repay funds owed to the bank. Country risk and concentration risk are included in credit risk*". The credit risk is for the largest part calculated using the Advanced Internal Ratings Based (AIRB) approach, which uses regulatory capital formulas based on Probability of Default (PD), Loss Given Default (LGD), Exposure at Default (EAD) and Maturity (M). The regulatory capital formula results in the Risk Weighted Assets from which 8% is held as regulatory capital.

1.1.4 PD model

The creditworthiness of a client is assessed through a scorecard model. This is a regression model that produces a risk score based on risk drivers as input. Equation 1.1 is an example of a scorecard linear regression model.

$$score = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n \tag{1.1}$$

X_n = Transformed financial or qualitative factors
β_0 = The constant term and $\beta_1, ... , \beta_n$ are the factor weights

A calibration function transforms these scores into the probability of defaults. The definition of this probability of default in a business context is: "*The likelihood that a counterparty will default within 1 year*". Equation 1.2 is an example of logistic calibration function

$$PD = \frac{1}{1 + \exp^{-\alpha - \gamma score}} \tag{1.2}$$

There is a high correlation between the β parameters from Equation 1.1 and the α & γ parameters from Equation 1.2. If the score is unchanged from Equation 1.1 and the β parameters are estimated with maximum likelihood estimation for logistic regression, the α and γ parameters from the calibration will be redundant. In that case they will be estimated to be 0 and 1 respectively, which means that the calibration function is simply a non-linear transformation of the score, given by Equation 1.3. These α and γ parameters are used if the scores are transformed and/or if the β parameters are not estimated by logistic regression.

$$PD = \frac{1}{1 + \exp^{-score}} \tag{1.3}$$

Non-linearity

Logistic regression is used to map the dependent variable on a $[0, 1]$ scale based on the explanatory variables. This is useful for PD modelling, since probabilities are also on a $[0, 1]$ scale. The advantage of logistic regression is that it does not assume the risk drivers to be linearly related to the probability of default. The disadvantage is that the risk drivers have to be linearly related to the log-odds of PD, the linear-in-log-odds assumption. The log odds are defined as shown in Equation 1.4.

$$\text{Log-odds of PD} = log(\frac{PD}{1 - PD}) \tag{1.4}$$

Currently both continuous and categorical variables are transformed in the process of creating a PD model to ensure the validity of the assumption. For example, a five category risk driver, with scores {1,2,3,4,5}, can be transformed to have scores

{1,1.5,3,4,4.5} to ensure that an increase in category score has a linear effect on the log-odds. The problem is that even though the individual variables all satisfy the linear-in-log-odds assumption, the scores of the combined model can still have a non-linear relationship with the log-odds. The logistic regression will give a linear scalar to the model scores which fails to capture this non-linear relationship to the log-odds.

The consequence is that predictions of log-odds and therefore PD will be inaccurate. This PD is used for the pricing of products, loan provisioning and regulatory capital calculation. Inaccurate PD predictions may therefore lead to the mispricing of financial products and suboptimal capital allocation.

1.2 Research proposal

1.2.1 Problem statement

PD modelling can be divided in two parts, calculating the risk score and PD calibration. Part one consists of transferring clients risk drivers to a risk score. This risk score is used to rank clients from creditworthy to likely-to-default and is therefore used to discriminate bad clients from good ones. This is used in the decision to provide lending services to clients and their performance in separating good and bad clients is currently very high. The second part consists of transforming the risk score to a PD. This PD is used within the AIRB approach for calculating regulatory capital. The PD predictions do not always align with the observed defaults. The implication of over- or underestimating the PD can be inefficient capital allocation where the bank is holding too much or not enough capital for default losses. The logistic regression used within the calibration function assumes a linear relationship between risk drivers and log-odds, which is often violated and the cause of misleading PD results (Mcdonald, Sturgess, and Smith, 2012).

1.2.2 Research goal

The goal of my research is to investigate whether predicted PDs and observed defaults are consistently misaligned and what the magnitude of this deviation is. In the case of a misalignment the research extends to the investigation of the cause and whether the linear relationship assumption of logistic regression has been violated. Final solutions should provide guidance on how to correct for the deviation, and concurrently improve PD estimation. This will lead to more efficient capital allocation.

1.2.3 Research Questions

Main research question: *What is the effect of non-linearity on the accuracy of PD prediction and how can this be controlled for in loan portfolios?*

Sub-questions

- *Which causes of non-linearity in log odds are identified in literature?*
- *What is the magnitude of the deviation between predicted and observed default rates?*
  - *How to measure the magnitude of the deviation?*
  - *Is there a significant difference in predicted and observed default rates?*
  - *What is the cause of the deviation between predicted and observed default rates within Rabobank?*
- *How to incorporate non-linearity in PD modelling procedures?*
  - *How and when to test for non-linearity within the PD model development timeline?*
  - *What are the possibilities to correct for the non-linearity?*
- *Is there an impact on the regulatory capital?*

1.2.4 Outline

TABLE 1.1: Thesis outline

Chapter 2	Literature review
Chapter 3	Data-analysis
Chapter 4	Results and recommendations
Chapter 5	Missing value analysis
Chapter 6	Conclusions
Chapter 7	Discussion

Chapter 2

Literature review

2.1 Non-Linearity of Scorecard log-odds

Mcdonald, Sturgess, and Smith, 2012 investigated the accuracy of the inferred log-odds of an event, for example probability of default. Within a Credit Risk environment the scorecard model produces a creditworthiness score for every client based on certain characteristics. These scores rank every client from good to bad and determine whether credit can be provided. The scores can also be used to predict the actual probability of default with a logistic regression that yields the log-odds of default.

According to the findings of Mcdonald, Sturgess, and Smith, 2012 the ability to discriminate between good and bad clients is high. The problem lies in predicting the actual PD of these good and bad clients and this prediction tends to deviate from the actual probability. As seen in Figure 2.1, the inferred log-odds from the Lloyds Banking Group sample from the paper tends to overestimate the PD and the deviation from the inferred odds seems to be quadratic.

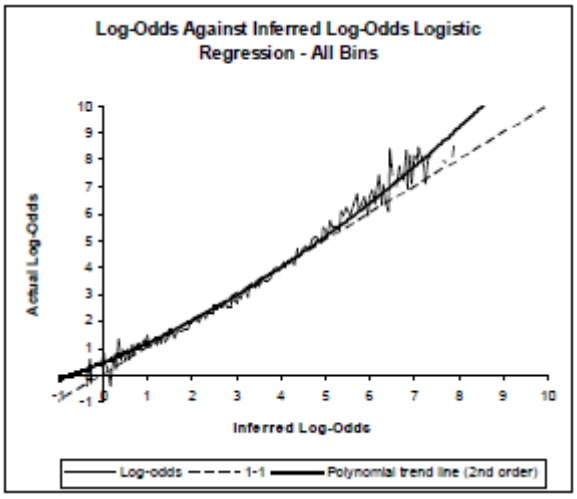
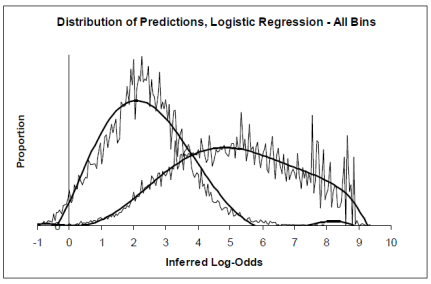


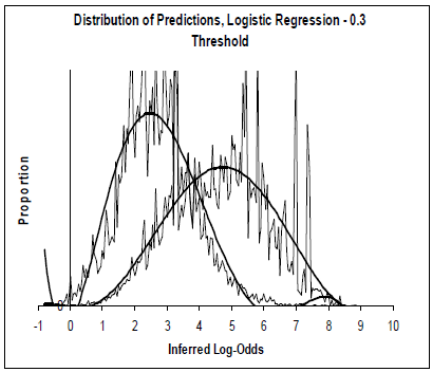
FIGURE 2.1: Actual and inferred log-odds, logistic regression (Mcdonald, Sturgess, and Smith, 2012)

Mcdonald, Sturgess, and Smith, 2012 show that the reason for the quadratic deviation of actual log-odds is due to bin correlation. Binning the variables is transforming the continuous variables into categorical variables. Each value on the continuous variable scale is assigned to a category. Bin correlation is the correlation between the different categorical variables. Bin correlation can be present due to

many different reasons, one example being 'missing value correlation'. Values from an individual that are missing in one characteristic are most likely also missing in another characteristic. Bin correlation produces skewed distributions as seen in Figure 2.2a. The figure shows the distributions of predicted log-odds for the two classes of data, clients who defaulted and client that did not default. The left distribution is the default-data and the right distribution is the no-default data. Figure 2.2b shows the distributions after removing correlated bins. It clearly shows that removing bins with a correlation higher than 0.3 produces less skewed distributions.



(A) All bins

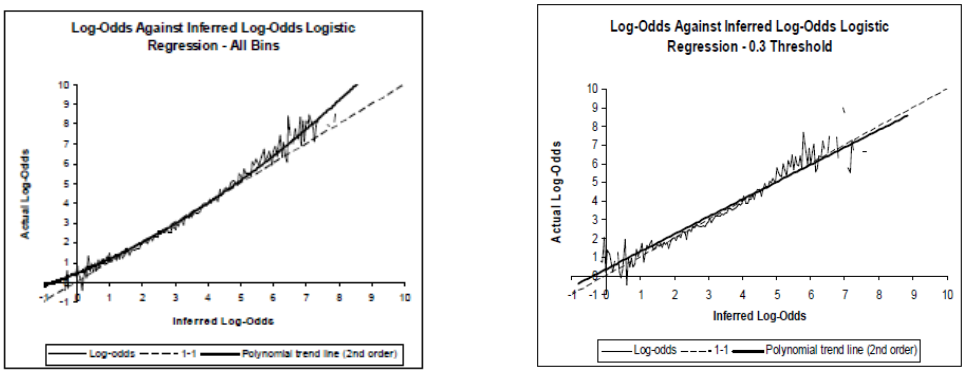


(B) Bins where correlation < 0.3

FIGURE 2.2: Transforming distribution of predictions (Mcdonald, Sturgess, and Smith, 2012)

Bin correlation specifically leads to the fact that coefficients of different characteristics can be used to 'dampen' each other, which in turn leads to different optimal solutions to the maximum likelihood equation of the logistic regression. In this case maximum likelihood optimization will find multiple sets of optimal solutions and will be unable to differentiate between these sets. Removing correlated bins also has an effect on the quadratic deviation as seen in Figure 2.3, but unfortunately also on the discriminatory power of the model due to removal of prediction power. Increasing the discriminatory power of the model will push the means of the two classes from Figure 2.2 apart, but the range of the distributions will remain largely the same. The change in means will therefore cause the distributions to become more skewed and therefore the linearity issues to become more present. A trade-off therefore exists between discriminatory power and accuracy of PD prediction. Mcdonald, Sturgess, and Smith, 2012 also found that PD models have higher quadratic deviations when their discriminatory power increases.

Mcdonald, Sturgess, and Smith, 2012 found that the curvature is due to a difference in variance between the two distributions. Given that the distribution are



(A) All bins (B) Bins where correlation < 0.3

FIGURE 2.3: Transforming inferred log-odds (Mcdonald, Sturgess, and Smith, 2012)

normally distributed then the following holds:

$$
\begin{aligned}
log(\frac{P(Class=2|s)}{P(Class=1|s)}) &= log(\frac{P(Class=2)}{P(Class=1)}) + log\frac{\sigma_1}{\sigma_2} + \frac{1}{2}((\frac{s-\mu_1}{\sigma_1})^2 - (\frac{s-\mu_2}{\sigma_2})^2) \\
&= log(\frac{P(Class=2)}{P(Class=1)}) + log\frac{\sigma_1}{\sigma_2} + \frac{1}{2}(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2}) + (\frac{\mu_2^2}{\sigma_2^2} - \frac{\mu_1^2}{\sigma_1^2})s + \frac{1}{2}(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2})s^2 \\
&= as^2 + bs + c \\
where& \\
a &= \frac{1}{2}(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}), \\
b &= (\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}), \\
c &= log(\frac{P(Class=2)}{P(Class=1)}) + log\frac{\sigma_1}{\sigma_2} + \frac{1}{2}(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2}) \\
s &= \text{model score}
\end{aligned}
$$

This means that the quadratic curvature of the model is based on the value of a, and this value increases as the difference between σ_1 and σ_2 gets larger. A larger difference between the variances of the two classes therefore will lead to more curvature. This explains why removing the curvature in one of the distributions led to a decrease in curvature, since skewed distributions have a higher variance. Mcdonald, Sturgess, and Smith, 2012 state that the amount of curvature is influenced by the discriminatory power of the model, due to higher score variances. This means that a stronger scorecard model will have more curvature. That begs the question whether the decrease in curvature from removing correlated bins in Figure 2.3 is due to the decrease in skewness of the model or simply because the model is less predictive and therefore has a higher variance. Unfortunately the authors did not investigate this.

The remedy according to Mcdonald, Sturgess, and Smith, 2012 is to perform a retrospective non-linear transformation to correct for the curvature. This solution does not affect the ranking performance, nor the Gini coefficient. The disadvantage would be that the remedy is retrospective. The correction will always be based on past data, so the correction might not be relevant for future cases. Changes in the underlying distribution and in particular changes in variance and skewness will

have an effect on the curvature of the deviation.

2.1.1 Does non-linearity matter in credit risk modelling?

Jagric, Kracun, and Jagric, 2011 analysed Slovenian banking data and studied whether non-linear relationships between credit risk and explanatory variables have a significant effect on model performance. Their research indicates that prediction models that include non-linear parts outperform the widely used logistic regression. The authors compared logistic regression with neural networks. Due to the improved handling of non-linear relationships and properties of categorical variables, their results show a significant increase of 8% in classification rate.

Lennox, 1999 investigated whether there are risk drivers in the Credit Risk model that have significant non-linear effects. He revealed that both cashflow and leverage have non-linear effects on the probability of default and that incorporating these effects improved the predictive accuracy.

2.2 Non-linear Logistic regression

2.2.1 Generalized Partial Linear Model

Müller and Härdle, 2003 examined the effects of adding a non-linear part to the logistic regression. They argued that the method of adding polynomial terms to logistic regression is an imprecise method of reflecting a non-linear relationship and an additional modelling step is needed to approximate the optimal polynomial order. The flexible method of representing a non-linear relationship is neural networks or classification trees, but these methods often do not reflect the underlying relationship between dependent and explanatory variables and are regarded as 'black box' style techniques.

Müller and Härdle, 2003 suggest to use a modification of the generalized linear model (GLM, Equation 2.2) from which the logit model is a special case. This generalized partial linear model (GPLM, Equation 2.3) preserves the 'easy to interpret' structure of logistic regression.

$$E(Y|X) = G(\beta' X) \tag{2.2}$$

$$E(Y|X) = G(\beta' X_1 + m(X_2)) \tag{2.3}$$

The link function G remains the logit function, which in the GLPM contains the parameter β and the non-parametric function $m(X_2)$. The explanatory variables matrix X is split up in two matrices, X_1 which is used in the parametric logit estimation and X_2 which is used in the non-parametric function. Function $m()$ is the smooth function, such as a kernel density function, which describes the effect of X_2 in a non-parametric fashion.

2.2.2 Principal component analysis

Mcdonald, Sturgess, and Smith, 2012 showed that a cause of the non-linearity is correlation between risk drivers. Principal Component Analysis (PCA) can be used to

transform the correlated risk drivers to uncorrelated principal components. Aguilera, Escabias, and Valderrama, 2006 performed a research in using PCA in combination with logistic regression and found that the model produced a similar goodness-of-fit with less model components.

2.3 Missing Values

A factor that strongly influences the correlation between the risk drivers of a PD model is missing values. This is due to the high probability that if one of the risk drivers is missing, other risk drivers are missing as well. This is one of the reasons that makes missing values a major problem in the financial industry and in particular PD modelling. There are many underlying causes of these missing values, including, but not limited to, fields that were not captured, discontinued fields, unavailability of the characteristic, intentionally not filled out by applicant or outliers that were removed. Statistical techniques such as random forests or decision trees are capable of handling datasets with missing values. Logistic regression needs a complete dataset, so either missing values have to be replaced or entries with missing values have to be removed (Siddiqi, 2006). Missing data rates of less than 1% are not a problem, 1-5% can be managed, 5-15% is problematic and >15% impacts any kind of interpretation (Acuña and Rodriguez, 2004).

2.3.1 Missing value treatment and classifier accuracy

Acuña and Rodriguez, 2004 identified four missing value treatments: Case Deletion (CD), Mean Imputation (MI), Median Imputation (MDI) and KNN Imputation (KNNI).

Case Deletion

Deleting all the cases with a missing value, optionally combined with deleting features with a high degree of missing values first. Sufficiently large sample size, low percentage of missing values and randomly generated missing values minimizes the effect of case deletion. If the missing data was not randomly generated, or the sample size is low, CD can produce biased estimates. According to Little and Rubin, 2002 CD should only be used when the missing data is completely randomly generated.

Mean Imputation or Median Imputation

MI involves replacing the missing values with the mean of the feature. The method is wildly used but the drawbacks consist of, but are not limited to, inflated sample size, underestimated variance, and negatively biased correlation. The advantage is that MI does give good performance in classification rates (Little and Rubin, 2002). MDI has the advantage over MI that it is not influenced by outliers.

KNN Imputation

KNN Imputation replaces the missing value using the cases that are most similar and do not miss the feature of interest. Advantages consist of flexibility in missing values in both categorical and continuous variables, multiple missing values within a case and correlation structure. Disadvantages are choice of distance function and

K-value, and required computational resources. KNNI outperforms the other methods, especially when the percentage of missing values increases. KNNI consists of the following algorithm (Acuña and Rodriguez, 2004):

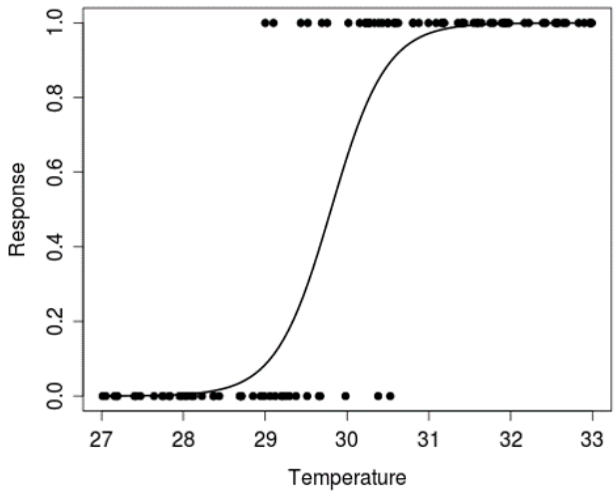
1. Divide the data set in two parts, D_m contains the cases with missing values and D_c contains all the complete cases.
2. For each case in D_m:
   (a) Divide the case in observed and missing vectors $x = [x_0; x_m]$
   (b) Calculate the distance between x_0 and the relevant features in D_c
   (c) Use K-nearest neighbours and perform majority voting estimate for the missing values x_m.

2.4 Identifying non-linearity

2.4.1 Linearity assumption logistic regression

Logistic regression is used to map the dependent variable on a $[0, 1]$ scale based on the explanatory variables. This is useful for PD modelling, since probabilities are also on a $[0, 1]$ scale. The second advantage is that the relationship between the explanatory variables and the dependent variable can be non-linear. This is often the case in practice were the marginal effect of increasing an explanatory variable is decaying as this variable increases. An example would be the effect of income on the PD. Improving the income of 0 to 100.000 has a larger effect on PD than improving from 100.000 to 200.000. This effect is shown in the example of Figure 2.4, where the probability of a chemical reaction is regressed on the explanatory variable temperature. The black line is the estimated probability off the chemical reaction for the temperature and the dots are the actual data points. The Figure clearly shows the non-linear relationship where the marginal effect starts to decay from a temperature of 30 and is almost obsolete after 31 degrees.

FIGURE 2.4: Logistic regression



Even though the relationship between the explanatory variables and the dependent binary variable can be non-linear, there is a still a very restrictive assumption on this relationship. The explanatory variables need to be linearly related to the log-odds of the dependent variable. This is shown in Equations 2.4, 2.5 & 2.6. 2.4 shows the logistic regression equation, 2.5 shows the odds ratio of the PD and 2.6 shows the equation of the log-odds. One can clearly see that the log-odds of the PD are a linear combination of the explanatory variables.

$$PD = \frac{1}{1 + e^{(\beta_0 + \beta_1 X_1 + ... + \beta_n X_n)}}, \tag{2.4}$$

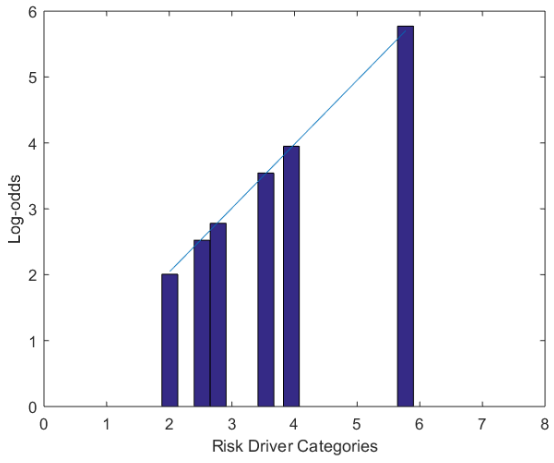
$$\frac{1 - PD}{PD} = e^{(\beta_0 + \beta_1 X_1 + ... + \beta_n X_n)}, \tag{2.5}$$

$$log(\frac{1 - PD}{PD}) = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n, \tag{2.6}$$

$X_1, ..., X_n$ = Explanatory variables,
β_0 = Regression constant,
$\beta_1, ..., \beta_n$ = Explanatory variables coefficients

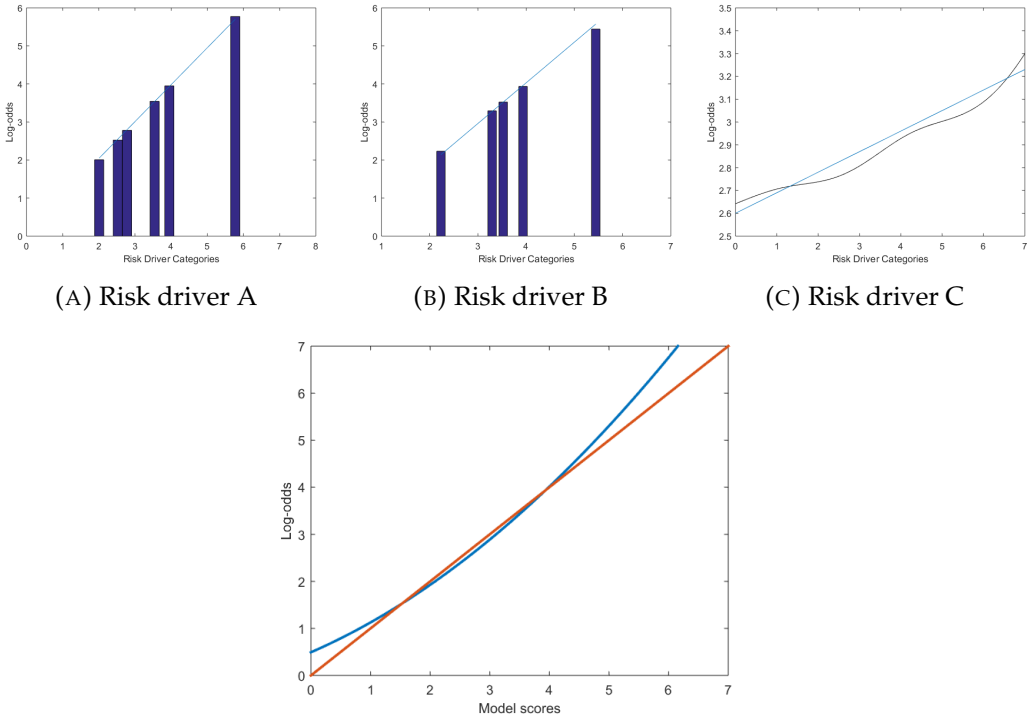
This log-odds linearity assumption is a very restrictive assumption, but can be satisfied for a single risk driver, even if initially it does not hold. For example, a categorical variable with six categories could have scores of {1, 2, 3, 4, 5, 6}. If after calculating the log-odds of each category it turns out an increase in category does not have a linear effect on the log-odds, the scores need to be transformed. The correct scores for each category could therefore be {2.0, 2.5, 2.8, 3.5, 4.0, 5.8}. A correctly transformed variable is shown in Figure 2.5. The figure shows a categorical risk driver with six categories, the x-axis shows the risk score of each category and the y-axis shows the log-odds. Instead of evenly spacing the risk scores, the difference in score reflects the decrease in risk. The linear relationship between score and log-odds is therefore satisfied.

FIGURE 2.5: Transformed explanatory variable



If the risk driver initially has a U-shaped relationship with the log-odds, it is slightly more complicated. It is still possible to transform the risk driver correctly by rearranging the categories or transforming the continuous variable, but the interpretation of the risk drivers effect on the log-odds might be hard to interpret and explain.

The real problem arises when combining multiple risk drivers in the PD model, as shown in Figure 2.6. Even if all the individual risk drivers are linearly related to the log-odds, the PD model predictions can still deviate non-linearly from the observed log-odds.



(A) Risk driver A (B) Risk driver B (C) Risk driver C

(D) PD model

FIGURE 2.6: Effect of combing correctly transformed risk drivers

2.4.2 Significance of regression parameters

The maximum likelihood estimator of logistic regression returns the value of the coefficients of the different risk drivers and the Wald test can be used to test their significance. The Wald statistic test the hypothesis of the coefficient being equal to zero. A rejection of the null hypothesis infers that the corresponding risk driver has a significant effect on the dependent variable (PD).

Wald test: (Wasserman, 2010)

$$H_o : \theta = 0 \text{ versus } H_1 : \theta \neq 0 \tag{2.7}$$

$$\frac{\sqrt{n}(\hat{\theta} - 0)}{\hat{se}} \xrightarrow{d} N(0,1) \tag{2.8}$$

Reject H_0 when $|W| > z_{\alpha/2}$

$$W = \frac{\hat{\theta}}{\hat{se}} \tag{2.9}$$

$\hat{\theta}$ = Estimated coefficient,
θ = True coefficient,
$\hat{se}$ = Estimated standard error
$z_{\alpha/2}$ = Standard deviations from mean at confidence level α of standard normal variable

For Equation 2.8 to hold, the estimator of the maximum likelihood estimation $\hat{\theta}$ needs to be asymptotically normal. For logistic regression this is theoretically the case (Ngunyi, Mwita, and Odhiambo, 2014). Appendix A contains evidence on the asymptotic behaviour of the estimated parameters.

AIC and SBIC statistic

Even though the individual parameters are significant, it is also worth checking if the model has an increase in goodness-of-fit that justifies the use of an extra parameter. Adding extra parameters to your model will always improve your fit, but it will also increase the noise of your estimate. The Akaike Information Criterion (AIC) compares the the quality of two different models based on a trade-off between goodness-of-fit and the number of parameters, as shown in Equation 2.10a.
A second measure, similar to the AIC statistic, is the Schwarz Criterion (SBIC). This measure also includes the number of data points in the estimate, as shown in Equation 2.10b. In both cases the model with the lower score is preferred.

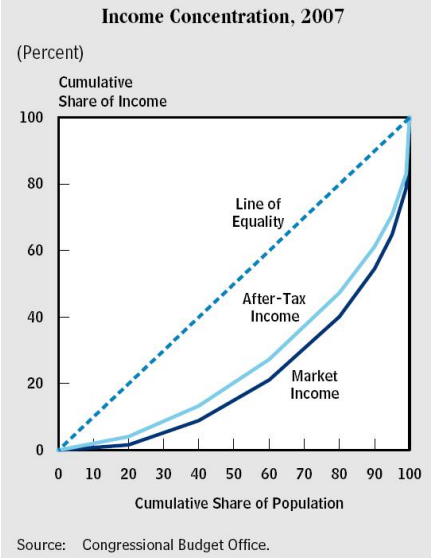
$$AIC = 2k - 2L \tag{2.10a}$$
$$SBIC = -2L + k\log(n) \tag{2.10b}$$

k = number of parameters, L = Log likelihood value, n = number of data points

2.4.3 Gini coefficient

When developing a model for probability of default, we are interested in how well it can discriminate between a client that is going to go into default, and one that is not. This is the ranking performance of a model and this is usually measure by the Gini coefficient. The Gini coefficient is is extensively used as a measure of income equality, where it is defined as the area between the Lorenz curve and the diagonal line representing perfect income equality (Lerman and Yitzhaki, 1984). An example is shown in Figure 2.7, where the Lorenz curve is given by the cumulative income for the cumulative share of the population.

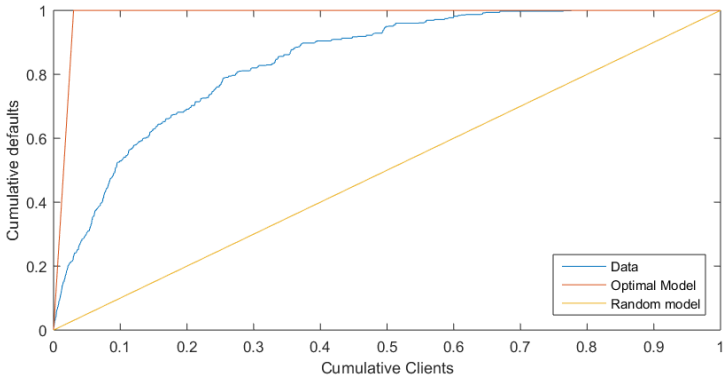
FIGURE 2.7: Income equality (Taylor, 1970)



From an income perspective the Lorenz curve should be as close to the diagonal line as possible, representing perfect income equality. In a PD modelling environment, the Lorenz curve is given by the cumulative number of defaults for the cumulative number of clients, as shown in Figure 2.8. From this perspective, the optimal model discriminates perfectly

between defaults and no-defaults and therefore all the defaults are in the first set of clients, represented by the orange line in Figure 2.8. The yellow line is the worst model possible, since the defaults are randomly spread across the data and there is no discrimination. The blue line is an example of the ranking performance of a PD model. The Gini coefficient is given by the area between the blue and the yellow line, divided by the area between the orange and yellow line.

FIGURE 2.8: Ranking Performance



2.5 Impact on regulatory capital

Regulatory capital is not only impacted by the risk parameters, but also by the Margin of Conservatism (MoC). MoC adds a conservative layer to the model, by capitalising on model risk. The risk model calculates the amount of capital a financial institution needs to hold to be able to cover unexpected losses, but the model itself is also affected by uncertainties. Examples are data issues, limited data availability, changing definitions definitions, lacking procedures, but also performance of the model (EBA, 2016). The amount of MoC is calculated with a scorecard with questions based on several aspects of the model such as uncertainties during model development, model implementation and business strategy. One of the aspects of uncertainties during model development is the calibration accuracy. This aspect is affected by the accuracy of predicted values and realized values within the data. For a PD model this mean that increasing the performance of the transformation of score to PD lowers the margin of conservatism of the model and therefore regulatory capital.

1

Chapter 3

Data Analysis

3.1 Sample Data

To analyse whether the non-linearity issues are present within the models of the Rabobank, we analyse the datasets used for the development and employment of the model. There are several datasets available from which the largest was chosen which contains residential mortgages. The total dataset contains more than 1.5 million records containing loans from 2012 until 2016. The dataset contains four risk drivers, whose values are scores on four different client and loan characteristics.

3.1.1 Causes of non-linearity in the data

Mcdonald, Sturgess, and Smith, 2012 identified two causes for the non-linearity they found in the predicted log-odds. These causes are correlation between binned risk drivers and difference in variance between the distributions of the two classes. The two classes consist of the class where the event happens and the class where the event does not happen.

Difference in distributions between classes

The data can be dividend into two classes, the data from the clients that defaulted (Class 1) and data from the client that did not default (Class 2). Difference in distributions and especially unequal variances, according to Mcdonald, Sturgess, and Smith, 2012, influence the accuracy of PD predictions and leads to violation of the non-linearity assumption of logistic regression.

TABLE 3.1: Mortgage statistics

	Mean	Variance	Skewness	Difference of variances	P-value
Class 1 (Event)	3.15	4.51	-0.70	3.01	<0.0001
Class 2 (No event)	5.73	1.50	-0.32		
Total	5.71	1.58	-0.50		

The variances of two different distributions is compared with the two-sample F-test. The null hypothesis that the two variances are equal is tested against the alternative hypothesis that variances are unequal. Distribution characteristic and test statistics can be found in Table 3.1. The P-value is extremely small, so the hypothesis of equal variances can be rejected with a high significance level.

As shown in Figure 3.1, the distributions of the two classes have more irregularities. The difference in means is due to the nature of score estimation, where the model uses the data to discriminate between the default and non-default classes to be able to predict defaults. The difference in skewness of the two classes is also clearly visible where there is especially a large left skew in the default data. The effect of skewness in score distributions on non-linearity issues is further treated in Chapter 4

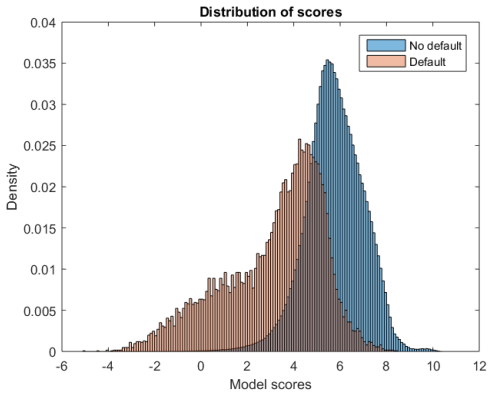
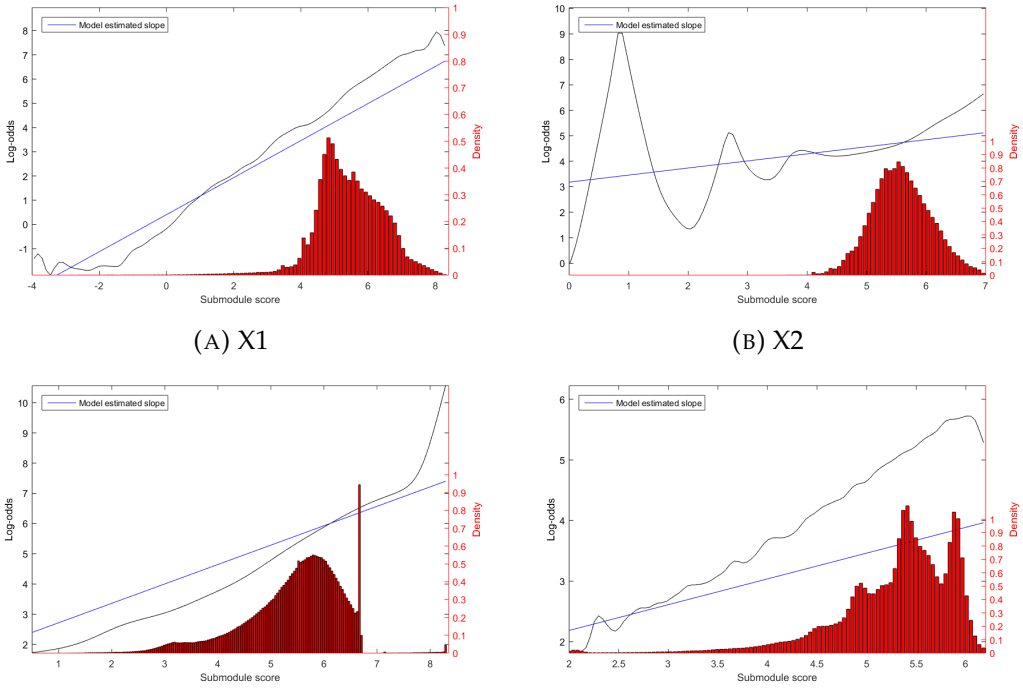


FIGURE 3.1: Class distributions

Individual risk drivers

To find causes of some non-linearity, the individual relationships of the risk drivers and the ODR are investigated. In the mortgage data the risk drivers are model scores of the sub-models. These sub-models are PD models based on different categories of client characteristics and shown in Figure 3.2. The red bars consists of the proportional density, so the significance of different parts of the log-odds relationship can be observed. The added blue slope is the slope of the parameter that is estimated by the model with all risk drivers included. With this graph whether the the scores and log-odds on an individual level have a linear relationship. If the individual risk drivers already have a non-linear relationship with the log-odds, this could potentially explain the non-linearity in the final PD model. The density of the data is also added to the graph to be able to verify the significance of the irregularities in the data. For example, Figure 3.2b shows very irregular log-odds on the left side, but since there is almost no data density the effect will be very limited.



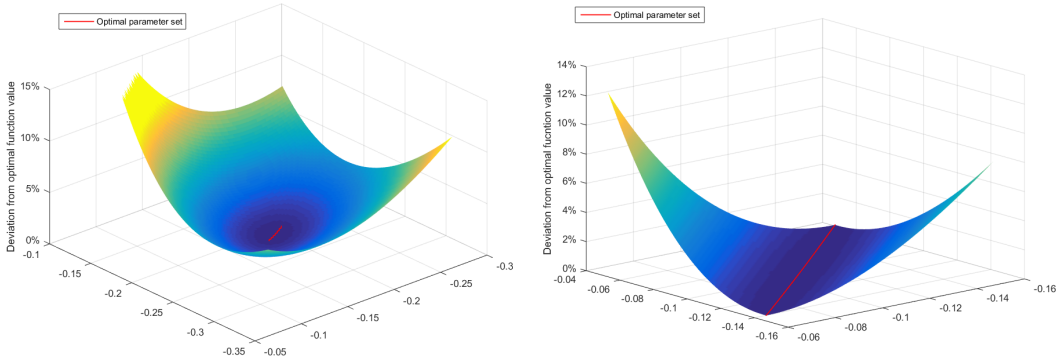
(A) X1 (B) X2

(C) X3 (D) X4

FIGURE 3.2: Log-odds vs. risk drivers

Correlation between risk drivers

The problem with correlation of the risk drivers is that it has an effect on the convergence of the maximum likelihood estimate (MLE). Increasing correlation leads to risk drivers becoming substitutes of one another. Even though very high correlated risk drivers are usually removed in the selection process, the remaining risk drivers are still correlated. This effect is illustrated in Figure 3.3.



(A) Low correlation (0.18)

(B) High correlation (0.58)

FIGURE 3.3: Example of effect of correlation on MLE conversion

Figure 3.3 shows the deviation from the optional function value from the likelihood function. The x-axis and y-axis contain the values of two different parameters and the z-axis contains the deviation from optimal function value. A fast converging MLE means that a change in parameter value from each of the two parameters leads to a significant effect on the optimal function value from the likelihood function, as shown in Figure 3.3a. On the other hand, slow converging means that a change in parameter value leads to an insignificant effect on the optimal function value, as shown in The Figure 3.3b. Correlated variables and therefore slow MLE convergence leads to different combinations of parameters that lead to almost identical likelihood function values. This means that the MLE will be indifferent to different parameter sets and it is a random draw which parameter value are eventually estimated.

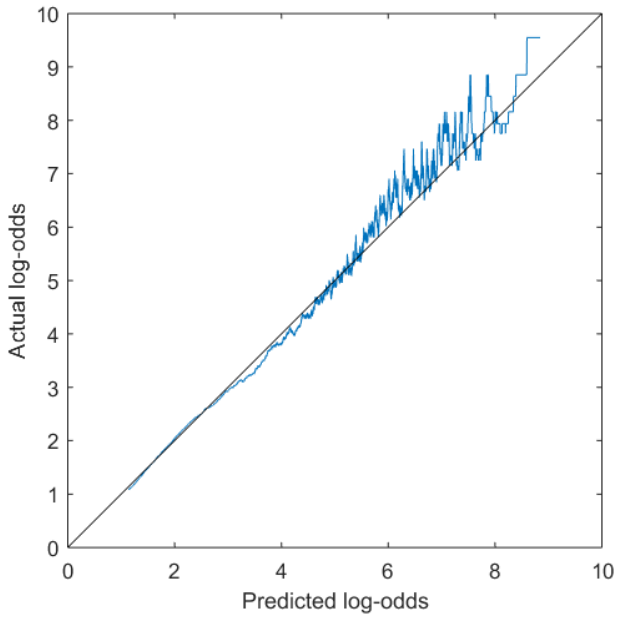
TABLE 3.2: Mortgage Correlation

	X1	X2	X3	X4
X1		**0.2460**	**0.3153**	**0.3774**
X2	0.2460		0.2033	0.1117
X3	0.3153	0.2033		0.2389
X4	**0.3774**	0.1117	0.2389	

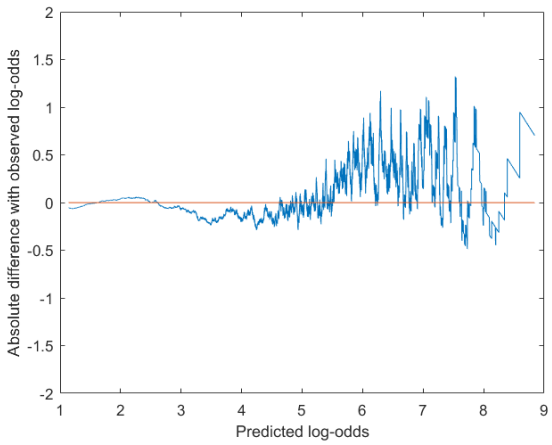
The correlation between the risk drivers from the mortgage data is given in Table 3.2. Mcdonald, Sturgess, and Smith, 2012 used a threshold of 0.3 to remove the highest variable correlation and the mortgage data contains only one variable pair with a slightly higher correlation. This would suggest that the influence of non-linearity within inferred PDs from the mortgage data due to variable correlation is limited, but might still be present.

3.1.2 Actual vs. predicted log-odds

To investigate the linearity in log-odds assumption in the mortgage data a comparison of predicted and actual log-odds is made. Figure 3.4 shows a comparison between the actual log-odds versus the observed log-odds. In Figure 3.4a the x-axis contains the moving average of predicted log-odds from the model and the y-axis contains moving average of the observed log-odds. Higher log-odds represent clients with lower probabilities of default. With a perfect model the predicted average is exactly equal to the observed average. In Figure 3.4a, the noise around the linear line for the higher log-odds is expected, since the defaults in the data are getting rarer. When the moving average drops or adds a default it has a large effect since the amount of defaults in the average is low. What is not expected is the consistent under prediction of the log-odds in the higher region, because this means the PDs of these clients are overestimated. Figure 3.4b is a plot of this difference between the predicted and observed log-odds. In this Figure it becomes extra clear that there is a consistent under prediction for the higher log-odds. The consistent deviation from the linear line could be causes by non-linearity issues from the model.



(A) Actual vs. inferred log-odds



(B) Difference

FIGURE 3.4: Comparison actual vs. inferred log-odds

The average predicted PD will exactly match the average observed PD, by construction of the logistic regression by which they are derived. The problem is that at smaller sub-sections of the data the predicted PD can be significantly different from the observed PD. By dividing the data in 10 buckets, from worst to best clients, we can see the consistent over-prediction of PD for the higher log-odds. This is shown in Table 3.3, where especially buckets 5-9 deviate significantly. The ODR represents the Observed Default Rate (ODR) from the data for the different buckets, indexed due to confidentiality reasons. The other two columns represent the indexed prediction and prediction error.

TABLE 3.3: PD prediction performance

Bucket	ODR index	Predicted	Prediction error
1	100.0	98.3	-1.67%
2	22.2	19.5	-12.00%
3	13.1	12.6	-3.91%
4	8.92	9.18	2.88%
5	6.03	6.90	14.30%
6	3.79	5.14	35.41%
7	2.48	3.71	49.36%
8	1.87	2.55	36.30%
9	1.21	1.62	33.67%
10	0.74	0.80	8.88%

Chapter 4

Results

4.1 Adjustment to PD transformation

We investigated the potential violation of the linearity assumption of logistic regression and suggest an adjustment to the score-to-pd conversion, as shown in Equation 4.1. Two extra parameters are added, one for the squared transformation of the score and an intercept for correction purposes. With the γ_2 parameter we will be able to identify and measure the significance of the non-linear deviation.

$$PD = \frac{1}{1 + \exp^{-\gamma_0 - \gamma_1 score - \gamma_2 score^2}} \tag{4.1}$$

The maximum likelihood estimator will give an estimate of the gamma parameters and we are able to establish their significance. In the case the non-linear effects are absent and the linearity assumption holds, the maximum likelihood would need to approximate the values in Table 4.1

TABLE 4.1: Regression output if linearity holds

Parameter	Value	Significance
Constant (γ_0)	≈0	Not significant
Score (γ_1)	≈1	Significant
Score^2 (γ_2)	≈0	Not significant

A significant γ_2 parameter means that the accuracy of the PD prediction can be improved by making an adjustment to the score. The new score is given by Equation 4.2. This adjusted score can then be linearly transformed to a PD as presented in Equation 1.2.

$$\text{Adjusted score} = \gamma_0 + \gamma_1 score + \gamma_2 score^2 \tag{4.2}$$

As Mcdonald, Sturgess, and Smith, 2012 had pointed out, the nice property of applying a correction to the score is that the ranking performance of the model is not affected and therefore the Gini coefficient is not affected. This means that if client A was more creditworthy than client B in the original model, this will still be the case in the adjusted model. The only thing that is affected is the absolute credit score of client A and B, translating in different PDs.
Applying corrections to the original data can also result in more accurate PD prediction but at the same time lower the ranking performance, which is undesirable in a business environment. For example, in an extreme case, one could drop the data and only use gender as a risk driver which would result in two different credit scores for the male and female clients. Predicted default rates for the two genders would simply be their observed averages and prediction accuracy will be extremely high, but the practical use of this model is nil.

4.1.1 Identification of non-linearity

A γ_2 parameter that is significantly different from zero suggests that the linearity assumption of logistic regression is violated. It also implies that the model that includes score^2 inherently outperforms in goodness-of-fit the linear model that includes only the score.

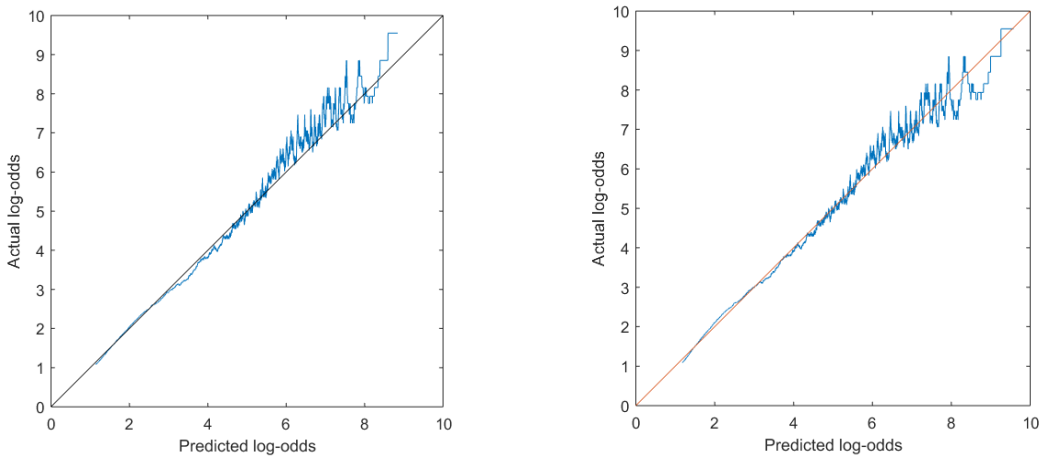
TABLE 4.2: Regression output

Parameter	Estimation	Estimated S.E	z-statistic	p-value
$\hat{\gamma_0}$	0.153	0.025	6.07	<0.0001
$\hat{\gamma_1}$	0.846	0.015	55.5	<0.0001
$\hat{\gamma_2}$	0.025	0.0023	10.5	<0.0001

For our data the results of the regression output are shown in Table 4.2. The Wald test is used to test the hypothesis that $H_0 : \gamma = 0$ versus $H_1 : \gamma \neq 0$. The z-score is the amount of standard deviations from the mean from which the p-value is derived. For $\hat{\gamma_2}$ the p-value is very small, so we can reject H_0 and conclude that γ_2 is significantly different from zero. Therefore the linearity assumption is indeed violated.

4.1.2 Correction non-linearity

The effect of this transformation is visualized in Figure 4.1. Figure 4.1a is the original comparison of observed and predicted log-odds calculated with the original score and Figure 4.1b is the same comparison but with the scores corrected for the non-linear deviation. The linear line represents perfect prediction again. With the corrected scores, the difference between predicted and observed log-odds is smaller.



(A) Original Score (B) Adjusted Score

FIGURE 4.1: Comparison actual vs. inferred log-odds

Effect on performance

The exact increase in performance is given by Table 4.3. For the first nine buckets the performance of the prediction increases and only in the last bucket the prediction error is larger than for the original model. In the last bucket the exponent in the score transformation is causing the log-odds to be overestimated and therefore PD to be underestimated for the most creditworthy clients. In reality this effect will be limited due to a mandatory floor for the PD. The regulators issued a floor of 3 basis points to the PDs so all calculated PDs lower than 3

basis points will be corrected. A PD of 3 basis points corresponds to log-odds of approximately 8.1. These results are all based on in-sample results, so there is a risk of over-fitting on the sample dataset. We reproduced these results for ten out-of-sample estimations. These results are based on ten different sets of 90% training data and 10% validation data. We found that the average prediction error improved significantly in out-of-sample predictions, from 21% to 4%, when the scores are corrected.

TABLE 4.3: Bucket performance

		Linear		Non-linear	
Bucket	ODR index	Predicted	Prediction error	Predicted	Prediction error
1	100.0	98.3	-1.67%	100.3	**0.31%**
2	22.2	19.5	-12.00%	20.2	**-8.88%**
3	13.1	12.6	-3.91%	12.6	**-3.77%**
4	8.92	9.18	2.88%	8.92	**0.01%**
5	6.03	6.90	14.30%	6.50	**7.73%**
6	3.79	5.14	35.41%	4.67	**23.14%**
7	2.48	3.71	49.36%	3.23	**29.95%**
8	1.87	2.55	36.30%	2.10	**12.06%**
9	1.21	1.62	33.67%	1.23	**1.71%**
10	0.74	0.80	**8.88%**	0.54	-27.01%
	Mean error		16.3%		**3.5%**
	Mean absolute error		19.8%		**11.5%**

The difference in performance of the two models is due to the the adjustments in log-odds that lead to certain clients receiving a higher PD and certain clients a lower PD. Because of the construction of logistic regression, the average PD of both models will be exactly the same and equal to the observed defaults in the dataset. The differences in indexed PDs for different types of clients are made visible in Table 4.4.

TABLE 4.4: Indexed differences for PD, LGD & EAD

Bucket	Original PD	Transformed PD	% Difference	LGD	EAD
1	98.3	100.3	2.0%	100.0	100.0
2	19.5	20.2	3.6%	89.6	97.7
3	12.6	12.6	0.0%	79.8	94.7
4	9.18	8.92	-2.8%	70.4	91.9
5	6.90	6.50	-5.8%	61.0	87.6
6	5.14	4.67	-9.1%	53.6	81.6
7	3.71	3.23	-12.9%	48.6	73.1
8	2.55	2.10	-17.6%	46	62.1
9	1.62	1.23	-24.1%	46.7	48.7
10	0.80	0.54	-32.5%	70.8	31.0

The results show that the two buckets of least creditworthy clients receive a higher PD while the rest of the clients receive a lower PD. Especially the PD of the most creditworthy clients is significantly lower, since this is where the non-linear deviation is the greatest. The largest exposures and LGD's are in the buckets which increase in PD, so the expectation is that this will have a negative affect on the regulatory capital. The clients that 'gain' in terms of a lower PD have lower LGD's and less exposure.

AIC and SBIC test results

To test whether the increase in fit of the adjusted model justifies the use of an extra parameter we use the AIC and SBIC test as explained in Subsection 2.4.2. We compare the goodness-of-fit from the original model with the new model and penalize the number of parameters used. The results of the AIC and SBIC tests are shown in Table 4.5. The AIC and SBIC are indexed, because the actual values mean nothing on their own and are simply used to rank the models. The AIC and SBIC statistics provides an indication about the difference between two models, but both can still be useless (Snipes and Taylor, 2014). Adding the squared parameter improves the goodness-of-fit significantly enough to justify the extra parameter of the model.

TABLE 4.5: AIC and SBIC test results

	AIC index	SBIC index
Linear model	1	1
Non-linear model	0.999	0.999

4.2 Alternative methods

The method given in Section 4.1 is a correction that is applied after the clients scores have been calculated. This of course has the convenient property that the ranking performance and the Gini coefficient are unaffected. An alternative to making a correction to the scores is to make an alteration to the score transformation itself. This could improve the PD prediction but could also have an effect on the ranking performance of the model. Another alternative is to make an adjustment or transformation to the original data. Using Principal Component Analysis (PCA) to remove correlation or forcing a normal distribution to the data could also have an effect on the PD prediction.

4.2.1 Adjustment to score transformation

Penalty for minimum risk driver

In Figure 3.4 the comparison was made between the predicted log-odds and the observed log-odds and this figure shows that the largest deviation of predicted and observed PDs exists for the most creditworthy clients. This means that the clients who score very well on all the different risk drivers have a PD that is even lower than the model predicted. An intuitive approach to correct for the underestimation of PD due to averaging a very bad score with good ones, is to add a factor that penalizes the clients more that score bad on one or more of the risk drivers. This can be done by adding the minimum of the risk drivers as an extra risk driver of the model. This would change Equation 1.1 in an alternative score calculation given by Equation 4.3.

$$score = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \beta_{n+1} \min(X_1, X_2, ..., X_n) \tag{4.3}$$

X_n = Transformed financial or qualitative factors
β_0 = The constant term and $\beta_1, ... , \beta_{n+1}$ are the factor weights

The effect of adding the minimum risk driver as an extra risk driver in the model can be analysed by estimating the β parameters from Equation 4.3. The weight of the extra parameter is an indicator of the significance of the minimum risk driver in predicting PD. The results of the parameter estimation is shown in Table 4.6.

TABLE 4.6: Regression output of penalty model

Factor	$\hat{\beta}$	P-value	Original $\hat{\beta}$
Constant	5.1	<0.0001	5.7
X1	0.58	<0.0001	0.76
X2	0.27	<0.0001	0.28
X3	0.52	<0.0001	0.63
X4	0.38	<0.0001	0.43
$\min(X_1, X_2, X_3, X_4)$	0.28	<0.0001	-

The results indicate that the minimum risk driver is highly predictive of a default. This would suggest adding the minimum risk driver as a parameter adds predictive power, but taking a look at discriminatory power of the model, the Gini coefficient is unaffected (Δ = 0.02%). This would suggest that due to the high correlation with the other risk drivers, the extra parameter is acting as a substitute. This correlation of this extra parameter with $\beta_1, \beta_2, \beta_3, \beta_4$ is 0.65, 0.33, 0.69, 0.52. The high correlation also poses a potential problem. We are trying to correct for the non-linear devation in predicted log-odds by adding an extra parameter, but due to the high correlation with the other risk drivers it might be also cause a deviation. We use the identification method from Section 4.1 to test for non-linearity issues in the clients scores. The results of this regression are given in Table 4.7.

The parameter γ_2 is highly significant and only fractionally lower than the 0.025 from the original model. The method of adding a factor that penalizes the clients that score bad on

TABLE 4.7: Regression output

Parameter	Value	P-value
γ_0	0.15	<0.0001
γ_1	0.85	<0.0001
γ_2	0.024	<0.0001

one or more of the risk drivers does not have enough impact on the clients scores in order to correct the deviation between predicted and observed log-odds. The bucketing performance of this model for ten equal buckets even performs worse than the original model, with a mean error of 24% vs 16%.

Squared transformation of every risk driver

An alternative to making a score correction with the squared transformation of the score as shown in Equation 4.1 is to include the squared transformation of every risk driver in the score transformation. This will results in the score transformation given by Equation 4.4.

$$score = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \beta_{n+1} X_1^2 + \beta_{n+2} X_2^2 + ... + \beta_{2n} X_n^2 \tag{4.4}$$

X_n = Transformed financial or qualitative factors
β_0 = The constant term and β_1, ... , β_{n+1} are the factor weights

The estimated parameters are shown in Table 4.8. Interesting observation is that X2 and the squared transformation of X2 have quite a high p-value, where in the original model X2 was highly significant. The rest of the risk drivers and squared transformations are highly significant.

TABLE 4.8: Regression output of squared model

Factor	β	P-value
Constant	1.76	<0.43
X_1	-1.14	<0.0001
X_2	1.11	0.10
X_3	0.97	<0.0001
X_4	0.31	<0.001
X_1^2	0.025	<0.0001
X_2^2	-0.10	0.05
X_3^2	-0.16	<0.0001
X_4^2	-0.087	<0.0001

The improved accuracy of the model is visualized in Figure 4.2. The noise around the linear line remains, but the average predicted log-odds is very close to the observed log-odds. The average prediction error for ten buckets is 5.41% when the original model had an average prediction error of 16.3%.

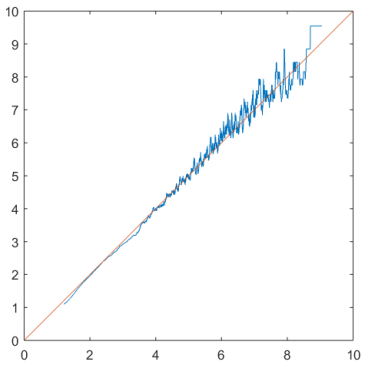


FIGURE 4.2: Actual vs. inferred log-odds squared model

The disadvantage of this model is that it will affect the ranking performance of the model and therefore the output will be harder to explain to the business side. For example, client A might have a lower loan to income than client B, but this model might conclude that client A is more creditworthy than client B based on the squared transformation of his loan to value.

4.2.2 Data transformation

Principal component analysis

The correlation between risk drivers is identified to be a large factor in the violation of the linearity assumption of logistic regression. Principal Component Analysis (PCA) is able to reduce the dimensionality of the data by transforming the variables to uncorrelated Principal Componants (PC). Jolliffe, 2010 defined in the following way:

> *The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables.*

Removing correlation from the data trough PCA therefore might be able to reduce the non-linear effect and improve PD prediction. To perform the PCA the steps from Smith, 2002 have been used. First step is to subtract the variable mean from each data point, so all the variables have a mean of zero. Second step involves calculating the covariance matrix, including the eigenvectors and eigenvalues. At this stage you can drop the 'lower' absolute eigenvalues to decrease the dimensionality of the data. Since our data contains five dimensions, four variables and an intercept, we are able to drop up to three variables. After lowering the dimension we form a feature vector from the remaining eigenvectors.

$$\text{Feature Vector} = (eig_1\ eig_2\ eig_3\ ...\ eig_n); \tag{4.5}$$

$$\text{Final Data} = \text{Feature Vector' x Adjusted Data} \tag{4.6}$$

The last step of the data transformation for the PCA is multiplying the data by the transposed Feature Vector. The final data can then be used to run the logistic regression and calculate the clients scores and predict their PD. The scores and Equation 4.1 are then used to identify the significance and size of the non-linear effects, which are shown in Table 4.9.

Neither dropping one, two or three principal components has a significant effect on the non-linear effects, since all γ_2 parameters are highly significant and their values are all close to the original value of 0.025.

TABLE 4.9: PCA non-linearity identification

Principal components dropped	γ_2 value	γ_2 P-value
1	0.0252	<0.0001
2	0.0242	<0.0001
3	0.0280	<0.0001

4.3 Impact on regulatory capital calculation

From a business perspective it is very interesting to see what effect this alternative method of PD transformation has on the regulatory capital that the financial institution needs to hold. The alternative method will directly impact the capital trough a change in calculated PD and indirectly through a change in MoC (Section 2.5).
Regulatory capital consists of the capital a financial institutions needs to hold for unexpected losses. The institution holds provisions for the expected losses and capital for the unexpected. These unexpected losses are calculated through the Vasicek formula based on Merton's model. The intuition of this formula is given by Equation 4.7. $PD_{downturn}$ is the worst case PD for a certain confidence bound, 99.9% in case of regulatory capital. The actual Vasicek formula is given by Equation 4.8 (BIS, 2005).

$$Capital = (PD_{downturn} - PD) * LGD * EAD \tag{4.7}$$

$$Capital = \left(\mathcal{N} \left(\frac{\mathcal{N}^{-1}(PD) + \sqrt{R} * \mathcal{N}^{-1}(0.999)}{\sqrt{1-R}} \right) - PD \right) * LGD * EAD \tag{4.8}$$

$\mathcal{N}$ = Cumulative distribution function for a standard normal random variable,
$\mathcal{N}^{-1}$ = Inverse cumulative distribution function for a standard normal random variable,
R = Correlation factor based on the type of exposure

By construction of logistic regression, the mean PD of both models will be exactly the same and equal to the observed default frequency of the dataset. The transformed model is correcting for an overestimation of the very creditworthy clients, so these will be assigned a lower PD and less creditworthy will be assigned a higher PD. The difference in capital therefore will be dependent on the exposure and LGD of both types of clients. These differences have already been shown in Table 4.4.
For our dataset, the transformed model results in an increase in regulatory capital of 0.26% due to the direct effect on the PD. The explanation resides in the difference in LGD and EAD between the most creditworthy and the least creditworthy clients. Half of the dataset containing the least creditworthy clients has a 61% higher LGD and a 58% higher EAD.
Even though the total regulatory capital is slightly higher, the capital per client now better represents the risk because of the increase in PD accuracy. The 0.26% also does not include the expected decrease in capital due to a lower MoC, but the exact MoC cannot be calculated since it is a qualitative add-on.

Chapter 5

Missing values and non-linearity

5.1 Approaches to missing value analysis

Missing values are a large problem in the financial industry and in particular PD modelling. There are many underlying causes of these missing values. Examples of these underlying causes include fields that were not captured, discontinued fields, unavailability of the characteristic, intentionally not filled out by applicant or outliers that were removed. Statistical techniques such as random forests or decision trees are immune to missing values, but logistic regression needs a complete dataset (Siddiqi, 2006).

Gelman and Hill, 2006 identified a framework containing four different types of missing values and how to adopt them within the regression framework.

- Missingness completely at random:

  The probability of a missing value is the same for every client.

- Missingness at random:

  The probability of a missing value is not equal for every client, but the information that is affecting this probability is known in the dataset. For example, males have a higher probability of missing values, but the gender is included in the dataset.

- Missingness that depends on unobserved predictors:

  The probability of a missing value is not equal for every client and it is now known what is affecting this probability.

- Missingness that depends on the missing value itself:

  The probability of a missing value is not equal for every client and the variable itself is affecting this probability (e.g. only high-earning persons are unwilling to provide income data).

5.1.1 Missing value bias

Logistic regression is the most widely used method for calculating the PD and because it needs a complete dataset to function, the missing values have to be dealt with. The most conventional choices of coping with missing values are the following three methods (McKnight, 2007):

- Mean imputation, replacing the missing value with the mean of the remaining values.
- Median imputation, replacing the missing value with the median of the remaining values.
- Discard row, removing the records which contain missing values for at least one of the variables.

The problem with these methods is that they could have an effect on the estimated parameters and therefore could inflict a bias in the model.

Parameter estimation

We can take our original mortgage data and randomly impute missing data in the dataset. This means we can compare the original parameters of the model with the parameters estimated for the model with missing values and determine if a bias might be present. Table 5.1 show the results of this analysis. For different percentages of missing values we calculate the estimated $\hat{\beta}$ parameters from the regression and find their average deviation from the basis scenario. The value of 1 would mean that there is no bias in the model, since the estimated parameters are exactly equal to the model without missing values. Values higher or lower than 1 means the β parameters are overestimated or underestimated.

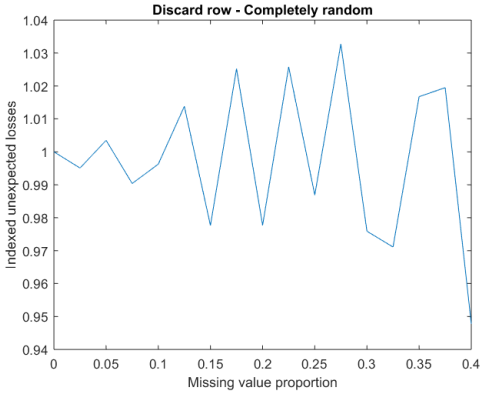
TABLE 5.1: Effect of method on average indexed parameter

Missing value percentage	Mean imputation	Median imputation	Discarded record
5%	1.08	1.05	0.97
10%	1.11	1.15	1.00
20%	1.14	1.21	0.98
30%	0.99	1.43	1.04
40%	0.79	1.47	1.11

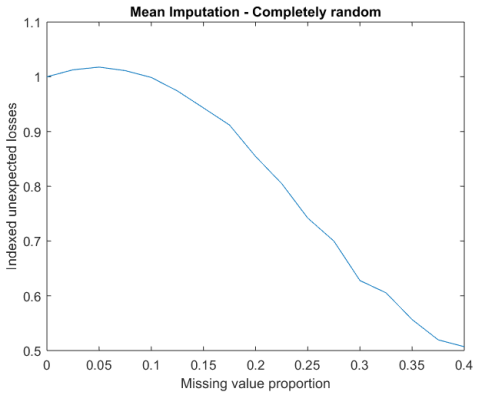
The table presents evidence that, under the assumption that the missing data in a dataset is completely random, imputing the mean or median will introduce a bias in your model. Even at only 5% missing values this deviation in parameter estimation can be quite large. Discarding the record seems to be the best method to avoid a bias in your model. The disadvantage is that you will need to throw away a lot of data, which might affect the ranking performance of the model.

Unexpected losses

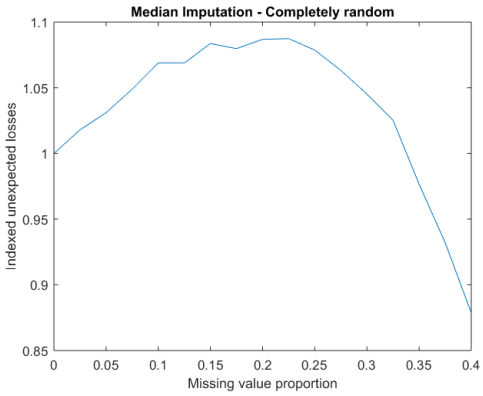
Another interesting aspect to look at is the effect of missing data on the calculated capital requirements. We know that imputing mean or median to cope with missing values has an effect on the estimated parameters, therefore we also expect it to have an effect on the unexpected losses. The unexpected losses are part of the loss distribution for which a financial institution needs to hold capital. Figure 5.1 shows the effect of each method on the indexed unexpected losses. The y-axis show the indexed unexpected losses and the x-axis the proportion of missing values.



(A) Discard record



(B) Mean imputation



(C) Median imputation

FIGURE 5.1: Effect of method on indexed unexpected losses

The method of discarding the record is, like expected, the most unbiased. Figure 5.1b shows that for high percentages of missing values, the unexpected losses will be underestimated by a large portion when using mean imputation. The mean imputation method from Figure 5.1c increases the unexpected losses for the lower missing percentages, but ends op with an overoptimistic estimate for the higher percentages. The effect is present because the imputation of the mean will decrease the variance of the distributions. Potential outlier will be replaced with the mean of the distribution and this will lead to the underestimation of the risk. The difference of effects between mean and median imputation will be mostly based on the skewness of the distribution. Skewed distribution will have a different mean and median and this difference is either positive or negative depending on a left or right skewed distribution. When faced with the choice of imputing the mean or median in the case of missing values it will be wise to look at the skewness of the variable distributions.

TABLE 5.2: Imputing mean effect on non-linearity

Missing value percentage	Non-linearity parameter	P-value
5.0%	0.024	<0.0001
10.0%	0.060	<0.0000
15.0%	0.060	<0.0000
20.0%	0.020	0.0003
22.5%	0.013	0.0157
25.0%	0.000	0.47
27.5%	-0.016	0.0027
30.0%	-0.026	<0.0001
35.0%	-0.062	<0.0001

Non-linearity and missing values

By randomly inserting missing values in our dataset and replacing them by the mean or median, the distributions of the default and non-default data are altered. From Section 2.1 we know that the difference in distribution between the two classes is a known driver of non-linearity in the data. Table 5.2 shows the results the effect of missing values and imputing the mean on the γ_2 parameter from Equation 4.1 which identifies the significance and magnitude of the non-linearity in the data.

The results show that adjusting the data by imputing mean has a large effect on the non-linear deviation of predicted log-odds. A missing values percentage of 25% even produces a distribution for which the linearity assumption of logistic regression holds. The results illustrate the effect of the distributions on the accuracy of the PD prediction, but actually randomly imputing the mean is never a solution to the non-linear deviation. A lot of information in the data will be lost and the discriminatory power of the model will be negatively affected. These results should be taken into account when developing a model. In the process of risk driver selection, choices have to be made about which risk drivers end up in the final model. When faced with two highly correlated risk drivers with equal prediction power, one very skewed and irregular and one almost normally distributed, these results suggest choosing the latter.

Chapter 6

Conclusion

In this thesis we tried to answer the question whether a non-linear deviation between predicted and observed log-odds is present and how this deviation can be corrected. Logistic regression has the linearity-in-log-odds assumption which is very restrictive. The results indicate that for datasets used in PD prediction, the log-odds are not linearly related to the PD and that enforcing a linear relationship results in suboptimal PD prediction accuracy.

Unfortunately the problem cannot be avoided by transforming each individual risk driver, because the combined model is still subject to the assumption being violated. This is due to correlation between the different risk drivers and differences between the distribution of default and non-default data. These differences lead to a non-linear relationship of the log-odds and model scores and therefore to the overestimation of the most creditworthy clients.

There are several methods to identify whether a PD model is affected by this non-linearity. Mcdonald, Sturgess, and Smith, 2012 identified two causes as the largest drivers of the linearity assumption being violated, high correlation between the risk drivers and large variance differences between the default and non-default data distributions. An initial indication can therefore be provided by calculating the correlation and variance difference.

The method for identification we suggest is to add a parameter to the transformation of client score to PD, as shown in Equation 4.1. The significance of the parameter identifies whether the non-linearity is significant and the size of the parameter represents the magnitude. The advantage is that the estimated parameters of this step can be used to adjust the original log-odds and therefore correct the deviation. The other significant advantage is that the ranking performance based on the creditworthiness of the clients remains the same and therefore the model does not lose discriminatory power. The correction is applied before the transformation to probability, so only the absolute value of the PD is affected to improve the prediction accuracy.

The accuracy of these corrected log-odds was much higher on subsets of the data, even though the total average PD is still equal. The average error for ten subsets of the data improved from 16% to 4% by correcting the log-odds. This means that the calculated PDs for all clients is more representative of the corresponding risks, which is convenient for efficient capital allocation and RAROC measures.

The regulatory capital the financial institution needs to hold is affected by the value and the accuracy of the PD prediction. Due to the accuracy improving, the Margin of Conservatism that the financial institution needs to hold as extra capital is lower. The effect of the value change of PD is less straightforward. Since the correction that is applied is parabolic, there will be clients who receive a higher PD and clients who receive a lower PD. Dependent on the LGD and exposure of these clients the regulatory capital is adjusted up- or downwards.

Returning to the main research question, our results suggest that the accuracy of probability of default calculations for loan portfolios is significantly impacted by the restrictions of logistic regression. Disregarding the fact that linearly fitting the log-odds can have a negative impact on the accuracy of the PD prediction, will lead to suboptimal capital allocation. Therefore for all financial institutions that use logistic regression it is beneficial to add this extra step to their PD transformation and identify whether the linearity assumption holds and make a correction to the log-odds if the assumption is violated.

Chapter 7

Discussion

7.1 Limitations

The outcomes of PD models have to be interpretable, since PD models have to be approved by the European Central Bank. The problem is that this is a large limitation in a prediction environment. Machine learning provides very useful techniques for default prediction which significantly outperforms any form of logistic regression, but 'black box' style modelling techniques are not likely to be approved. Sirignano, Sadhwani, and Giesecke, 2016 investigated whether the use of neural networks improved the performance of predicting mortgage delinquency. The neural networks had an 8% improvement in fit to the empirical distribution over logistic regression. They let both models choose their own 100.000 loan portfolio and the neural network had a portfolio with 20% fewer defaults and 50% fewer number of mortgages with prepayments (also a risk for mortgage providers). The problem is explaining the outcomes of a neural network model. This is the reason for making an alteration to logistic regression and not using another non-linear technique for predicting PDs.

Our research has focused on a portfolio dataset containing a fairly large amount of 'safe' clients, so a low average PD. This means the amount of historical defaults is very low and the impact of overestimation of creditworthy clients has an extra impact. Extending the research to other type of portfolios, for example credit-card loans, would have been an interesting addition. These types of portfolios have more historical defaults and a more even distribution of predicted PDs.

When comparing the performance of the original PD model and the model with our suggested correction, the predicted PD is measured against the observed PD. To avoid overfitting on the training dataset, we made a comparison based on validating the performance on an out-of-sample dataset. Since the data suggest that currently the PDs are overpredicted for the clients with a high credit score, our research suggests an improvement on the method to predict PD. An interesting addition to this research would be to back-test whether the use of the original logistic regression actually led to the overestimation of PD for the creditworthy clients.

7.2 Suggestions for further research

Our research is primarily focused on the prediction of PD using logistic regression, but the underlying problem that is identified could hold for all applications of logistic regression. Logistic regression is used for all kinds of binary classifications in other fields such as social and medical studies. We focused on the statistical properties of logistic regression and the problems with the restricting linear-in-log-odds assumption. A suggestion for further research would therefore be to apply the techniques from this research to datasets from other fields.

Another suggestion for further research is the underlying causes of the linearity assumption being violated. The paper from Mcdonald, Sturgess, and Smith, 2012 indicates correlation and difference in variances as causes, because in this paper the issues with the linearity assumption disappeared when risk drivers with a correlation higher than 0.3 are removed. For the dataset in this research, this method was not sufficient, which indicates that the cause

is not just correlation. As described in Section 5.1.1, changing the distributions of the default and non-default data also has a large impact on the non-linearity issues, but isolating a single cause and studying the effect is difficult. For example, generating data to investigate the effect of correlation or variance is not possible. The assumption that the actual portfolio data are similar to the normally generated data will not hold.

Appendix A

Properties of logistic regression parameters

A.1 Asymptotic Normality

An estimator is asymptotically normal if Equation A.1 holds. So as n tends to infinity, the difference between the estimate $\hat{\beta}$ and the true parameter β will converge in probability to a normal distribution with mean 0 and variance equal to the estimated covariance matrix (Naima and Mamunur, 2012).

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, I(\beta)^{-}1) \tag{A.1}$$

Asymptotic normality does not only state that the estimated parameter converges to the true parameter as n grows, but also that it converges at a rate that is fast enough, $\frac{1}{\sqrt{n}}$ (Panchenko, 2006). To verify that our n is large enough, we can test what our sample behaviour is by bootstrapping 2000 different samples from our dataset and estimate the β parameters for each sample. Figures A.2 & A.1 show the Quantile-Quantile (Q-Q) plots of the residuals from the β parameter estimation. The plots verify that the asymptotic normality of the estimated parameters hold. This is important for the usability of the Wald estimator in Section 2.4.2, since it allows us to efficiently estimate the significance of the estimated parameters.

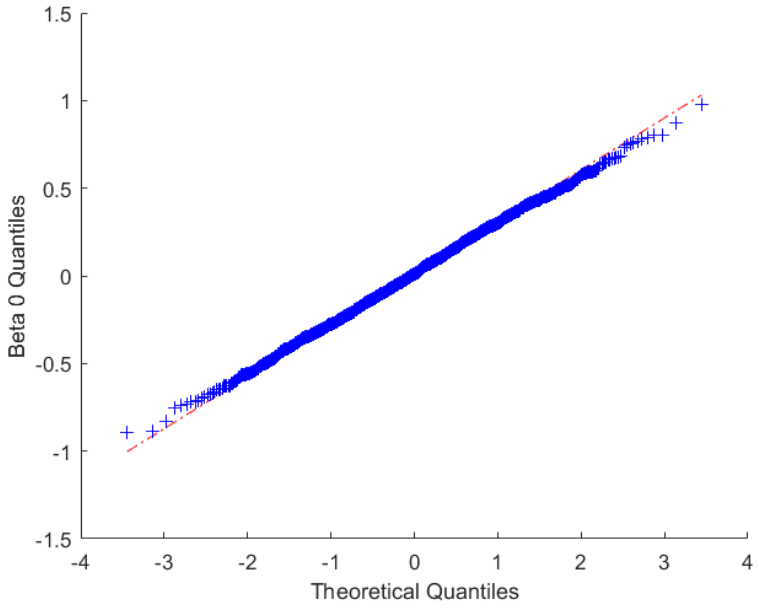
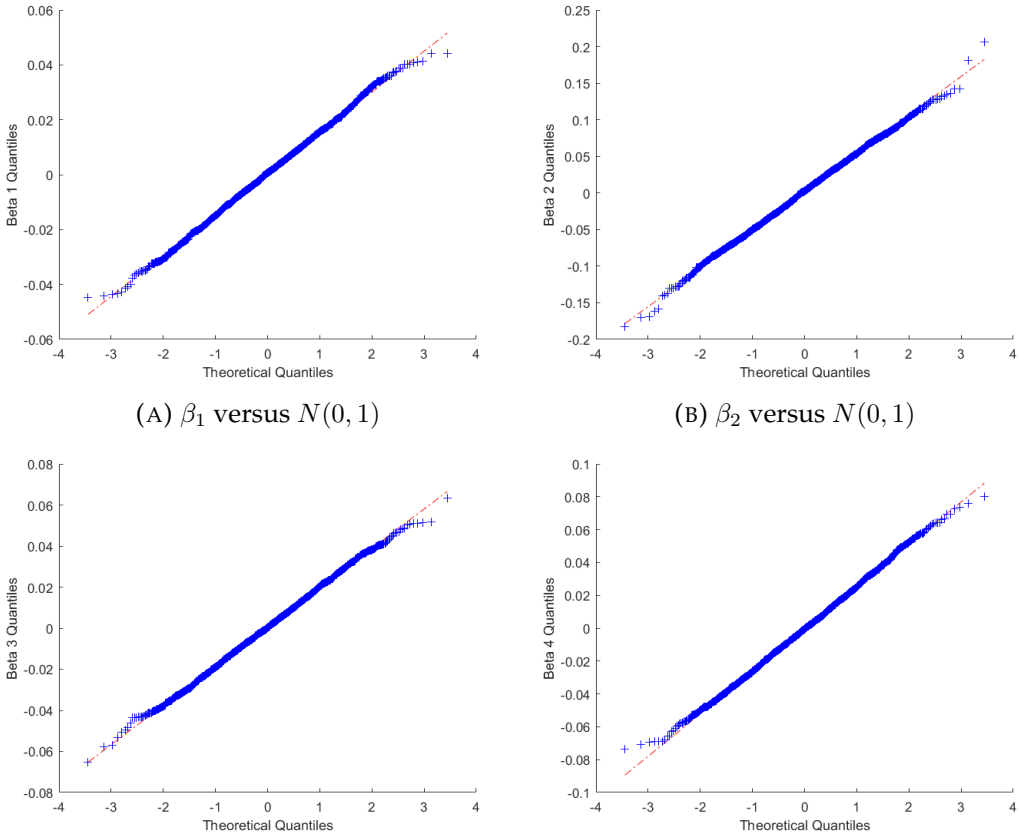


FIGURE A.1: Q-Q plot of the intercept versus $N(0, 1)$



(A) β_1 versus $N(0,1)$

(B) β_2 versus $N(0,1)$

(C) β_3 versus $N(0,1)$

(D) β_4 versus $N(0,1)$

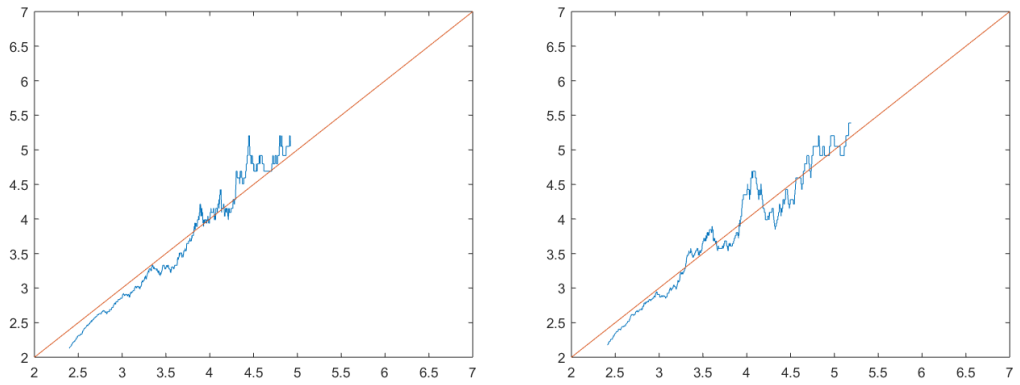
FIGURE A.2: Q-Q plot of the β parameter residuals versus standard normal distribution

Appendix B

Normally generated data

B.1 Non-linearity in normally generated data

To investigate the effect of correlation within risk drivers used for PD modelling we use a randomly generated dataset. The first step is to generate the data required and analyse the accuracy of the PD prediction. The second step is to add correlated variables to the dataset and calculate whether there is a significant difference. Figure B.1 shows that there is not a significant difference. The process was repeated as to ensure the validity of the result.



(A) No correlation (B) Correlation

FIGURE B.1: Regression including and excluding correlated variables

Correlation between risk drivers is identified to be a large driver of deviation from predicted log-odds, but when data is generated normally the effect is not present. This presents further evidence for the importance of the distributions of the data, especially the variance and skewness of the default and non-default data, on the violation of linearity assumption of logistic regression.

Bibliography

Acuña, Edgar and Caroline Rodriguez (2004). "The Treatment of Missing Values and its Effect on Classifier Accuracy". In: *Classification, Clustering, and Data Mining Applications*, 639–647.

Aguilera, Ana M., Manuel Escabias, and Mariano J. Valderrama (2006). "Using principal components for estimating logistic regression with high-dimensional multicollinear data". In: *Computational Statistics and Data Analysis* 50.8, 1905–1924.

BIS (2005). "An Explanatory Note on the Basel II IRB Risk Weight Functions". In: *Bank for International Settlements*.

EBA (2016). *Guidelines on PD estimation, LGD estimation and treatment of defaulted assets*.

Gelman, Andrew and Jennifer Hill (2006). "Missing-data imputation". In: *Data Analysis Using Regression and Multilevel/Hierarchical Models*, 529–544.

Jagric, Vita, Davorin Kracun, and Timotej Jagric (2011). *Does Non-linearity Matter in Retail Credit Risk Modeling*.

Jolliffe, I. T. (2010). *Principal component analysis*. Springer.

Lennox, Clive (1999). "Identifying failing companies: a re-evaluation of the logit, probit and DA approaches". In: *Journal of Economics and Business* 51, 347–364.

Lerman, Robert I. and Shlomo Yitzhaki (1984). "A note on the calculation and interpretation of the Gini index". In: *Economics Letters* 15, 363–368.

Little, Roderick J. A. and Donald B. Rubin (2002). *Statistical Analysis with Missing Data*.

Mcdonald, Ross A., Matthew Sturgess, and Keith Smith (2012). "Non-linearity of scorecard log-odds". In: *International Journal of Forecasting* 28.1, 239–247.

McKnight, Patrick E. (2007). *Missing data: a gentle introduction*. Guilford Press.

Müller, Marlene and Wolfgang Härdle (2003). "Exploring Credit Data". In: *Credit Risk Contributions to Economics*, 157–173.

Naima, Shifa and Rashid Mamunur (2012). *Monte Carlo Evaluation of Consistency and Normality of Dichotomous Logistic and Multinomial Logistic Regression Models*.

Ngunyi, Anthony, Peter Nyamuhanga Mwita, and Romanus O. Odhiambo (2014). "On the Estimation and Properties of Logistic Regression Parameters". In: *IOSR Journal of Mathematics* 10.4, 57–68.

Panchenko, Dmitry (2006). *Properties of MLE: consistency,asymptotic normality, Fisher information.*

Rabobank Group, Rabobank (2014). -.

Siddiqi, Naeem (2006). *Credit risk scorecards developing and implementing intelligent credit scoring*. Wiley.

Sirignano, Justin, Apaar Sadhwani, and Kay Giesecke (2016). *Deep Learning for Mortgage Risk*.

Smith, Lindsay I (2002). *A tutorial on Principal Components Analysis*.

Snipes, Michael and D. Christopher Taylor (2014). "Model selection and Akaike Information Criteria, An example from wine ratings and prices". In: *Wine Economics and Policy* 3, 3–9.

Taylor, Timothy (1970). "What's a Gini Coefficient?" In: *Conversable Economist*.

Wasserman, Larry (2010). *All of statistics: a concise course in statistical inference*. Springer.