

LUND UNIVERSITY  
DEPARTMENT OF ECONOMICS

MASTER THESIS

---

MITIGATING DEFAULT RISK IN THE  
CONSUMER CREDIT MARKET

---

*Author:*  
JACOB HALÉN DAHLSTRÖM

*Supervisor:*  
BIRGER NILSSON, DOC.

August 24, 2016

## Abstract

This paper aims to evaluate recent policy updates in a credit scoring model and determine if the new model is efficient, as well as further investigate other potential risk factors. In order to evaluate the policy changes, the proprietary dataset is first categorized and estimated by a logistical regression model and secondly the dataset is transformed according to new policies and then simulated in a second regression. The choice of variables is further tested to ensure robust result of the identified risk factors and best fitting of the model. The discoveries points towards efficient implemented policy changes to the scoring model, and the identification of other potential risk factors which leads to a managerial suggestion at the end of this paper.

*Keywords:* Credit scoring models, Probability of default, Logistical regression, Maximum likelihood, Information value, Consumer loans

## Acknowledgements

I would like to thank my supervisor Birger Nilsson (Lund University), the bank on which this investigation is based on, and finally my family and friends who have supported me during the past years at Lund University.

August 17, 2016

Jacob Halén Dahlström

# CONTENTS

- INTRODUCTION ..... 7
  - 1.1 BACKGROUND ..... 7
  - 1.2 PURPOSE ..... 8
  - 1.3 DISPOSITION ..... 9
- BACKGROUND ..... 10
  - 2.1 HISTORY..... 10
  - 2.2 REGULATORY IMPLICATIONS: BASEL II & III ..... 11
- THEORETICAL FRAMEWORK ..... 12
  - 3.1 SUMMARY ..... 12
  - 3.2 CREDIT SCORING MODELS..... 13
    - 3.2.1 Discriminant analysis ..... 14
    - 3.2.2 Regression..... 14
    - 3.2.3 Logistic regression ..... 15
    - 3.2.4 Mathematical programming methods..... 15
    - 3.2.5 Recursive partitioning ..... 15
    - 3.2.6 Expert systems ..... 16
    - 3.2.7 Neural networks..... 16
    - 3.2.8 Smoothing nonparametric methods ..... 16
    - 3.2.9 Time varying models..... 16
- METHODOLOGY ..... 17
  - 4.1 EMPLOYED REGRESSION AND FRAMEWORK..... 17
  - 4.2 RESEARCH PURPOSE ..... 17
  - 4.3 DATA SOURCE ..... 18
  - 4.4 MODEL SPECIFICATION ..... 18
    - 4.4.1 The logistic regression model..... 18
    - 4.4.2 Estimated model ..... 20
  - 4.5 ECONOMETRIC ASSUMPTIONS..... 21
    - 4.5.1 Omitted variables ..... 21
    - 4.5.2 Covariate measurement error ..... 22
    - 4.5.3 Response misclassification..... 22
  - 4.6 VARIABLES ..... 23
    - 4.6.1 Descriptive statistics..... 24
    - 4.6.2 Dependent variable..... 24
    - 4.6.3 Explanatory variables..... 25
- EMPIRICAL FINDINGS ..... 27
  - 5.1 REGRESSION MODELS ..... 27
    - 5.1.1 Demographic ..... 28
    - 5.1.2 Financial..... 29
    - 5.1.3 Behavioral..... 31
  - 5.2 IMPLEMENTED POLICY CHANGES – MODEL 2 ..... 32
    - 5.2.1 Demographic ..... 34
    - 5.2.2 Financial..... 35
    - 5.2.3 Behavioral..... 36

5.3 INFORMATION VALUE – EXPLANATORY VARIABLE REDUCTION.....	37
5.3.1 Akaike Information Criterion (AIC).....	40
5.3.2 Pearson’s chi square test.....	41
5.3.3 Log-likelihood ratio (LR) test.....	42
5.4 INTERPRETATION OF RESULTS .....	42
5.4.1 Model evaluation .....	42
5.5 DISCUSSION .....	43
5.5.1 Pre policy changes – Model 1.....	43
5.5.2 Post policy changes – Model 2.....	44
5.5.3 Policy performance.....	45
5.5.4 Remaining risk factors .....	46
<b>CONCLUSION.....</b>	<b>47</b>
6.1 CONCLUDING REMARKS .....	47
6.2 LIMITATIONS.....	47
6.2.1 Sample selection bias.....	47
6.2.2 Model simulation .....	47
6.2.3 Sample size .....	48
6.3 MANAGERIAL SUGGESTIONS.....	48
<b>REFERENCES .....</b>	<b>50</b>
<b>APPENDIX.....</b>	<b>53</b>
8.1 VARIABLES .....	53
8.2 DESCRIPTIVE STATISTICS.....	54
8.3 REGRESSIONS.....	55
8.4 INFORMATION VALUE.....	60
8.5 SELECTION TESTS.....	61

## TABLES AND FIGURES

TABLE 1: VARIABLE DEFINITION.....	23
TABLE 2: DESCRIPTIVE STATISTICS .....	24
TABLE 3: MODEL 1 – DEMOGRAPHICS .....	28
TABLE 4: MODEL 1 – FINANCIALS .....	29
TABLE 5: MODEL 1 – BEHAVIORALS .....	31
TABLE 6: POLICY DEVELOPMENT .....	33
TABLE 7: MODEL 2 – DEMOGRAPHICS .....	34
TABLE 8: MODEL 2 – FINANCIALS .....	35
TABLE 9: MODEL 2 – BEHAVIORALS .....	36
TABLE 10: INFORMATION VALUES .....	37
TABLE 11: SELECTED VARIABLE REGRESSIONS .....	39
TABLE 12: AIC & BIC VALUES.....	41
TABLE 13: CHI SQUARE TEST.....	41
TABLE 14: LR TEST.....	42
TABLE 15: TEST EVALUATION .....	42
TABLE 16: POLICY PERFORMANCE .....	45
FIGURE 1: LOGIT FUNCTION GRAPH.....	20
FIGURE 2: DEFAULT DISTRIBUTION .....	33
EQUATION 1: DEMOGRAPHIC VARIABLES .....	20
EQUATION 2: FINANCIAL VARIABLES.....	21
EQUATION 3: BEHAVIORAL VARIABLES .....	21

# Chapter 1

## Introduction

### 1.1 Background

Somewhere near 4 million of the Swedish population are indebted by loans and credits besides mortgage loans, of these are individuals with the lowest incomes indebted up to 296 % of their annual income (Winstrand and Ölcer, 2014, p.4). Credit institutions and banks issue credits and loans in order to allow consumption smoothing for individuals, this in turn plays an important role for producers and also enable sustainable economic growth for a society. Over the years' individuals and households have become more prone to use consumer credits and other financing alternatives for everyday consumption. This liberalization of credits has allowed for new actors on the credit market, and while the facilitation of credits increases the purchasing power of individuals and households and there by sales, it also includes a certain risk taking. With an increase in volume comes increased risk, the risk of customers not paying back on time or at worst not at all, this is called defaulting. This increased demand has created the need for efficient and accurate credit scoring in order to keep the risk exposure minimized and in accordance with Basel III rules, by trying to predict customer defaults.

The term 'credit' is referred to as an amount that is borrowed by the consumer from a bank or financial institution which must be repaid over a period of time with interest, often a predetermined number of months or years. Credit scoring refers to the method of classifying the loan applicants into different categories dependent of their probability of default or non-

default (Hand and Henley, 1997, p.531). This method is usually based on historical data of previous loans gone bad and the demographic, financial and behavioral patterns underlying the default. These data are usually used in a logistic regression analysis that links the probability of default given the historical data, which in turn can be applied on new application data. This is used to classify applicants into accept or reject categories (Cramer, 2004, p. 365), although in this paper I will focus on an existing risk-scoring model and evaluation of its efficiency.

### 1.2 Purpose

This paper investigates the efficiency of a credit scoring model in a Swedish niche-bank, which mainly issues consumer credits within Sweden. With recent changes in their credit model and structural changes within the bank they wish to evaluate their risk model in accordance with both previous and current credit policy. Since the aim is to evaluate the fallout of implemented changes, I perform an analysis on approved application data in accordance with both old and new credit policy and regress both demographic, financial and behavioral variables to a number of defaults over the time period 2013-2015.

The data consist of 7577 approved loans in full sample, there among are 183 defaulted loans over the time period 2013-2015. The data is a proprietary dataset from a Swedish bank whom have chosen to stay anonymously in this paper due to the nature of the disclosed information. The analysis is carried out using a binary logistic regression model and the data is treated as panel data, where the dependent variable is the binary default/non-default variable and the independent variables are divided into demographical, financial and behavioral variables.

Based on this I have formed an explicit research purpose of this paper, which will be developed further in chapter 4. The purpose of this paper is to *evaluate the performance of the implemented policy changes with regards to specific individual variables which may predict a default.*

### 1.3 Disposition

The remainder of the paper will be the following: chapter 2 reviews the background of the subject, and exemplifies some of the historic implications. Chapter 3 reviews the theoretical background and previous literature on the subject. Chapter 4 discloses the data and the dataset structure, while also looking at the model specification and the describing the variables. Chapter 5 describes the empirical findings throughout the different regressions and check the results robustness. Finally, chapter 6 finishes off with some concluding remarks and some managerial suggestions.

# Chapter 2

## Background

### 2.1 History

The first credit scoring system implemented for banks and other creditors was introduced as early as in the fifties in the U.S, at that point handling only minor loans and credit card issuing. Mortgage loans issuers did not implement this kind of automated system until the 1990s as the market for automated statistical credit scoring as a method for underwriting and approving credits had grown considerably (Straka, 2000, p.211). The first implemented and widely used credit scoring model was the multiple discriminant credit scoring analysis, since then other models have been developed where previous research points at four specific methods when creating a credit scoring model: *i) linear probability model, ii) logit model, iii) probit model* and finally *iv) multiple discriminant analysis model* (Allen et al, 2004). These models can all differentiate default probabilities of consumer loans by identifying underlying financial and non-financial variables. The main task of these models is to minimize credit risk and avoid issuing credits to “bad consumers” and not to reject “good consumers”.

Lack of applicable credit scoring models can have horrific outcomes for both the lenders, debtors and society at large. During the subprime mortgage crisis in the United States 2007-2010, which was mainly driven by lack of application screening regarding their financial situation and improper use or lack of credit scoring for mortgage buyers. This resulted in a nation-wide, and soon thereafter, international financial crises where lenders throughout the world had accumulated losses of somewhat 4 percent out of the \$23.21-trillion credit market (Bianco, 2008). This occurrence points towards the importance of a formalized and unbiased credit scoring process, and moreover the practical implications of a computerized scoring system that is both time- and cost effective.

## 2.2 Regulatory implications: Basel II & III

The aim for Basel III as well as for Basel I and II is to maintain banks' solvency, this is done by strengthening the regulations, risk management and supervision of the banking industry. This means capital adequacy requirement which limit the amount of assets that a bank may have in comparison to its capital, i.e. that losses in the banks assets can be absorbed without crippling the rights of the creditors and other depositors (Edwards et al, 2010). Being able to state the level of credit risk is an important factor in Minimum Capital Requirements calculations which is an important part in all of the Basel Accords, as of the better ability one has to estimate the level of credit risk the lower capital requirements are needed. In order for banks to handle both greater lending volume and in a more time effective way, the ability to accurately predict credit defaults have become increasingly important for both the financial sector and the society at large. Although the institutions are allowed to use historical data to estimate default probabilities, a majority of lenders use risk-ratings provided by credit rating agencies (Edwards et al, 2010). Looking to the Swedish market Upplysningscentralen, UC, provides an efficient risk-scoring for both individuals and mid-sized companies whereas the risk-percentage given indicates the probability of the applicant defaulting within the next 12 months (Upplysningscentralen, 2016). The UC-score, as well as other agency ratings, is based on financial and demographic variables which are usually updated every second month. This may result in a lagged risk score, and is one of the reasons the Basel Committee have been criticized for not prohibiting solely external credit ratings for banks (BIS, 2016).

# Chapter 3

## Theoretical Framework

### 3.1 Summary

Long time traditional methods of deciding of whether to grant credit to an applicant have been through human decision in assessing the applicants default risk, most often based on previous experience (Hand and Henley, 1997 p. 523). This method has been proved successful over many years of practice, although with an ever growing demand in the credit markets with increased volumes and more efficient rulings, the need to make the application process more efficient and accurate have become crucial. Single consumer loans are usually quite small compared to the creditors total lending, and therefore it is not cost effective to assess each and every loan application individually, ergo the individual risk is minimal although the total credit portfolios risk is crucial to keep minimized and in check. Much of todays credit application channels are computerized and the granting process are usually based on sophisticated statistical models, e.g. logistical regression models calculate the probability of default given certain explanatory variables, that means using applicants financial, demographic and behavioral characteristics to predict the likelihood for an applicant not being able to repay the credit, ergo default. This is often called *credit scoring* and leads to an accept or reject decision for the creditor depending on the default risk threshold (Hand and Henley, 1997, p. 524).

## 3.2 Credit scoring models

Over the past 30 years the need to measure credit risk has evolved dramatically due to a number of secular forces that have increased the importance more than ever before. Forces that have facilitated this need have been more competitive margins on credits due to declining values of real assets (and thereby collateral) and a worldwide structural increase in number of bankruptcies. In light of these developments academics and practitioners started developing more sophisticated credit-scoring models/warnings-systems (Altman and Saunders, 1998, p.1722), the most widely used techniques for building score-card models have historically been discriminant analysis and linear regression models. Both these methods give a coefficient and a numerical score of the characteristic which is then combined into giving a single impact coefficient, this in turn is added to give an overall score.

In development of a score card the data often used comes from applicants with already granted credits, historic application data usually contains values of their characteristics and also information on previous repayment behavior and financial information from credit agencies. Other external influences can also change a person's propensity to default, e.g. socioeconomic changes or personal relations, these variables are most commonly also included in a more sophisticated behavioral scoring model. Even if default risk is the main focus when looking at the credit scoring industry, one can debate that risk measures is only one aspect of the overall credit granting decision (Hand and Henley 1997, p.524). For any bank or financial institution there is an aspect of maximizing their profits and in doing so they may take on a certain amount of risk in order to increase profitability, since profitability and risk is not always monotonically related. As to people with an exemplary economic behavior, those "high" risk consumers who do not pay their credit card bill in full each month and instead pay a higher interest rate tend to be more profitable for the creditor. The same goes for issuing loans, only accepting minimum risk consumers will effect profitability by not being able to collect on higher interest rate payments and late payment fees. Although banks are prone to accept certain amount of risk due to higher profitability, the need to accurately calculate the accepted risk is always constant, both in terms of being able to measure the total risk portfolio and through the Basel accords liquidity and asset regulations. The application of statistical methods as the credit scoring method has mainly been used for the application stage

in credit issuing, it can and nowadays is more commonly used in all stages of the credit life cycle. It is being used to measure, understand, predict, monitor and detect all aspects of consumer behavior (Anderson 2007).

Credit default predictors have been studied through many different financial and statistical models, where one of the most commonly used is the logistic regression model. This model is frequently used in many binary regression scenarios, although there is a limited amount of previous studies on the consumer credit market with an employed binary logistic regression model. Most of the previous studies are conducted on U.S and Asian data, where studies on the European and the Swedish market are negligible. However, the earliest literature on probability of default and CSM came around the time the first consumer loan was granted. It is proven in more recent studies that the automation of credit scoring and the development of more sophisticated CSMs reduces the number of defaults (Straka, 2000, p.215). Below I will discuss some of the most often used models in CSM.

### 3.2.1 Discriminant analysis

Discriminant analysis was first encountered by Durand (1941) as it showed to produce good predictions of credit repayment. A critical assumption when using the discriminant analysis method is that the variable samples are treated as multivariate normally distributed. This assumption dictates that the variables are being significance tested. Since a significant part of both financial and demographic information is not normally distributed, practitioners have developed the use of regression analysis as an alternative method in credit scoring (Hand and Henley, 1995, p.533).

### 3.2.2 Regression

Orgler (1970) started using regression analysis for a loan scoring model for commercial loans, later on he used regression analysis to construct a score-card model to evaluate outstanding debt Orgler (1971). This way he used information on behavior to construct one of the first credit-scoring models with behavioral variables. His conclusion pointed to using behavioral variables before application characteristics.

### 3.2.3 Logistic regression

Logistic regression is binary regression method which employs a binary dependent variable and a set of independent variables. In theory this method might seem more appropriate when employing a credit scoring purpose since it evaluating two discrete classes, namely default and non-default (Henley, 1995). One study concluded that logistic regression was not significantly better than ordinary linear regression when looking at applicants between a certain risk-interval. In comparison to discriminant analysis logistic regression, when applied to credit scoring, gives far better results according to a study by Wiginton (1980) (Hand and Henley, 1997, p.534). The discussion surrounding credit scoring and its efficiency have long been regarding the use and cost efficiency with practical use. Logistic regression models are usually more complex and therefor costlier to employ in scoring models.

### 3.2.4 Mathematical programming methods

This can be used given a certain objective criterion to optimize, e.g. minimize the perceptron criterion which is a linear function of the sum of the points corresponding to the applicants who are misclassified by using linear programming (Hand, 1981). Other areas of use can be to minimize the number of incorrectly classified loan applicants.

### 3.2.5 Recursive partitioning

Recursive partitioning, also called decision tree methods, are used throughout many areas there among managerial decision trees and credit scoring models. The method attempts to correctly classify members of a population based on a binary response variable<sup>1</sup> (Boye et al, 1992, p.10).

---

<sup>1</sup> Boye et al (1992)

### 3.2.6 Expert systems

This system is a sort of online questionnaire guidance, by system generated questions one is guided to a correct answer, in CSM terms questions to determine good or bad applicants. This model makes it very convenient to get feedback on why an applicant was rejected in the process.

### 3.2.7 Neural networks

This type of model is sometimes employed in credit scoring problems and can be described as a statistical model involving linear combinations of nested sequences of non-linear transformations of linear combinations of variables, as described by Hand and Henley (1997). This is a model with nowadays rare practical applications, creditors chose to employ more well established CSMs.

### 3.2.8 Smoothing nonparametric methods

Nonparametric methods have been used in a much broader range when it comes to credit scoring applications by e.g. Chatterjee and Barcun (1970) (Henley and Hand 1996). One of the first applications was on personal loans where the applications was classified on the basis of the proportion of the cases with identical characteristic vectors which belonged to the same category. A renowned model is the nearest neighbor method with a dynamical feature of adding applicants to the model when their true class becomes known, and then by dropping older cases. It is not widely used due to its computational demand.

### 3.2.9 Time varying models

As most credit scoring models attempt to distinguish good debtors from bad, this may not be the best solution at all times for a profit maximizing organization. An applicant's propensity to default will vary over time as their circumstances vary. A time varying model is a profit based approach to determine if the creditor should accept a new loan on the basis of timely updating the default probability (Henley and Hand, 1996).

# Chapter 4

## Methodology

### 4.1 Employed regression and framework

Logit and probit models are both used when investigating binary response models. Both models share similar properties and in the following analysis I have chosen to use the logit model. The binary logistic model (logit) estimates the probability of the binary response based on the predictive variables (characteristics).

Since this paper investigates the underlying risk characteristics of the banks customers in order to evaluate the implemented changes in the banks credit policy and risk model, I have chosen to divide the predictive variables into three categories. The first contains variables connected to customers' direct financial ability to repay the lender. Secondly a set of variables with indirect causality on repayment ability, e.g. demographic characteristics, and finally a set of behavioral variables.

### 4.2 Research purpose

Looking to previous literature on the subject from both academia and practitioners, given a sufficient sample size, one should be able to detect some of the underlying drivers with significant impact on probability to default among the credit applicants. This way I would be able to evaluate the implemented changes regarding the banks updated credit policy and risk model. This leads to the purpose of this paper, which is to *evaluate the performance of the implemented policy changes with regards to specific individual variables which may predict a default.*

### 4.3 Data source

The used dataset for all analysis in this paper is a proprietary dataset provided by the Swedish bank for which I evaluate their risk model. This bank offers a set of services to private customers, there among savings- and transaction accounts, credit cards and private consumption loans (further denoted as credits). The bank offers credits from the amount of 10 000 SEK to 350 000 SEK and markets itself through different loan application channels with different customer groups, and by its own brand-name. The data consist of application data from 7577 approved credits between January 2013 and January 2016. The bank employs application possibility in the whole of Sweden and have in general a diverse customer stock. Through the application form and the Swedish Credit Rating Agency, Upplysningscentralen, the dataset contains approximately 20 individual variables stretching from credit specific information to personal financial and demographic variables. Out of this dataset of 7577 credits 183 have defaulted, which gives us a total default ratio of 2.302 % during the timespan of three years.

### 4.4 Model specification

#### 4.4.1 The logistic regression model

In the following analysis I use a binary logistic model, logit, as the dependent variable can take on one of two values, default and non-default. Since I use binary variables, I can not assume the distribution of the variables to be normally distributed, hence I am using a logit model. If  $y_i = 1$  (default) or  $y_i = 0$  (non-default) we have the probabilities:

$$\begin{aligned}\Pr(Y_i = 1) &= \pi_i \\ \Pr(Y_i = 0) &= 1 - \pi_i\end{aligned}$$

Here  $y_i$  is a realization of a random variable  $Y_i$  that can take the values one and zero with probabilities  $\pi_i$  and  $1 - \pi_i$ . This can be written in a more compact form:

$$\Pr(Y_i = y_i) = \pi_i^{y_i} (1 - \pi_i)^{1-y_i},$$

for  $y_i = 0, 1$ .

Next we assume that we can model this probability with a single linear regression:

$$\pi_i = x_i' \beta,$$

where  $\pi_i$  takes a value between 0 and 1. Although this model is possible to use it might not be optimal since the righthand side can take any value, while the lefthand side takes a value between 0 and 1, this is a problem when trying to predict  $\pi_i$  given the estimated model. This is where a logit transformation comes in handy, where the regression model is defined as:

$$\pi_i = \frac{\exp(x_i' \beta)}{1 + \exp(x_i' \beta)},$$

where  $\pi_i$  and  $x_i' \beta$  is according to the linear model. Further, to be able to estimate the model we maximize the log-likelihood function<sup>2</sup>:

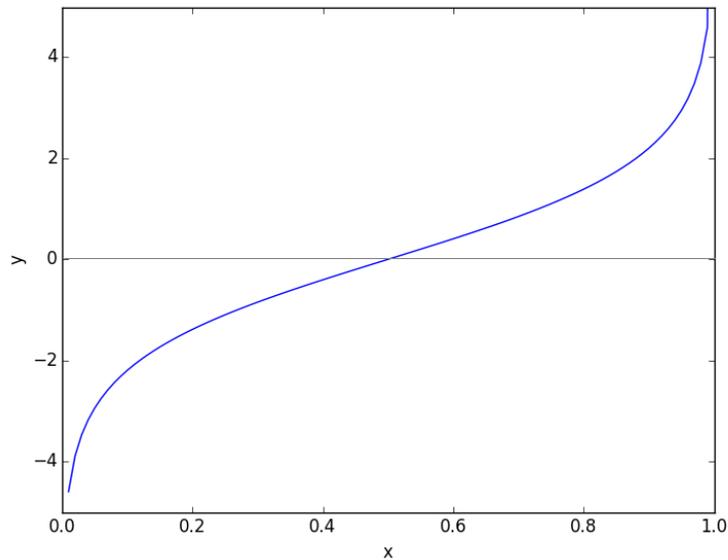
$$\begin{aligned} \ln L(\beta) &= \sum_i [y_i \ln(\pi_i) + (1 - y_i) \ln(1 - \pi_i)] \\ &= \sum_i \left[ y_i \ln \left( \frac{\exp(x_i' \beta)}{1 + \exp(x_i' \beta)} \right) + (1 - y_i) \ln \left( 1 - \frac{\exp(x_i' \beta)}{1 + \exp(x_i' \beta)} \right) \right] \end{aligned}$$

---

<sup>2</sup> Verbeek, 2012, p.208

**Figure 1: Logit function graph**

Plot of logit function where on the x-axis the probability  $\pi_i$  takes values from 0 to 1.



#### 4.4.2 Estimated model

The first estimated model, model 1, contains the full dataset from my employed bank, divided into binary and categorical variables. The second model, model 2, contains the revised dataset according to the newly implemented credit policies. Both models consist of three regression, 1<sup>st</sup> demographic variables which provide information on indirect financial ability, 2<sup>nd</sup> financial variables which provide information on direct financial ability and 3<sup>rd</sup> behavioral variables which look at the applicants behavioral pattern that might have impact on their repayment behavior.

Model 1 and 2:

#### **Equation 1: Demographic variables**

$$DEFAULTED = \beta_0 + \beta_1(AGE) + \beta_2(GENDER) + \beta_3(MAJORCITY) + \beta_4(NRCHILDREN) + \beta_5(SIZEHOUSEHOLD) + \varepsilon_i$$

**Equation 2: Financial variables**

$$DEFAULTED = \beta_0 + \beta_1(ANNUALINCOME) + \beta_2(BLANCO) + \beta_3(CREDIT) + \beta_4(INCREASEDRISK) + \beta_5(KALP) + \beta_6(MORTGAGELoAN) + \beta_7(PAYPLAN) + \beta_8(SCORE) + \varepsilon_i$$

**Equation 3: Behavioral variables**

$$DEFAULTED = \beta_0 + \beta_1(COSIGNER) + \beta_2(LOANSIZE) + \beta_3(MONTHLY) + \beta_4(PHONE) + \beta_5(REPAYPERIOD) + \beta_6(YEARSUMPERINCOME) + \varepsilon_i$$

## 4.5 Econometric assumptions

When employing a binary logistical regression model one have to make certain assumptions of the validity of the dataset and the specification of the structural model. Maximum likelihood estimations of the coefficients require that the model is in fact a good description of the behavior of the intended population, and that the dataset reflects the characteristics of the population sample that is available. Problems that may however arise involves leaving out relevant explanatory variables or dealing with misspecification due to observation problems, some of the most common issues are discussed below.

### 4.5.1 Omitted variables

Omission of relevant explanatory variables in binary models often leads to inconsistent likelihood estimations. The magnitude of the bias depends on the strength of the relationship among the included and excluded variables, whereas if the excluded variables are uncorrelated with included variables there is a case of unbiasedness although the opposite is quite more common (Menard, 2002, p.69). In this study most of the available variables from the complete dataset have been used in model 1 and 2 in accordance with previous CSM studies. The full sample model is then compared to other models with dropped variables in order to study potential bias estimations.

#### 4.5.2 Covariate measurement error

The problem that usually occurs when instead of measure observation  $X$  we observe  $W$ , where  $W$  is a function of  $X$  and  $U$ ,  $W = X + U$ , where  $U$  is an error term. This may effect the intercept of the variable and contribute to a loss of the underlying trend due to increased number of error terms. In this sample the data has already been provided by the bank and other credit agencies which makes measurement errors less likely and also not possible to adjust if present (Esmeralda et. al., 2012, p.6).

#### 4.5.3 Response misclassification

This problem appears when the response variable (dependent variable) have been formalized in the wrong way, the parameters are defined as the probability of observing 1 (0) when the actual response is 0 (1). This problem however is less likely in this dataset where the response variable is a careful study of individuals that has defaulted on their loans, although it may contain some loans that has defaulted after the banks internal default study (Esmeralda et. al., 2012, p.7).

## 4.6 Variables

**Table 1: Variable definition**

The table presents the definition for all variables used in the following analysis. The variables are defined as either categorically divided which indicates that the variable can take on three or more values, or as a dummy where the variable takes on binary value of zero or one. The variables are then sorted according to demographic, financial or behavioral categorize.

Variable	Definition	Demographic	Financial	Behavioral
AGE	categorical, age of applicant	x		
ANNUALINCOME	categorical, self-reported annual income of applicant (in SEK)		x	
BLANCO	categorical, previous private loan-amount (in SEK)		x	
COSIGNER	dummy, takes value 1 if two applicants			x
CREDIT	dummy, takes value 1 if applicant has one or more credit cards		x	
DEFAULTED	dummy, takes value 1 if applicant has defaulted the loan			x
GENDER	dummy, takes value 1 if applicant is a man	x		
INCREASEDRISK	dummy, takes value 1 if applicant would not pass new application by 1601		x	
KALP	categorical, applicants monthly income excess minus budget (in SEK)		x	
LOANSIZE	categorical, the amount of loan (in SEK)			x
MAJORITYCITY	dummy, takes value 1 if applicant lives in one of the 3 major cities	x		
MONTHLY	categorical, the amount the applicant repays every month (in SEK)			x
MORTGAGELoan	dummy, takes value 1 if the applicant has a mortgage loan		x	
NRCHILDREN	categorical, nr of children in household	x		
PAYPLAN	categorical, previous repayment plans amount (in SEK)		x	
PHONE	dummy, takes value 1 if applicant has provided a phone number			x
REPAYPERIOD	categorical, number of years to repay the loan			x
SCORE	categorical, the score the applicant has received by the lender		x	
SIZEHOUSEHOLD	categorical, nr of persons in the household	x		
YEARSUMPERINCOME	categorical, the early repayment amount to yearly income (in SEK)			x

## 4.6.1 Descriptive statistics

**Table 2: Descriptive statistics**

The table presents the descriptive statistics for the sample in full. The columns show total number of observations of a specific variable as well as the mean, standard deviation and min/max value of the variable. See appendix for full variable-category table.

Variable	Number of obs	Mean	Std. Dev.	Min	Max
AGE	7391	49.46	11.40	29	79
ANNUALINCOME	7394	359480.70	52927.81	150000	400000
BLANCO	7392	247280.80	91744.43	100000	350000
COSIGNER	7342	0.21	0.41	0	1
CREDIT	7394	0.60	0.49	0	1
DEFAULTED	7394	0.01	0.11	0	1
GENDER	7389	0.70	0.46	0	1
INCREASEDRISK	7305	0.11	0.31	0	1
KALP	7361	6865.13	4707.60	1000	16000
LOANSIZE	7394	209223.70	96156.87	70000	370000
MAJORITYCITY	7305	0.03	0.18	0	1
MONTHLY	7394	1958.82	727.68	500	3000
MORTGAGELOAN	7394	0.81	0.39	0	1
NRCHILDREN	7394	0.76	0.90	0	7
PAYPLAN	7379	1278.29	1150.99	500	3000
PHONE	7394	0.71	0.45	0	1
REPAYPERIOD	7394	8.71	2.85	1	12
SCORE	7394	0.01	0.01	0.005	0.05
SIZEHOUSEHOLD	7394	2.03	1.09	1	9
YEARSUMPERINCOME	7394	0.06	0.04	0	0.93

## 4.6.2 Dependent variable

Throughout the following analysis the dependent variable is `DEFAULTED` which is a binary variable with two outcomes, default if the customer has no repayment ability or non-default if the loan is still active or has been paid in full. The subject bank uses the same definition of default as which the Basel accords states, the loan has defaulted if the payments are more than 90 days overdue. Past this point the bank sells the defaulted loan to a debt-collecting agency.

In this particular bank the default levels are low, with a mean value of the default variable at 0.01 and standard deviation of 0.11, meaning that the absolute most of the customer repay their loans.

### 4.6.3 Explanatory variables

#### *Demographic*

*Age* is the first demographic variable and have been categorized into six groups with applicants between the ages 18-29, 30-39, 40-49, 50-59, 60-60 and 70-79. The mean age of the applicants is 49 years of age. *Gender* is the second variable and it is a dummy variable where it takes the value of 1 if the applicant is a man and 0 if a woman. The mean value is 0.70 which means that 70 % of the applicants are men. *Major city* represents if the applicants live in one of Sweden's 3 largest cities, Stockholm, Gothenburg or Malmö, if true it takes the value of 1. This variable tries to control for potential higher living cost that comes from living in a major city. *Nr children* is a categorized variable of how many children lives full or part-time in the household. Many children would indicate a higher living cost and therefore higher tendency of a marginalized economy. The mean applicant has 0.76 children. *Size household* also controls for cost increase but also a potential higher total income in the household. The mean applicant lives in a household of 2.

#### *Financial*

*Annual income* is the categorized yearly income of applicants into six groups ranging from less than 150 000 SEK to more than 400 000 SEK. This information is self reported, but if this information differs more than 30 % from the reported taxed income the applicant must prove its income. The mean income of the sample is almost 360 000 SEK which is in terms of purchasing power a decent yearly salary. *Blanco* is previous unsecured loan variable reported by Upplysningscentralen, which is here divided into six amount-groups ranging from less than 100 000 SEK to more than 350 000 SEK. The mean value of the sample is 247 000 SEK, indicating that the average applicant has high amounts of unsecured loans. *Credit* is a dummy variable of value 1 if the applicant has one or more credit cards, the mean value tells us that 60 % of all approved applicants have access to one or more credit cards. *Increased risk* is a dummy variable of all sampled applicants credit score by 1<sup>st</sup> of January 2016 of which

customers with a credit score above 5 % are given the value 1, as these customers would not pass a new loan application. *Kalp* is the categorized difference between the customers' income excess and the budgeted minimum excess on a monthly basis. *Kalp* stretches from less than 1000 SEK to more than 16 000 SEK in income excess. *Mortgage loan* is a binary variable which takes the value of 1 if the applicant has a mortgage loan, this might indicate that the customer has had the financial possibility to buy a home which would indicate some level of economical wellbeing. The mean value is 0.81 which tells us that 81 % of the sample has a mortgage loan. *Pay plan* is a categorical variable showing previous repayment plans ranging from less than 500 SEK to more than 3000 SEK, with a mean value of almost 1300 SEK. *Score* is the banks risk score measured as the credit score reported by Upplysningscentralen. It is categorized from 0.5 % up till 5 % with a mean value of 1 %.

### *Behavioral*

*Co-signer* is a dummy variable taking the value of 1 if the applicant has a co-signer for the loan, usually needed if the applicant's financial situation is not sufficient. In this sample only 21 % of the loans are co-signed. *Loan size* is a categorical variable of the size of the loan ranging from less than 70 000 SEK to 370 000 SEK with a mean value of 209 000 SEK. A high amount in a unsecured loan might suggest car purchase or other consumption financing, neither with any return on investment in general. *Monthly* is what the applicant is going to repay every month on the loan categorized between 500 and 3000. Lower amounts speak to the customers general financial wellbeing. *Phone* is a dummy variable taken the value 1 if the applicant chose to provide a phone number as contact information for the bank. The choice to provide voluntary contact information might indicate an agenda to repay the lender. *Repay period* is the number of years the customer should repay the in, usually longer repayment periods mean higher risk for the lender, this samples mean is 8.7 years. *Year sum per income* is a ratio variable of the chosen repayment amount per year compared to the applicant's annual income. A high variable value suggests a higher share of the applicant's annual income being spent on loan payments, which is more unlikely. This sample shows a mean of 0.06, meaning that the mean applicants should repay loan amounts corresponding to 6 % of the annual salary.

# Chapter 5

## Empirical findings

### 5.1 Regression models

The first regression model analyzes the full application dataset in order to identify default characteristics among the applicants as of previous credit policy within the bank. The second regression model analyze the revised dataset, according to new implemented credit policies, in order to evaluate the efficiency of the implemented policy changes and contribute to further improvements. Both regression models use the full sample of defaulted loans which then is supplemented by the explanatory samples of increasing size of the non-defaulted loans (Cramer, 2004, p.369). Starting with the 183 defaulted loans compared to 183 non-defaulted loans, then adding the randomized sample of non-defaulted loans according to  $K=1$  for 183 non-defaulted loans, and further  $K=2, 5, 20$  & *full sample*.

The first three regression tables shows the base sample before implemented credit policy changes for  $k=1, \dots, all$  with the dependent variable **DEFAULTED** and the three explanatory variable categorize. The different sample sizes of non defaulted loans (good loans) have been randomized and the grouping of the categorized variables is based on previous studies within scoring models.

### 5.1.1 Demographic

**Table 3: Model 1 – Demographics** The first regression model is based on the three variable categorize: demographic, financial and behavioral. The sample contain random selections for  $k=1, \dots, all$  and has been regressed using a logit model in the STATA 12 software. The table shows the categorized and binary results for the demographic regression and presents the coefficients as marginal effects and the p-value for significance level where \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ . Robust standard errors in appendix table.

DEFAULTED	<b>K=1</b>	P> z	<b>K=2</b>	P> z	<b>K=5</b>	P> z	<b>K=20</b>	P> z	<b>K=all</b>	P> z
<i>Num bad loans</i>	183		183		183		183		183	
<i>Num good loans</i>	183		366		915		3660		7394	
<b>Demographic</b>										
AGE										
29	0.242	**	0.24	***	0.021	***	0.634	***	0.319	***
39	0.102		-0.008		-0.012		-0.001		-0.002	
49	-0.042		-0.16	**	-0.089	**	-0.022	**	-0.01	*
59	-0.232	**	-0.24	***	-0.121	***	-0.034	***	-0.019	***
69	-0.182		-0.128		-0.095	**	-0.023	**	-0.016	***
79	0.074		0.198		0.092		0.008		-0.007	
GENDER	0.095		0.061		0.028		0.009	*	0.004	*
MAJORITY					0.398		-0.23		-0.007	
NRCHILDREN	0.033		0.095		0.046		0.013		0.002	
SIZEHOUSEHOLD										
2	-0.374	***	-0.267	***	-0.174	***	-0.052	**	-0.009	
3	-0.421	***	-0.363	***	-0.206	***	-0.062	**	-0.015	**
4	-0.613	***	-0.443	***	-0.245	***	-0.071	***	.	
5	-0.443		-0.445	***	-0.24	***	-0.071	***		

The first regression uses demographic explanatory variables where AGE shows highly significant coefficients of increased default probabilities for younger applicants between the age of 18-29. Higher age tends to decrease the probability of default throughout all sample sets, these results are consistent with theory of younger people are in more need of consumption financing and have less financial stability and therefore a higher risk of not

being able to repay you loans. GENDER is significant when looking at the larger sample sets, where it points towards slightly higher default probabilities for men in difference to women. NRCHILDREN is insignificant for all sample sets and is therefore inadmissible. SIZEHOUSEHOLD further shows that the larger household size of the applicant the less likely is the probability of default, this might be explained by having afford of having a large family and therefor the financial wellbeing of repaying your loans.

### 5.1.2 Financial

**Table 4: Model 1 – Financials** The sample contain random selections for  $k=1, \dots, \text{all}$  and has been regressed using a logit model in the STATA 12 software. The table shows the categorized and binary results for the financial regression and presents the coefficients as marginal effects and the p-value for significance level where \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ . Robust standard errors in appendix table.

DEFAULTED	K=1	P> z	K=2	P> z	K=5	P> z	K=20	P> z	K=all	P> z
<i>Num bad loans</i>	183		183		183		183		183	
<i>Num good loans</i>	183		366		915		3660		7394	
<b>Financial</b>										
ANNUALINCOME										
200000	-0.553	***	0.032	*	0.053		0.009		0	***
250000	-0.336	***	0.07	*	0.073		0.009		0	***
300000	-0.096		-0.029		0.103		0.01		0.002	***
350000	-0.356	***	-0.091		0.067		0.004		0	***
400000	-0.335	***	-0.039		0.066		0.008		-0.001	***
BLANCO										
150000	-0.447	***	-0.052	***	0.018	***	-0.003	***	-0.002	***
200000	0.105	***	0.129	***	0.031	***	0.007	***	0.004	***
250000	-0.085	***	-0.006	***	0.022	***	0.004	***	0.002	***
300000	-0.205	***	-0.032	***	-0.026	***	-0.008	***	-0.001	***
350000	-0.249	***	-0.028	***	-0.023	***	-0.009	***	-0.002	***
CREDIT	-0.255	***	-0.136	***	-0.064	***	-0.017	***	-0.006	***
INCREASEDRISK					0.398		-0.23		-0.007	
KALP										
4000	-0.073		-0.123	*	-0.039		-0.012	*	-0.002	
7000	-0.239	*	-0.105		-0.03		-0.01		0.001	
10000	-0.016		-0.111		-0.042		-0.009		0.001	
13000	0.182		-0.186	**	-0.054		-0.016	*	-0.002	
16000	-0.061		-0.16	*	-0.036		-0.014	*	0.002	
MORTGAGELOAN	-0.403	***	-0.178	***	-0.082	***	-0.018	***	-0.006	***
PAYPLAN	0	***	0	***	0	***	0	***	0	***
SCORE	22.277	***	13.48	***	5.251	***	1.209	***	0.485	***

The second regression uses explanatory variables of direct financial ability of the applicant, where ANNUALINCOME only shows significance for smaller and large sample sets with different coefficient signs which makes this variable inconsistent. BLANCO which tells how much the previous unsecured loans the applicant has shows significance for all sample sets and where having loans between 150 001-200 000 SEK shows highest probability of defaulting. CREDIT shows strong significant coefficients of decreased default probability for applicants having one or more credit cards. KALP which tells us about the income excess on a monthly basis compared to the banks budgeted excess demand, the significant coefficients show a slight decrease of default the higher the excess is. MORTGAGELoAN shows a decrease in default probability if the applicant has a mortgage loan, ergo own a home. This might be explained by a more likely well-managed personal economy and an asset to liquidate if needed. PAYPLAN shows an indifference in probability impact on default. SCORE which is the banks risk measure of an applicant, shows as expected a highly increased default probability when given a higher score.

5.1.3 Behavioral

**Table 5: Model 1 – Behaviorals** The sample contain random selections for  $k=1, \dots, all$  and has been regressed using a logit model in the STATA 12 software. The table shows the categorized and binary results for the behavioral regression and presents the coefficients as marginal effects and the p-value for significance level where \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ . Robust standard errors in appendix table.

DEFAULTED	<b>K=1</b>	P> z	<b>K=2</b>	P> z	<b>K=5</b>	P> z	<b>K=20</b>	P> z	<b>K=all</b>	P> z
<i>Num bad loans</i>	183		183		183		183		183	
<i>Num good loans</i>	183		366		915		3660		7394	
<b>Behavioral</b>										
COSIGNER					-0.488	***	-0.331	***	-0.177	***
LOANSIZE										
130000	0.343	**	0		-0.046	***	0.142	**	0.046	**
190000			-0.007		-0.239	***	0.142	**	0.056	**
250000			-0.005		0.06	**	0.387	***	0.124	**
310000					-0.353	***	0.178	**	0.026	*
370000			-0.981		-0.437	***	0.39	***	0.059	**
MONTHLY										
1000					0.064		-0.127		0.007	
1500					0.398		-0.23		-0.007	
2000					0.301		-0.213		0.035	
2500					0.615	*	-0.191		0.069	
3000					0.622	*	-0.143		0.039	
PHONE			-0.087		-0.428	***	-0.321		-0.184	***
REPAYPERIOD	0.082		0.009		0.028		-0.008		-0.002	
YEARSUMPERINCOME	7.091		0.35		-2.404		0.677		0.291	

Unfortunately, some of the sample sets contained insufficient amount of information given some of the variables. COSIGNER shows highly significant coefficients of decreased default probability given a second applicant on a loan. This can be explained by the solidary responsibility of repaying the loan and the financial backup of two persons. LOANSIZE overall shows twice the probability of loans going bad for amounts of 190 001-250 000 SEK. This might point to an approval discrepancy or the number of loans ranging this amount. MONTHLY gives insignificant and inconsistent results, PHONE on the other hand shows significant probability decrease for applicants providing voluntary information, e.g. phone number. REPAYPERIOD and YEARSUMPERINCOME are both insignificant and therefore excluded in the discussion.

## 5.2 Implemented policy changes – Model 2

The second three regression tables shows the altered sample according to the new implemented credit policy changes for  $k=1, \dots, all$  with the dependent variable `DEFAULTED` and the three explanatory variable categorize. The different sample sizes of non defaulted loans (good loans) have been randomized, as in model 1, and the grouping of the categorized variables is based on previous studies within scoring models. The purpose of the newly implemented credit policy changes is to lower the probability of default for approved applicants, a successful implementation would reflect in fewer significant coefficients and less decisive risk probabilities.

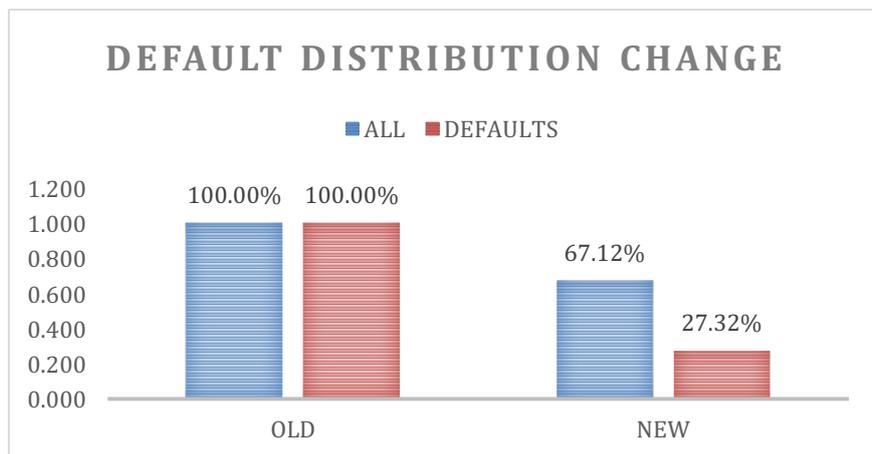
**Table 6: Policy development**

Implemented credit policy changes according as from January 1<sup>st</sup> 2016 compared to previous policy.

	Old	New
<b>Income measure</b>	Self reported complemented by taxed	Only taxed income
<b>KALP budget</b>	Individually calculated minimum level	Individually calculated minimum level + 2000
<b>Debt levels</b>	Maximum of 100 % Blanco to annual income	Dynamic level according to new income and Kalp policy
<b>Risk score</b>	According to credit reporting agency	Previous + debt level + payment reminder + age 71-75
<b>Application channel</b>	10 different application channels	9 application channels, 1 has been removed

As the table reveals has model 2’s dataset been modified according to new policy. This results in out of 183 previous defaults only 50 defaults have been included in the new sample due to avoided “bad” loan approvals. The total sample shrunk from 7577 to 4963 due to the changes and looking at the default distribution within the samples it originated in 2.302 % before policy changes and is at 1.008 % as of the revised sample.

**Figure 2: Default distribution**



5.2.1 Demographic

**Table 7: Model 2 – Demographics** The second regression model is based on the three variable categorize: demographic, financial and behavioral. The sample contain random selections for  $k=1, \dots, all$  for the revised sample according to new credit policy and has been regressed using a logit model in the STATA 12 software. The table shows the categorized and binary results for the demographic regression and presents the coefficients as marginal effects and the p-value for significance level where \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ . Robust standard errors in appendix table.

DEFAULTED	K=1	P> z	K=2	P> z	K=5	P> z	K=20	P> z	K=all	P> z	
<i>Num bad loans</i>	50		50		50		50		50		
<i>Num good loans</i>	50		100		250		1000		4963		
<b>Demographic</b>											
AGE											
	29	0.039	0.012	**	0.533	**	0.171		0.063		
	39	0.001	0.257		0.058		0.01		0.004		
	49	-0.15	0.087		0.024		-0.002		0.001		
	59	-0.127	0.019		-0.06		-0.02		-0.004		
	69	-0.182	-0.077		-0.058		-0.021		-0.004		
	79										
GENDER	0.047		0.039		0.012		0.015		0.002		
MAJORCITY	0.091		0.1		-0.006		-0.005		0		
NRCHILDREN	0.533	**	0.4	**	0.171	**	0.063	***	0.012	**	
SIZEHOUSEHOLD											
	2	-0.448	***	-0.567	***	-0.319	**	-0.17	**	-0.043	*
	3	-0.719	***	-0.707	***	-0.369	**	-0.186	**	-0.046	*
	4	-0.827	***	-0.77	***	-0.426	***	-0.193	**	-0.048	**
	5										

Looking at the second regression table and the demographic variables we see that AGE only shows significant values for ages 18-29 with increased default probabilities for these ages. This is consistent to model 1 and still shows overrepresentation of default probabilities within this age group. GENDER nor MAJORCITY shows any significant coefficients. SIZEHOUSEHOLD gives significant but indifferent coefficient values indicating resolved risk factor.

5.2.2 Financial

**Table 8: Model 2 – Financials** The sample contain random selections for  $k=1, \dots, all$  for the revised sample according to new credit policy and has been regressed using a logit model in the STATA 12 software. The table shows the categorized and binary results for the financial regression and presents the coefficients as marginal effects and the p-value for significance level where \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ . Robust standard errors in appendix table.

DEFAULTED	K=1	P> z	K=2	P> z	K=5	P> z	K=20	P> z	K=all	P> z
<i>Num bad loans</i>	50		50		50		50		50	
<i>Num good loans</i>	50		100		250		1000		4963	
<b>Financial</b>										
ANNUALINCOME										
200000										
250000										
300000	0.317		0.14							
350000	0.738		0.336							
400000	0.524		0.255							
BLANCO										
150000	-0.684	***	-0.355	***	-0.259	**	-0.047	**	-0.006	**
200000	0.062		-0.045	***	-0.038	***	-0.004	**	0.003	**
250000	-0.284	***	-0.026	***	-0.149	***	-0.029	**	-0.001	**
300000	-0.348	***	-0.134	***	-0.164	***	-0.03	**	-0.002	**
350000	-0.465	***	-0.224	***	-0.215	**	-0.041	**	-0.003	**
CREDIT	-0.819	***	-0.311	**	-0.099	**	-0.031	***	-0.006	***
INCREASEDRISK	0.017		0.038		0.045		0.035		-0.038	
KALP										
4000	0		-0.19		-0.016		-0.063		-0.004	
7000	0.418	*	0.153		0.007		-0.002		0	
10000	0.39	*	0.51	***	0.028		0.012		0.001	
13000	0.448		0.459	**	0.024		-0.006		-0.001	
16000	0.091		0.1		-0.006		-0.005		0	
MORTGAGELOAN	-0.314		-0.147		-0.084	*	-0.008		-0.004	*
PAYPLAN	0	**	0	***	0	***	0	**	0	*
SCORE	47.05	***	22.747	***	10.177	***	2.137	***	0.42	***

BLANCO shows significant values for all groups but indifference to model 1 the coefficients do not point towards any particular risk increase for any amounts. CREDIT still shows less likely probability of defaulting when having one or more credit cards. KALP gives few

significant values and have changed signs, now showing indifference in KALP amounts effect on default probabilities. MORGAGELOAN still contribute to a slight decrease in default probability as of when owning a home. PAYPLAN do not effect the default risk either way, and finally SCORE shows a strong causal relationship with increased default risk.

### 5.2.3 Behavioral

**Table 9: Model 2 – Behaviorals** The sample contain random selections for  $k=1, \dots, all$  for the revised sample according to new credit policy and has been regressed using a logit model in the STATA 12 software. The table shows the categorized and binary results for the behavioral regression and presents the coefficients as marginal effects and the p-value for significance level where \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ . Robust standard errors in appendix table.

DEFAULTED	<b>K=1</b>	P> z	<b>K=2</b>	P> z	<b>K=5</b>	P> z	<b>K=20</b>	P> z	<b>K=all</b>	P> z
<i>Num bad loans</i>	50		50		50		50		50	
<i>Num good loans</i>	50		100		250		1000		4963	
<b>Behavioral</b>										
COSIGNER	-0.781	***	-0.438	***	-0.175	***	-0.042	***	-0.007	
LOANSIZE										
130000	0.158		0.159		0.038		0.013		0.002	
190000	-0.16	***	0.161		-0.026		0.016		0.002	
250000	-0.406	**	0.131		-0.017		0.02		0.003	
310000	-0.523		0.189		-0.038		0.017		0.002	
370000	-0.488		0.133		-0.058		0.009		0.002	
MONTHLY										
1000	0.086		0.045		0.035		-0.038		-0.004	
1500	0		-0.19		-0.019		-0.063		-0.007	
2000	0.309		-0.171		0.033		-0.059		-0.006	
2500	0.531		-0.174		0.042		-0.06		-0.006	
3000	0.666	*	-0.061		0.11		-0.056		-0.005	
PHONE	-0.063		-0.13		-0.08	**	-0.015	*	-0.002	
REPAYPERIOD	0.109	**	0.015		0.009		0		0	
YEARSUMPERINCOME	1.801		0.691		-0.576		-0.065		-0.016	

COSIGNER is consistent to previous results in model 1 showing that a second applicant on a loan decreases the default risk. LOANSIZE gives few significant values and is therefore indecisive as well as MONTHLY. PHONE still shows increased default risk when applicants choosing not to provide voluntary contact information.

### 5.3 Information value – Explanatory variable reduction

Not much previous literature is found on the use of information value (IV) for variable selection when constructing credit scoring models (CSM), although it originates from information theory. It has resembling features of the *Chi Square value* which is further discussed below. The use of the information value is convenient when looking at independent variable selection as it provides a clear rule of thumb (UC Analytics, 2016).

Table 10: Information Values

Model 1 Sample		Model 2 Sample	
Variable	Information Value	Variable	Information Value
SCORE	1.121899	SCORE	0.921293
PAYPLAN	0.330926	MORTGAGELOAN	0.382071
ANNUALINCOME	0.324082	AGE	0.344872
AGE	0.225491	PAYPLAN	0.237877
COSIGNER	0.169369	GENDER	0.169151
LOANSIZE	0.149894	SIZEHOUSEHOLD	0.154272
BLANCO	0.116913	BLANCO	0.148848
KALP	0.067547	LOANSIZE	0.087307
MORTGAGELOAN	0.067263	COSIGNER	0.025881
GENDER	0.054121	ANNUALINCOME	0.016774
PHONE	0.047173	CREDIT	0.015844
CREDIT	0.028718	PHONE	0.014925
SIZEHOUSEHOLD	0.015755	MONTHLY	0.007973
NRCHILDREN	0.009004	NRCHILDREN	0.007448
MONTHLY	0.001715	KALP	0.000814
MAJORITY	0.000000	YEARSUMPERINCOME	0.000000
INCREASEDRISK	0.000000	REPAYPERIOD	0.000000
REPAYPERIOD	0.000000	MAJORITY	0.000000
YEARSUMPERINCOME	0.000000	INCREASEDRISK	0.000000

The table displays the independent variables individual information values for model 1, the full baseline dataset, and model 2 which contains the revised dataset. The values are divided into four categories of predictive power where  $< 0.02$  are not applicable for prediction,  $0.02 - 0.1$  are weak predictors,  $0.1 - 0.3$  medium predictors and  $0.3 - 0.5$  are strong predictors.

Above 0.5 is suspiciously strong and indicates a high level of correlation with other independent variables, only SCORE holds this property and fits the description. The IV-model is defined as the distribution between good/bad, in our case default/non-default and looks like this<sup>3</sup>:

$$IV = \sum (Dist. NonDefault_i - Dist. Default_i) \times WOE,$$

where the Weight of Evidence (WOE) is a logit component:

$$WOE = \ln \left( \frac{Dist. NonDefault_i}{Dist. Default_i} \right)$$

As the table indicates the information value between model 1 and 2 have change and more of the explanatory variables have become less useful for predicting default. This is explained by the implemented policy changes in model 2's dataset, which have resulted in elimination of more than half of the defaulted loans. In order to further investigate the validity of the results for model 2, I will regress the variables holding strong<sup>4</sup> and medium strong predictive power in model 3 and the weaker predictive variables (including ANNUALINCOME, CREDIT and PHONE) in model 4. The Akaike Information Criterion (AIC) will further help to decide the quality of the different models. The AIC helps selecting models that have the best fitted values without over-fitting the model and the risk of losing the underlying trend among the increasing standard errors, its doing so by adjusting the RSS for the sample size (n) and the number of dependent variables (K), (Baltagi, 2013, p.195).

---

<sup>3</sup> Hand and Henley, 1997, p 529

<sup>4</sup> Including SCORE as a strong variable in model.

**Table 11: Selected variable regressions**

The table shows the reduced variable regressions, model 3 and 4 according to their individual information values. The coefficients are interpreted as the marginal effect from the logit regression and the asterisk indicates the level of significance, \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ . Robust standard errors in appendix table. AIC-value indicates the quality of the model, where higher value equals a better fitted model.

Model 3	dy/dx		Model 4	dy/dx	
SCORE	0.3614	***	LOANSIZE		
MORTGAGELoAN	-0.0052	**	130000	0.0036	
AGE			190000	0.0013	
29	0.0711	**	250000	0.0024	*
39	0.0052	*	310000	0.0023	
49	0.0024		370000	0.0024	
59	-0.0017		COSIGNER	-0.0096	***
69	-0.0022		ANNUALINCOME	0.0000	
79			CREDIT	-0.0023	*
PAYPLAN	0.0000	***	PHONE	-0.0032	**
GENDER	0.0038	*			
SIZEHOUSEHOLD					
2	-0.0041	*			
3	-0.0021				
4	-0.0075	***			
BLANCO					
150000	-0.0057	**			
200000	0.0020	**			
250000	-0.0010	**			
300000	-0.0024	**			
350000	-0.0025	**			
<b>AIC</b>	<b>462.7371</b>			<b>409.1429</b>	
<b>BIC</b>	<b>568.3321</b>			<b>474.1312</b>	
<b>Pseudo R2</b>	<b>0.1766</b>			<b>0.0641</b>	

Looking at the regression results for model 3 and 4 almost all coefficients show significant values at some level. The SCORE variable shows a high marginal effect on the default probability, confirming its sole purpose. MORTGAGELoAN tends to decrease the default probability, while age still shows significant higher default probabilities for ages 18-29. The GENDER coefficient tends to increase the probability of default by almost 0.4 % if the applicant is a man. Having a large household tends to decrease the default probability, while applicants with more than 150 000 SEK in BLANCO but less than 250 000 SEK have a

higher default probability. Model 3's set of independent variables explains about 17.6 % of the total default behavior. In model 4 I chose to include variables with low predictive power (according to the IV estimation) which also shows in the pseudo R<sup>2</sup> value of only 6.41 %. The LOANSIZE in it self barely seem to affect the default probability, although if the loan is issued to two persons the COSIGNER coefficient reduces the default probability by almost 1 %. Having previous credit cards still shows a negative relationship with the dependent variable, and finally for applicants providing voluntary contact information seem to reduce the probability of default.

### 5.3.1 Akaike Information Criterion (AIC)

AIC measures the relative quality of the different regression models by providing an optimal trade-off between the goodness of fit and the number of parameters used, in this way it provides a selection tool for different models (Baltagi, 2013, p.196). More direct it rewards the goodness of fit and penalizes an increasing number of parameters, thus additional number of parameters almost always increases the goodness of fit. Hence, in our model selection the sought of value is the model with the lowest AIC value. In difference to AIC, the *Bayesian Information Criterion*, BIC is also based on the likelihood function and is closely related to the AIC, but penalizes the number of parameters more strongly than the AIC measure (Anderson et. al., 2002, p. 914). The model with the lowest BIC value is preferred here as well. The below table shows the AIC and BIC values for the first policy changed model (model 2) regressed on all explanatory variables, and the selected variable models (model 3 and 4) according to their information values.

Model 2 shows the lowest AIC value but a significant higher BIC than the rest, ergo it provides the best fitted values although BIC raises the value due to the many explanatory variables. Model 3 is the result of the highest informative variables, including SCORE which is a result of some of the other variables, this model gives a high pseudo R<sup>2</sup> value while also using many explanatory variables, thus both the AIC and the BIC value are the highest. Model 4 only uses explanatory variables with low predictive power according to their information value. This is also shown in the pseudo R<sup>2</sup> value, but instead highlight some of

the behavioral variables, which according to Orgler (1971) possess the best explanatory power within credit scoring. Hence, the AIC value is the second lowest and the BIC value favours the few number of variables and is therefore the lowest among the three models. According to AIC the first policy implemented model best predicts the default probabilities among customers, whereas the BIC prefers model 4 with the low information variables.

**Table 12: AIC & BIC values**

	AIC value	BIC value
Model 2	354.946	582.985
Model 3	462.737	568.332
Model 4	409.143	474.131

### 5.3.2 Pearson's chi square test

The Pearson's chi square test assesses both the goodness of fit of the models distribution and the independence of the observations (Howell, 2016, p.1). As for the test result, the values for the three models are all highly significant and therefore efficient. Model 2 shows the highest value.

**Table 13: Chi square test**

	Pearson Chi-Square	Prob > chi2	df
Model 2	92.94	0.000	37
Model 3	91.940	0.000	16
Model 4	26.650	0.000	9

### 5.3.3 Log-likelihood ratio (LR) test

The table below shows the results from the log-likelihood ratio test (LR test) which is performed by running each of the constrained models, 3 and 4, against the unconstrained baseline model, 2, which gives the value of 149.79 when comparing model 2 with model 3 with 21 degrees of freedom. Further when comparing model 2 with model 4 we have the likelihood ratio of 110.20 with 28 degrees of freedom. Both ratios are highly significant and according to the LR test the best model is the nested model 2 (Idre – UCLA, 2016).

**Table 14: LR test**

	LR chi2	Prob > chi2	df	Obs
Model 3	149.790	0.000	21	4851
Model 4	110.200	0.000	28	4830

## 5.4 Interpretation of results

### 5.4.1 Model evaluation

Looking at the results from the four previous efficiency and selection tests model 2 is the clear choice for a CSM and the best model for identifying other risk factors not already employed by the newly implemented credit policy changes. The test results witness of a need for both the highly predictive variables such as SCORE and MORTGAGELOAN as well as some of the less informative behavioral variables to construct an efficient model.

**Table 15: Test evaluation**

Test	Model 2	Model 3	Model 4
AIC	x		
BIC			x
Chi2	x		
LR	x		

---

## 5.5 Discussion

This section will further discuss the results from the first two main models and the performance of the selected variable models 3 and 4 using the theoretical framework and previous research discussed in chapter 3. I will also try to develop on the specified research statement in a concise manner.

### 5.5.1 Pre policy changes – Model 1

In the first model with the full baseline dataset, prior to any implemented model changes, there is indications of significant higher default probabilities given the AGE variable between the age group of 18-29. GENDER shows a slight increase in default probability, although with a sample with a majority of male applicants this could be explained by biased sample selection. The SIZEHOUSEHOLD shows a slight decrease in risk tendency if the applicant is part of a bigger household. ANNUALINCOME shows little impact on default probabilities when looking at the full sample, whereas having previous BLANCO loans in the range of 150 001-200 000 SEK increase default probability significantly. This may be explained by the previous income – loan ratio rule of 100 %, where the median of annual incomes lies within this range, in other words it takes a higher income to carry a higher amount of BLANCO, and applicants pushing a 100 % loan ratio of unsecured debt shows indication of a certain behavior. Applicants with previous CREDITS, usually in the form of credit cards, shows a decreased default probability which is inconsistent with previous studies and theory and cannot be explained. Income excess in terms of KALP shows some higher risk tendency for applicants with less than 4 001 SEK in income excess per month. MORGAGELOAN shows a quite high decrease in risk, in Sweden banks do not lend more than 85 % of the total price of the property which inclines the buyer to finance the rest with their own capital, this means the applicant have a certain level of assets (Konsumenternas.se). SCORE which is given by the credit reporting agency *Upplysningscentralen* is their risk measure of the applicant, and as shown it gives a good estimate of applicants with higher default probabilities. If the applicant has a COSIGNER, the general probability to default decreases when the responsibility of the loan is shared. The LOANSIZE shows much higher probability of going bad for amounts of 190 000 – 250 000 SEK, which may be explained by people who do not have the inclination

to repay the loan, or for people who live highly indebted but have the intention to repay. Applicants providing voluntary information, in this case PHONE number, shows significant lower probabilities of default. For most people who voluntarily provide contact information also have the intention of fulfilling their repayment obligations, thus this shows a negative relationship with “bad” costumers.

### 5.5.2 Post policy changes – Model 2

In the second model the dataset has been transformed according to the new credit policy changes, this mean an exclusion of more than two thirds of the defaulted loans and approximately 2 400 of non defaulted loans. Looking at the first variable AGE there is still an increased default probability for applicants in the age span of 18-29, this is natural since this variable have not been directly effected by the new policy. The effect of GENDER is now insignificant, probably due to the smaller dataset. MAJORCITY still gives insignificant values. NRCHILDREN shows an increased probability of default for applicants with more than one child. The trend of SIZEHOUSEHOLD is the same as in model 1, bigger household reduce the probability of defaulting. ANNUALINCOME have been automatically dropped from most of the regressions due to missing values, this is the result of only accepting taxed income numbers in the application process instead of previous self reported. The BLANCO variable gives approximately the same results as in model 1, showing an increase for amounts of 150 001 – 200 000 SEK. CREDIT shows the same significant trend of lower default probability for applicants with previous credit cards, which is inconsistent with theory. The KALP variable now exhibits no clear probability increase for any excess level, the KALP variable’s minimum threshold have been raised by 2 000 SEK as of the implemented policy changes. MORTGAGELOAN still shows a slight decrease in default probability for applicants who own a home, but most of the effect has disappeared with the policy changes. SCORE is still the strongest coefficient among them all, yet again proving its importance, here the SCORE is based on *Upplysningscentralen* score plus a combinations of additional risk factors according to the specified policy changes (see table 5). For applicants with a COSIGNER continues to show lower tendency of defaulting. The only remaining behavioral variable showing any kind of continues significance is the voluntary contact information

variable PHONE, which still shows a slight decrease in default probability when providing a phone number.

### 5.5.3 Policy performance

Looking at the table below of the potential risk factors and model 1 and 2's specific risk, the results of the implemented policy changes suggest that the previous identified risk factors have either been resolved to the limit of which they are no longer significantly represented in the population sample, or improved as to having less (more) of a marginal effect on the response variable.

**Table 16: Policy performance**

Potential risk factors	Model 1 risk	Policy change	Model 2 risk	Effect
AGE	x		x	<u>remaining</u>
GENDER	x			resolved
MAJORITYCITY				
NRCHILDREN			x	
SIZEHOUSEHOLD	x		x	<u>remaining</u>
ANNUALINCOME	x	x		resolved
BLANCO	x	x	x	improved
CREDIT	x		x	<u>remaining</u>
INCREASEDRISK				
KALP	x	x		resolved
MORTGAGELoAN	x		x	<u>remaining</u>
PAYPLAN				
SCORE	x	x	x	improved
COSIGNER	x		x	<u>remaining</u>
LOANSIZE	x			resolved
MONTHLY				
PHONE	x		x	<u>remaining</u>
REPAYMENTPERIOD				
YEARSUMPERINCOME				

#### 5.5.4 Remaining risk factors

In order to further evaluate the remaining risk factors from model 2, two new regressions were performed with regards to variables with high and low information values, model 3 and 4. Both the models provided approximately the same results as model 2, although with some exceptions where GENDER showed a high information value and gave significant results in model 3. By testing the validity of model 2's results by categorizing the variables according to a different test, I can present more of a robust trend among the risk factors and exclude that the pattern is not presented by chance. After evaluating the two new models with our benchmark model 2, I concluded that the most efficient and best fitted model is the benchmark model 2.

# Chapter 6

## Conclusion

### 6.1 Concluding remarks

The recent years' liberalization of consumer credit has allowed new banks and financial institutions to enter the market, many of which consumer credits might be unknown territory. With an increasing amount of loans and the efficiency of which they must be processed, the need for an accurate and sophisticated credit scoring model has never been more important. In this paper I have evaluated a recent implemented change in a banks credit scoring model, although the majority of the implemented policy changes appears to have resolved the intended risk factors others still remain. Two new models have been developed in effort to establish the robustness of the results of these risk factors, my concluding opinion is that the results given the population sample and the below discussed limitations are robust and true.

### 6.2 Limitations

As to most studies I have faced some limitations which may or may not question the validity of the results. Below follows a short description of the major limitations.

#### 6.2.1 Sample selection bias

A limitation in this study which gives rise to some level of sample selection bias in the dataset, is that the dataset only contains approved loan applications. To include a full set of rejected applications would have been ideal. One problem that, on the other hand, arises with reject data included is the inference of rejection, e.g. it is not possible to predict the outcome if the rejected loan would have been accepted.

#### 6.2.2 Model simulation

One issue of using a dataset prior to new changes have been implemented, and use this to simulate these new implications is that previous rejected applications may now have been

approved. This problem affects both the policy simulation and creates a somewhat biased simulation sample. This could have been avoided by using datasets both prior and post the policy implementation.

### 6.2.3 Sample size

The size of my sample may give rise to some concern, especially the ratio between defaulted and non-defaulted loans. Although the sample is small, it has been tested in equivalent sample combinations ( $K=1,2,\dots,all$ ) without any major difference in outcome, this put some confidence in my discoveries.

## 6.3 Managerial suggestions

In light of the findings in this analysis I would recommend the following addition to the existing credit policy:

- i. Increased marginal risk for applicants with age of 18-29.
- ii. Reduced marginal risk for applicants who lives in households of more than two.
- iii. Increased marginal risk for applicants who do not own a home.
- iv. Develop voluntary contact information further and add marginal risk for those who do not supply information.

As to the focus of this paper these are the suggestions I make. Since this has been a isolated study of a limited dataset from which a limited number of variables could be extracted, there may be other underlying risk factors not taken into consideration in this analysis. Thus, I would suggest further studies on this subject, although that is outside the scope of this paper.



# References

## Literature

Altman, E.I., Saunders, A. (1998). *Credit risk measurement: Developments over the last 20 years*. Journal of Banking and Finance, Vol. 21, p. 1721-1742.

Anderson, D.R., Burnham, K.P. (2002). *Avoiding Pitfalls When Using Information-Theoretic Methods*. The Journal of Wildlife Management, Vol. 66, No. 3, pp. 912-918.

Anderson, R. (2007). *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*.

Baltagi, B.H. (2013). *Econometric Analysis of Panel Data*. 5<sup>th</sup> Edition. John Wiley & Sons Publication.

Bianco, K.M. (2008). *The Subprime Lending Crisis: Causes and Effects of the Mortgage Meltdown*, CCH Mortgage Compliance Guide and Bank Digest.

Boyes, W. J., Hoffman, D. L. & Low, S. A. (2002). *An econometric analysis of the bank credit scoring problem*. Journal of Econometrics, Vol. 40, Issue 1, p. 3-14.

Cramer, J.S. (2004). *Scoring bank loans that may go wrong: a case study*, Statistica Neerlandica Vol. 58, nr. 3.

Esmeralda, A.R., Joaquim, J.S.R. (2012). *Alternative versions of the RESET test for binary response index models: a comparative study*.

Hand, D. J. (1981) *Discrimination and Classification*. Chichester: Wiley.

Henley, W. E. (1995) *Statistical aspects of credit scoring*. PhD Thesis. The Open University, Milton Keynes.

Henley, W. E. and Hand, D. J. (1996) A k-nearest-neighbour classifier for assessing consumer credit risk. *Statistician*, 45, p. 77-95.

Henley, W. E. and Hand, D. J. (1997). *Statistical Classification Methods in Consumer Credit Scoring: a Review*, Journal of the Royal Statistical Society.

Howell, D.C. (2016). *Chi-square test - analysis of contingency tables*. University of Vermont, p. 1-4.

Menard, S. (2002). *Applied Logistic Regression Analysis*. 2<sup>nd</sup> Edition, Sage University Paper of Bank Research 2 (1): 31-37.

Orgler, Y. E. (1971). *Evaluation of Bank Consumer Loans with Credit Scoring Models*. Journal

Orgler, Y.E. (1970). *A Credit Scoring Model for Commercial Loans*. Journal of Money, Credit & Banking (Ohio State University Press) 2.4.

Straka, J. W. (2000). *A Shift in the Mortgage Landscape: The 1990s Move to Automated Credit Evaluations*. Journal of Housing Research, Vol. 11, Issue 2, p. 207-232.

Verbeek, M. (2012). *A Guide to Modern Econometrics*. 4<sup>th</sup> Edition, John Wiley & Sons Publication.

Winstra, J. and Ölcer, D. (2014). *Hur skuldsatta är de svenska hushållen*, Ekonomiska kommentarer – Riksbanken.

## Internet

Bank for International Settlements (2016), *Basel III: international regulatory framework for banks*. Available at: <http://www.bis.org/bcbs/basel3.htm?m=3%7C14%7C572>

Edwards, J., Vine, P. Gray, K. (2010): *An introduction to Basel III - its consequences for lending*, Norton Rose Fulbright. Available at: <http://www.nortonrosefulbright.com/knowledge/publications/31077/an-introduction-to-basel-iii-its-consequences-for-lending>

Idre – UCLA (2016). *The likelihood ratio test*. Available at: [http://www.ats.ucla.edu/stat/mult\\_pkg/faq/general/nested\\_tests.htm](http://www.ats.ucla.edu/stat/mult_pkg/faq/general/nested_tests.htm)

Konsumenternas. (2016). *Om Bolån*. Available at: <http://www.konsumenternas.se/lana/olika-lan/om-bolan>

Upadhyay, R. (2016). *Information Value (IV) and Weight of Evidence (WOE) – A Case Study from Banking*. Available at: <http://ucanalytics.com/blogs/information-value-and-weight-of-evidencebanking-case/>

Upplysningscentralen (2016), *UC Risk*. Available at: <https://www.uc.se/foretag/tjanster/kreditinformation/kreditupplysningar/uc-risk.html>

## Databases

Proprietary database from a Swedish bank.

Proprietary database from a Swedish credit reporting agency, *UC*.

# Appendix

## 8.1 Variables

<b>Variable</b>	<b>Definition</b>
AGE	categorical, age of applicant
ANNUALINCOME	categorical, self-reported annual income of applicant (in SEK)
BLANCO	categorical, previous private loan-amount (in SEK)
COSIGNER	dummy, takes value 1 if two applicants
CREDIT	dummy, takes value 1 if applicant has one or more creditcards
DEFAULTED	dummy, takes value 1 if applicant has defaulted the loan
GENDER	dummy, takes value 1 if applicant is a man
INCREASEDRISK	dummy, takes value 1 if applicant would not pass new application by 160101
KALP	categorical, applicants monthly income excess minus budget (in SEK)
LOANSIZE	categorical, the amount of loan (in SEK)
MAJORITYCITY	dummy, takes value 1 if applicant lives in one of the 3 major cities
MONTHLY	categorical, the amount the applicant repays every month (in SEK)
MORTGAGELoAN	dummy, takes value 1 if the applicant has a mortgage loan
NRCHILDREN	categorical, nr of children in household
PAYPLAN	categorical, previous repayment plans amount (in SEK)
PHONE	dummy, takes value 1 if applicant has provided a phone number
REPAYPERIOD	categorical, number of years to repay the loan
SCORE	categorical, the score the applicant has received by the lender
SIZEHOUSEHOLD	categorical, nr of persons in the household
YEARSUMPERINCOME	categorical, the early repayment amount compared to yearly income

## 8.2 Descriptive statistics

Variable	Number of obs	Mean	Std. Dev.	Min	Max
AGE	7391	49.46	11.40	29	79
ANNUALINCOME	7394	359480.70	52927.81	150000	400000
BLANCO	7392	247280.80	91744.43	100000	350000
COSIGNER	542	0.21	0.41	0	1
CREDIT	7394	0.60	0.49	0	1
DEFAULTED	7394	0.01	0.11	0	1
GENDER	7389	0.70	0.46	0	1
INCREASEDRISK	7305	0.11	0.31	0	1
KALP	5761	6865.13	4707.60	1000	16000
LOANSIZE	7394	209223.70	96156.87	70000	370000
MAJORCITY	7305	0.03	0.18	0	1
MONTHLY	7394	1958.82	727.68	500	3000
MORTGAGELOAN	7394	0.81	0.39	0	1
NRCHILDREN	7394	0.76	0.90	0	7
PAYPLAN	7379	1278.29	1150.99	500	3000
PHONE	7394	0.71	0.45	0	1
REPAYPERIOD	7394	8.71	2.85	1	12
SCORE	7394	0.01	0.01	0.005	0.05
SIZEHOUSEHOLD	7394	2.03	1.09	1	9
YEARSUMPERINCOME	7394	0.06	0.04	0	0.93

# 8.3 Regressions

## Model 1

	<b>K=1</b>	P> z	Std. Err.	<b>K=2</b>	P> z	Std. Err.	<b>K=5</b>	P> z	Std. Err.	<b>K=20</b>	P> z	Std. Err.	<b>K=all</b>	P> z	Std. Err.
Defaulted	183			183			183			183			183		
Num bad loans	183			183			183			183			183		
Num good loans	183			366			915			3660			7394		
<b>Demographic</b>															
AGE															
29	0.039	0.239	0.687	0.012	0.239	0.79	0.533	0.239	0.023	0.171	0.239	0.04	0.063	0.239	0.002
39	0.102	0.344	0.108	-0.008	0.921	0.083	-0.012	0.776	0.043	-0.001	0.959	0.012	-0.002	0.788	0.006
49	-0.042	0.694	0.106	-0.160	0.035	0.076	-0.089	0.023	0.039	-0.022	0.040	0.011	-0.010	0.074	0.006
59	-0.232	0.017	0.097	-0.240	0.001	0.071	-0.121	0.001	0.038	-0.034	0.001	0.010	-0.019	0.000	0.005
69	-0.182	0.115	0.115	-0.128	0.165	0.092	-0.095	0.024	0.042	-0.023	0.048	0.012	-0.016	0.006	0.006
79	0.074	0.796	0.286	0.198	0.462	0.269	0.092	0.610	0.181	0.008	0.835	0.040	-0.007	0.675	0.016
GENDER	0.095	0.180	0.071	0.061	0.170	0.045	0.028	0.168	0.021	0.009	0.099	0.005	0.004	0.100	0.002
MAJORITY	omitted						0.398	0.362	0.073	-0.23	0.009	-0.001	-0.007	0.214	0.073
NRCHILDREN	0.033	0.760	0.108	0.095	0.178	0.071	0.046	0.168	0.033	0.013	0.174	0.009	0.002	0.670	0.016
SIZEHOUSEHOLD															
OLD															
2	-0.374	0.000	0.098	-0.267	0.003	0.091	-0.174	0.004	0.060	-0.052	0.014	0.021	-0.009	0.112	0.006
3	-0.421	0.009	0.161	-0.363	0.001	0.114	-0.206	0.004	0.073	-0.062	0.011	0.024	-0.015	0.017	0.006
4	-0.613	0.000	0.111	-0.443	0.000	0.091	-0.245	0.000	0.061	-0.071	0.001	0.022			
5	-0.443	0.166	0.320	-0.445	0.000	0.095	-0.240	0.000	0.069	-0.071	0.002	0.022			
<b>Financial</b>															
ANNUALINCOME															
OME															
200000	-0.553	0.000	0.023	0.032	0.109	0.020	0.053	0.370	0.059	0.009	0.336	0.009	0.000	0.001	0.000
250000	-0.336	0.003	0.114	0.070	0.078	0.040	0.073	0.359	0.080	0.009	0.336	0.009	0.000	0.001	0.000
300000	-0.096	0.191	0.073	-0.029	0.161	0.021	0.103	0.342	0.108	0.010	0.335	0.011	0.002	0.001	0.002
350000	-0.356	0.001	0.109	-0.091	0.214	0.073	0.067	0.362	0.073	0.004	0.338	0.004	0.000	0.001	0.000
400000	-0.335	0.003	0.115	-0.039	0.169	0.028	0.066	0.362	0.073	0.008	0.336	0.009	-0.001	0.001	0.001
BLANCO															
150000	-0.447	0.000	0.030	-0.052	0.000	0.010	0.018	0.000	0.004	-0.003	0.000	0.001	-0.002	0.001	0.001
200000	0.105	0.000	0.017	0.129	0.000	0.017	0.031	0.000	0.007	0.007	0.000	0.002	0.004	0.001	0.001
250000	-0.085	0.000	0.007	-0.006	0.000	0.001	0.022	0.000	0.005	0.004	0.000	0.001	0.002	0.001	0.001
300000	-0.205	0.000	0.007	-0.032	0.000	0.006	-0.026	0.000	0.006	-0.008	0.000	0.002	-0.001	0.001	0.000
350000	-0.249	0.000	0.004	-0.028	0.000	0.005	-0.023	0.000	0.006	-0.009	0.000	0.002	-0.002	0.001	0.001
CREDIT	-0.255	0.009	0.097	-0.136	0.010	0.053	-0.064	0.002	0.021	-0.017	0.000	0.005	-0.006	0.002	0.002
INCREASEDRISK							0.398	0.197	0.309	-0.23	0.584	0.349	-0.007	0.374	2.703
KALP															
4000	-0.073	0.567	0.128	-0.123	0.100	0.075	-0.039	0.179	0.029	-0.012	0.075	0.007	-0.002	0.469	0.002
7000	-0.239	0.058	0.126	-0.105	0.205	0.083	-0.030	0.355	0.032	-0.010	0.157	0.007	0.001	0.759	0.003
10000	-0.016	0.920	0.159	-0.111	0.254	0.098	-0.042	0.247	0.036	-0.009	0.291	0.009	0.001	0.704	0.004
13000	0.182	0.293	0.173	-0.186	0.056	0.098	-0.054	0.191	0.041	-0.016	0.066	0.009	-0.002	0.576	0.004
16000	-0.061	0.732	0.179	-0.160	0.109	0.100	-0.036	0.391	0.042	-0.014	0.108	0.009	0.002	0.686	0.004
MORTGAGELAN	-0.403	0.000	0.101	-0.178	0.001	0.056	-0.082	0.000	0.021	-0.018	0.000	0.005	-0.006	0.002	0.002
OAN															
PAYPLAN	0.000	0.004	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.000	0.004	0.000

SCORE	22.277	0.000	3.631	13.480	0.000	1.978	5.251	0.000	0.714	1.209	0.000	0.167	0.485	0.000	0.074
<b>Behavioral</b>															
COSIGNER							-0.488	0.000	0.137	-0.331	0.000	0.083	-0.177	0.000	0.048
LOANSIZE															
130000	0.343	0.015	0.141	0.000	0.996	0.004	-0.046	0.003	0.015	0.142	0.023	0.062	0.046	0.050	0.024
190000				-0.007	0.996	1.404	-0.239	0.000	0.039	0.142	0.023	0.063	0.056	0.047	0.028
250000				-0.005	0.996	0.942	0.060	0.022	0.026	0.387	0.000	0.105	0.124	0.029	0.057
310000							-0.353	0.000	0.020	0.178	0.016	0.074	0.026	0.056	0.014
370000				-0.981	0.802	3.905	-0.437	0.000	0.008	0.390	0.000	0.105	0.059	0.046	0.030
MONTHLY															
1000							0.064	0.786	0.237	-0.127	0.564	0.220	0.007	0.931	0.084
1500							0.398	0.197	0.309	-0.230	0.398	0.272	-0.007	0.950	0.109
2000							0.301	0.414	0.368	-0.213	0.487	0.307	0.035	0.792	0.133
2500							0.615	0.070	0.339	-0.191	0.584	0.349	0.069	0.675	0.165
3000							0.622	0.073	0.347	-0.143	0.726	0.409	0.039	0.836	0.191
PHONE				-0.087	0.992	9.200	-0.428	0.000	0.102	-0.321	0.000	0.057	-0.184	0.000	0.030
REPAYPERIO	0.082	0.611	0.161	0.009	0.992	0.908	0.028	0.415	0.035	-0.008	0.711	0.022	-0.002	0.820	0.011
D															
YEARSUMPE	7.091	0.661	16.171	0.350	0.992	37.003	-2.404	0.374	2.703	0.677	0.679	1.634	0.291	0.700	0.755
RINCOME															

**Model 2**

	<b>K=1</b>	<b>Std Err</b>	<b>P&gt; z </b>	<b>K=2</b>	<b>Std Err</b>	<b>P&gt; z </b>	<b>K=5</b>	<b>Std Err</b>	<b>P&gt; z </b>	<b>K=20</b>	<b>Std Err</b>	<b>P&gt; z </b>	<b>K=all</b>	<b>Std Err</b>	<b>P&gt; z </b>
DEFAULTED															
Num bad loans	50			50			50			50			50		
Num good loans	50			100			250			1000			4963		
<b>Demographic</b>															
<b>AGE</b>															
29	0.039	0.239	0.687	0.012	0.239	0.79	0.533	0.239	0.023	0.171	0.239	0.04	0.063	0.239	0.002
39	0.001	0.251	0.996	0.257	0.155	0.099	0.058	0.088	0.513	0.010	0.026	0.695	0.004	0.005	0.449
49	-0.150	0.239	0.531	0.087	0.131	0.505	0.024	0.081	0.761	-0.002	0.024	0.936	0.001	0.005	0.911
59	-0.127	0.244	0.602	0.019	0.136	0.886	-0.060	0.076	0.434	-0.020	0.023	0.376	-0.004	0.004	0.371
69	-0.182	0.284	0.523	-0.077	0.141	0.585	-0.058	0.087	0.506	-0.021	0.025	0.395	-0.004	0.005	0.393
79															
GENDER	0.047	0.136	0.730	0.039	0.098	0.687	0.012	0.048	0.790	0.015	0.012	0.208	0.002	0.002	0.333
MAJORITYCITY	0.091	0.108	0.726	0.100	0.170	0.447	-0.006	0.185	0.919	-0.005	0.025	0.666	0.000	0.047	0.969
NRCHILDREN	0.533	0.234	0.023	0.400	0.169	0.018	0.171	0.084	0.040	0.063	0.020	0.002	0.012	0.004	0.002
<b>SIZEHOUSEHOLD</b>															
2	-0.448	0.167	0.007	-0.567	0.152	0.000	-0.319	0.159	0.044	-0.170	0.086	0.048	-0.043	0.024	0.077
3	-0.719	0.195	0.000	-0.707	0.170	0.000	-0.369	0.185	0.047	-0.186	0.087	0.033	-0.046	0.025	0.059
4	-0.827	0.108	0.000	-0.770	0.121	0.000	-0.426	0.144	0.003	-0.193	0.082	0.018	-0.048	0.023	0.041
5															
<b>Financial</b>															
<b>ANNUALINCOME</b>															
200000															
250000															
300000	0.317	2.233	0.887	0.140	0.273	0.608									
350000	0.738	1.801	0.682	0.336	0.493	0.495									
400000	0.524	2.505	0.834	0.255	0.425	0.549									
<b>BLANCO</b>															
150000	-0.684	0.043	0.000	-0.355	0.123	0.004	-0.259	0.115	0.024	-0.047	0.020	0.022	-0.006	0.003	0.019
200000	0.062	0.045	0.168	-0.045	0.006	0.000	-0.038	0.012	0.001	-0.004	0.002	0.016	0.003	0.001	0.018
250000	-0.284	0.102	0.005	-0.026	0.003	0.000	-0.149	0.056	0.008	-0.029	0.012	0.019	-0.001	0.001	0.018
300000	-0.348	0.102	0.001	-0.134	0.027	0.000	-0.164	0.063	0.009	-0.030	0.013	0.019	-0.002	0.001	0.018
350000	-0.465	0.078	0.000	-0.224	0.059	0.000	-0.215	0.090	0.016	-0.041	0.018	0.021	-0.003	0.001	0.018
CREDIT	-0.819	0.289	0.005	-0.311	0.126	0.013	-0.099	0.044	0.025	-0.031	0.010	0.001	-0.006	0.002	0.006
<b>INCREASEDRISK</b>															
KALP	0.017		0.393	0.038		0.371	0.045		0.902	0.035		0.673	-0.038		0.738
4000	0.000	0.248	0.998	-0.190	0.134	0.674	-0.016	0.042	0.837	-0.063	0.126	0.312	-0.004	0.003	0.635
7000	0.418	0.254	0.100	0.153	0.109	0.163	0.007	0.051	0.890	-0.002	0.010	0.812	0.000	0.002	0.875
10000	0.390	0.239	0.103	0.510	0.167	0.002	0.028	0.061	0.648	0.012	0.015	0.435	0.001	0.003	0.686
13000	0.448	0.285	0.116	0.459	0.230	0.046	0.024	0.075	0.750	-0.006	0.013	0.654	-0.001	0.003	0.625
16000	0.091	0.260	0.726	0.100	0.131	0.447	-0.006	0.058	0.919	-0.005	0.012	0.666	0.000	0.003	0.969
MORTGAGELOAN	-0.314	0.248	0.206	-0.147	0.134	0.273	-0.084	0.046	0.068	-0.008	0.009	0.391	-0.004	0.002	0.073
PAYPLAN	0.000	0.000	0.021	0.000	0.000	0.003	0.000	0.000	0.008	0.000	0.000	0.013	0.000	0.000	0.057
SCORE	47.050	11.736	0.000	22.747	4.940	0.000	10.177	2.278	0.000	2.137	0.443	0.000	0.420	0.086	0.000
<b>Behavioral</b>															
COSIGNER	-0.781	0.234	0.001	-0.438	0.120	0.000	-0.175	0.057	0.002	-0.042	0.018	0.018	-0.007	0.005	0.125
<b>LOANSIZE</b>															
130000	0.158	0.107	0.140	0.159	0.186	0.393	0.038	0.042	0.371	0.013	0.018	0.463	0.002	0.003	0.451
190000	-0.160	0.002	0.000	0.161	0.188	0.392	-0.026	0.032	0.409	0.016	0.022	0.462	0.002	0.003	0.451
250000	-0.406	0.207	0.049	0.131	0.159	0.411	-0.017	0.020	0.403	0.020	0.027	0.460	0.003	0.003	0.451

310000	-0.523	0.395	0.185	0.189	0.212	0.373	-0.038	0.047	0.416	0.017	0.023	0.461	0.002	0.003	0.451
370000	-0.488	0.332	0.142	0.133	0.161	0.409	-0.058	0.074	0.427	0.009	0.012	0.465	0.002	0.002	0.451
MONTHLY															
1000	0.086	0.154	0.578	0.045	0.369	0.902	0.035	0.082	0.673	-0.038	0.113	0.738	-0.004	0.013	0.777
1500	0.000	0.180	0.998	-0.190	0.467	0.684	-0.019	0.092	0.837	-0.063	0.125	0.612	-0.007	0.015	0.635
2000	0.309	0.249	0.214	-0.171	0.501	0.732	0.033	0.102	0.749	-0.059	0.126	0.642	-0.006	0.015	0.694
2500	0.531	0.356	0.136	-0.174	0.526	0.741	0.042	0.116	0.715	-0.060	0.128	0.642	-0.006	0.015	0.685
3000	0.666	0.408	0.103	-0.061	0.637	0.924	0.110	0.180	0.541	-0.056	0.131	0.668	-0.005	0.015	0.729
PHONE	-0.063	0.136	0.644	-0.130	0.085	0.126	-0.080	0.034	0.020	-0.015	0.009	0.080	-0.002	0.002	0.178
REPAYPERIOD	0.109	0.054	0.042	0.015	0.030	0.606	0.009	0.011	0.422	0.000	0.001	0.728	0.000	0.000	0.921
YEARSUMPERINCOME	1.801	4.783	0.706	0.691	2.263	0.760	-0.576	0.852	0.499	-0.065	0.153	0.671	-0.016	0.020	0.418

<b>Model 3</b>	<b>dy/dx</b>		<b>Model 4</b>	<b>dy/dx</b>	
SCORE	0.3614	***	LOANSIZE		
MORTGAGELOAN	-0.0052	**	130000	0.0036	
AGE			190000	0.0013	
29			250000	0.0024	*
39	0.0052	*	310000	0.0023	
49	0.0024		370000	0.0024	
59	-0.0017		COSIGNER	-0.0096	***
69	-0.0022		ANNUALINCOME	0.0000	
79			CREDIT	-0.0023	*
PAYPLAN	0.0000	***	PHONE	-0.0032	**
GENDER	0.0038	*			
SIZEHOUSEHOLD					
2	-0.0041	*			
3	-0.0021				
4	-0.0075	***			
BLANCO					
150000	-0.0057	**			
200000	0.0020	**			
250000	-0.0010	**			
300000	-0.0024	**			
350000	-0.0025	**			

## 8.4 Information value

Model 1 Sample		Model 2 Sample	
Variable	Information Value	Variable	Information Value
SCORE	1.121899	SCORE	0.921293
PAYPLAN	0.330926	MORTGAGELOAN	0.382071
ANNUALINCOME	0.324082	AGE	0.344872
AGE	0.225491	PAYPLAN	0.237877
COSIGNER	0.169369	GENDER	0.169151
LOANSIZE	0.149894	SIZEHOUSEHOLD	0.154272
BLANCO	0.116913	BLANCO	0.148848
KALP	0.067547	LOANSIZE	0.087307
MORTGAGELOAN	0.067263	COSIGNER	0.025881
GENDER	0.054121	ANNUALINCOME	0.016774
PHONE	0.047173	CREDIT	0.015844
CREDIT	0.028718	PHONE	0.014925
SIZEHOUSEHOLD	0.015755	MONTHLY	0.007973
NRCHILDREN	0.009004	NRCHILDREN	0.007448
MONTHLY	0.001715	KALP	0.000814
MAJORITY	0.000000	YEARSUMPERINCOME	0.000000
INCREASEDRISK	0.000000	REPAYPERIOD	0.000000
REPAYPERIOD	0.000000	MAJORITY	0.000000
YEARSUMPERINCOME	0.000000	INCREASEDRISK	0.000000

## 8.5 Selection tests

	<b>AIC value</b>	<b>BIC value</b>
Model 2	354.946	582.985
Model 3	462.737	568.332
Model 4	409.143	474.131

	<b>LR chi2</b>	<b>Prob &gt; chi2</b>	<b>df</b>	<b>Obs</b>
Model 3	149.790	0.000	21.000	4851
Model 4	110.200	0.000	28.000	4830

	<b>Pearson Chi-Square</b>	<b>Prob &gt; chi2</b>	<b>df</b>
Model 2	92.94	0.000	37
Model 3	91.940	0.000	16
Model 4	26.650	0.000	9

