UNIVERSITY OF NAIROBI

COLLEGE OF BIOLOGICAL AND PHYSICAL SCIENCES

SCHOOL OF MATHEMATICS

IMPROVED COX HAZARD MODEL FOR LOANS AND ADVANCES IMPAIRMENT

GEORGE KAMAU MBUTHIA

REG NO. I56/76652/2014

A dissertation submitted to the school of mathematics in partial fulfillment for a degree of Master of Science in Biometry

July, 2016

Declaration

Declaration by the Candidate

I hereby declare that this is my original work towards the award of MSC. Biometry degree and contains no materials previously published by another person and has not been presented for a degree in any other University except where due acknowledgement has been made in text.

GEORGE KAMAU MBUTHIA
REG. NO: I56/76652/2014

_____ _____
Sign Date

Declaration by the Supervisors

This dissertation has been submitted for examination with my approval as the university supervisor.

Prof. Moses M. Manene
School of Mathematics
University of Nairobi
P.O BOX 30197-00100

Dr.Ivivi Mwaniki
School of Mathematics
University of Nairobi
P.O BOX 30197-00100

_____ _____ _____ _____
Sign Date Sign Date

Abstract

In the report by CBK in March 2014, Non-Performing Loans (NPLs) had increased by 34.4% to KES 77.3 billion in June 2013 raising the percentage of the gross NPLs from 4.5% to 5.3% over the 2012/2013 period. The gross inflation characterised by an increase in prices of goods and services had also increased attributed to bad economic times.

The task of this thesis is to design an improved cox proportional hazard (ph) model to analyse the loans default by customers and thereby reduce the NPLs in order to maximize the net returns on loans. The objectives of the project are to determine the survival time of loans, assess if the survival time differs by the loan category, study the influence of predictors on survival of a loan and determine to what extent a cox ph model can aid in prediction of the loans default prediction by improving it. Various residual plots are also fitted to test for the goodness-of-fit, to identify possible outliers and influential observations in the models and check if the assumptions of the model hold. Such residuals are namely Schoenfeld, Martingale which is an improvement to the cox snell residuals, deviance, and score residuals and we applied those methods on loans data.The aim is to model the loan portfolio for the sampled bank with empirical data on customer credit information and compare the cox ph model versus the improved cox ph model to demonstrate how the cox ph model can be adjusted to fit the various management needs. The analyses has been implemented using the Excel and the R-Graphical User Interface software.

The results show that account balance and loan classification are highly significant in the improved cox ph model than credit amount and value saving stock in determining the default rates of loans. The average survival time for a loan is 16 months for the improved cox proportional model. The improved cox proportional model has an AIC value of 3325.881 and thus a better fit to the cox proportional model with an AIC value of 3343.359.

Keywords: Cox proportional hazard model, improved cox proportional hazard model, survival time, residuals.

Contents

List of Figures

List of Tables

List of Abbreviations

AIC
Akaike Information Criteria.

CBK
Central Bank of Kenya.

CI
Confidence Interval.

CRB
Credit Reference Bureau

HR
Hazard Ratio.

IRB
Internal Rating Based.

LGD
Loss Given Default.

NPL
Non-Performing Loan.

PD
Probability of Default.

PH
Proportional Hazard.

R-GUI
R-Graphical User Interface.

SBC
Schwarzs Bayesian Criterion/ Bayesian Information Criterion

SDF Survival Distribution Function

SME
Small and Medium Enterprise.

Dedication

This work is dedicated to my dear mother Madam Lydiah Gathoni, my lovely wife Silpha Waithira and my dear sons.

Acknowledgement

I wish to express first and foremost my profound gratitude to My Creator, Yahweh, for the life given to me, wisdom, knowledge and strength during my academic life especially during the period of this course.

I am grateful to my employer, my siblings and my entire family for their encouragement, concern and support.

I express my sincere gratitude to my supervisors, Prof. Moses M. Manene, Head, Statistics department, School of Mathematics, UON, for his guidance, direction and suggestions and Dr. Ivivi Mwaniki, Lecturer, School of Mathematics, UON for his guidance, optimum support and positive criticism during the course of this work. I also acknowledge the lecturers who taught me.

I thank everybody who contributed in one way or the other to the success of my thesis especially Dr. J. Nderitu, Lecturer, School of Mathematics, UON.

Chapter 1

Introduction

1.0 Banks

A bank is a money related foundation that makes credit by loaning foremost add up to a borrower and recouping the loaned principal with a charge. This lending can be done directly or indirectly through capital markets. There is Credit risk associated with this lending. This credit risk is the risk of loss of the lent capital and fee due to a account holder's non-payment or other credit extension, for example, early repayments, (Wikipedia.org, as of March 2009).

Credit suffers credit hazard when the indebted person can't meet their legitimate commitment according to the advance contract. Some of the defaults can occur in loans, mortgages and other financial financing. Globally, all banks are subject to minimum capital requirements as required by the Basel Accord (III) in order to operate smoothly. The banks in Kenya are regulated by central bank and must reserve a deposit with the central bank under which banks hold liquid resources equivalent to just a bit of their present liabilities. Thus as banks generate majority of its revenue through lending, it is important to reduce the risk associated with lending of money which is failure to repay the lent money and the lent fee as a result of banks lending to poor payers or borrowers who cannot fulfil their obligations. This risk can be reduced by spreading the risk, lending a reasonable amount to individual borrowers and being able to choose an investment opportunity with a high return for the principal lent.

1.1 Background

Since nineties, banks have been developing credit risk models as a means to measure the potential loss that a portfolio of credit exposures might suffer within a certain time. In Kenya, the central bank plays a major role in lending as it issues the lending rates which are aimed to benchmark what the banks will use to lend to its customers. Credit reference bureau (CRB) supplement the focal pretended by banks and other budgetary establishments in developing money related administrations inside an economy. CRB according to the central bank website help loan lenders settle on speedier and exact choices. They gather, oversee and share client data to banks within a given administrative structure defined in 2008 and operationalised on second February, 2009. Records on credit form the basis for advances underwriting and also allow borrowers to assume their acknowledgment history starting with one money lender then onto the next bank, in this way making loaning markets more engaged and focused. The CRB oversee risk and extortion through sharing of data between cash related establishments in appreciation to their client conduct and this positively affects the economy.

In the eighties and nineties, Kenyan financial lending sector was saddled with a large non-performing loanss (NPLs) portfolio. This prompted the breakdown of a few banks. In the latest years, other banks have collapsed due to poor profiling of their customers and even mismanagement. One of the major cause is the serial defaulters who borrow from various banks and do not repay. The management is also lax on playing its oversight role and is not usually thorough in performing a background check for its potential borrowers. Even with the CRB in place, some banks have collapsed and some have a high number of non performing loans. This then implies that each individual bank should develop a properly working system of evaluating the borrower to supplement the already functional CRB. Globally, according to the National Bureau of Economic Research,the late 2000's financial recession that influenced the major enormous economies was viewed as the longest monetary downturn since the 1930 great recession (NBER, 2010). This financial recession which had a negative impact on the economy especially on 2007 to 2008 was characterised by high levels of unemployment, declining real estate value, liquidations and foreclosures affected the financial institutions and prompted massive bank failures especially in the United States.

It led to collapse of 25 banks in 2000 and another 140 banks closed in 2009. The effect of this recession is still felt in the United States although the economy has improved since then. Banks continue to close with 157 banks closing in 2010. This are the highest closures since the collapse of savings and loans in 1992. In 2011, 92 banks closed and was followed by closure of other 51 in 2012.

China economy growth and devaluation of the Chinese Yen have also had an impact on developing countries which are heavy importers. Thus the financial services sector in any part of the world plays a major role in guaranteeing shareholders of a decent return in their equity value, assures depositors of security for their savings and easy access of funds to borrowers. Lack of this stability would results in high lending rates hence discouraging borrowers and a negative impact on the economy as a whole or part of it.

Kenyan economy is based on significant columns, for example, Tourism, Agriculture, Livestock and Fisheries, Wholesale and retail exchange, Manufacturing, Information Technology empowered administrations (already known as business procedure off-shoring), Financial administrations and Oil also, Gas, (Kenya Vision 2030). The banking sector supports all the others by financing them and enabling a timely and efficient transactions of business. Thus it is important to gather a deeper understanding of the financial markets specifically the banking industry and understand what causes the bank failures. In the recent year, 2015, the banking industry suffered a major blow with Dubai and Imperial banks being closed and recently Chase bank due to the large portfolio of insider lending was placed under receivership. This led to investors losing their savings while others have to wait for a long time before they can recover their savings. Thus there is need think of a more viable solution notwithstanding the regulations being implemented by central bank at the branch level to decrease the current financial crises and shield a repeat of the failures seen some time recently.

The ability of a bank to predict and detect an early warnings of failure is essential for the survival of the banking industry. Thus in this project, the bank data have the customers characterized using the readily available data with the variables which will be tested on their impact on loan default. The application of financial modelling have proved useful in predicting and determining the credit worthiness of a borrower. Studies of time to event outcomes have gotten to be basic in many areas in scientific research. The term survival analysis has been used in broad sense for analysis involving time to event of a certain

occasion. Survival data is used in such experiments. The event may be occurrence of a certain disease, time till school dropout, development of cancer in cigarette smoker, and others. Application of survival models have mainly been used in the biomedical reliability research and has also been extended to other fields. This has led to its development. In the engineering sciences it has been used in the study of machines and their process failure as failure time analysis.

1.2 Statement of the Problem

Bank failures can essentially influence the economy and the client or investor certainty and confidence. There are many factors which make it hard to have a banking system without failure. With a reliable statistical model, there can be a significant reduction of failure in the banking sector whether caused by incompetent management or loopholes in the policies which increases the failure rate. These analysis would help lenders to maintain an accurate and effective lending database which would in turn result into a more successful framework.

Capacity to foresee the bank failures as a result of default by having the appropriate statistical model to predict is an important step. Historical data is mostly used for prediction of failure times. However, the most suitable data is not always readily available as the historical data is normally collected under many conditions which have a significant impact on the model. These covariates may have a statistically significant influence on the decision made.

There are several techniques applied in financial markets which include credit score models, Brownian model of financial markets, binomial option models, heat equation, black model, black-scholes option model , stochastic volatility, financial maths, monte-carlo option model, jump diffusion, real options analysis, logistic regression among others. However at one point these models have been criticised as they are not able to model all the factors in the financial data and have a margin of error in predicting default rates. For instance the black scholes model which is a first generation model uses differential equation and its major limitation is the normality assumption of the model and thus does not capture extreme movements. In addition, various researchers such as Espen Gaarder Hang and Nassim Nicholar Taleb, 2008, contend that black scholes fit just recasts existing broadly utilized models in terms of all intents and purposes unimaginable element sup-

porting rather than risk to make them more proficient with standard neoclassical economy hypothesis. Credit scoring have also been widely used to decide the credit value of a client. Credit scoring have also been criticised as it does not predict the probability of a loan default or the loan failure time.

Thus a survival model will be more appropriate to fill in this gap as has been used in mortality investigations and default rates on loans by various researchers. The cox hazard model is very flexible and can be adjusted to fit in the loans modelling. It will be able to capture the factors which influence default rates, indicate good and bad lenders, calculate the probability for a loan survival up to a specified duration and default rates on banks personal loans and the projection of default rates.

1.3 Objectives

The general objectives are to build a risk prediction model for loans and advances using the cox regression model and study the factors which have impact on time it takes for a loan issued to default by using the influential factors.

The specific objectives are:

1. To decide the survival time of loans
2. To assess if the survival time differs by the loan category
3. To study the influence of predictors on survival
4. To determine to what extent a cox ph can aid in prediction of the loans default by improving it.

1.4 Justification

In the late times, there has been a positive development on the populace and the enthusiasm to get to financial services. Banks have also been popularised with banks having several branches to reach a wide customer base which have allowed small and medium enterprises (SMEs), institutional and non-institutional investors in addition to large companies to get to advance loan funds. The accessibility to these loans have grown gradually with an estimated banked population at 75% which is equivalent to 8 adults out of 10.

Kenya has the highest banked population above the global average of 62% especially due to the large uptake of the mobile banking. These population is both in the formal and informal sector. This means a huge responsibility for banks to make a profit margin while lending at a reasonable fee or interest rate and being able to compete effectively with other financial services providers.

Loans and advances impairments are a major cause of banks failure. In the recent years many banks have closed for many reasons leading to an increase in the unbanked and low subscriptions as most customers turn to Sacco and mobile money. The purpose of this project is to explore the effects of risk and factors on default or non-default of the loans in banking. To achieve this purpose, the cox proportional hazard regression model will be constructed for the advances data to compute the hazard ratio (HR) and the confidence intervals (CI) for these risk variables.

The results of this project will provide insights on the risk factors and the most influential covariate that have critical effect on credit and advances impairment and identify the default risk of loans to borrowers under those significant factors at different time period for various loan portfolio. It also seeks to equip regulators and investors with a model capable of predicting future failures.

The advantages of having survival models that estimates when clients default are:

1. The capacity to figure the profitability over a client's lifetime and perform profit scoring.
2. These models may give the bank an evaluation of the default levels after some time which is useful for debt provisioning.
3. The assessments may help choose the term of the credit.

1.5 Limitation

In this project, it is assumed that the audience have some basic knowledge on survival analysis. The model has been explained such that the reader have a great amount of understanding the model.

From the literature, survival analysis is defined as the time it takes for an event to happpen. This event for the loan case study is either default or non-default. However, in the sampled bank data, loans are classified as either normal, watch, substandard, loss or doubtful. This will therefore mean if a customer is classified as non-default, then the customer is normal and if classified as default, the customer is either watch, substandard, loss or doubtful. Our period will cover the time from granting the loan to a customer to the time the customer is deemed a defaulter or a non-defaulter, (Isik, Deniz and Taner, 2010). This is in line the standard practice.Analysis demonstrates that the default rate as a component of time the client has been with the bank develops at first and it is simply following twelve months that it begins to settle. A shorter period would be underestimating the default rate and not mirroring the full sorts of qualities that predict the default, (Thomas, 2000)

The study is conducted based on secondary data which might have incomplete and biased information.

This study is also based on baseline value of the variables of interest.

1.6 Research hypothesis

H_0: There is no significant relationship between the loan default and loan characteristics (credited amount, among others), customer demographic characteristics (residence, purpose sex and marital status, employment status among others) .

H_1: There is a significant relationship between the loan repayment and loan characteristics (credited amount, among others), customer demographic characteristics (residence, Purpose, sex and marital status, employment status among others).

Chapter 2

Literature review

2.1 Introduction

Banks play a major role in driving a country's economy. The main function of a bank is to lend money at a fee called principal fee in order to generate profit. However banks generate income from other areas which include insurance, share trading or investing in government bonds. The money used for these investments is mainly from customer deposits and profit from lending. The allocation of credit to the community is based on credit demand in the community according to the neoclassical financial hypothesis created in 1980. A credit demand increase in the community may inspire the loaning establishments to make this credit more available to the communities. Both individuals and firms seek for credit from the lending institutions. To qualify for a loan facility a firm or an individual must be creditworthy. For firms looking for an advance from a respectable financial institution, this will give them a mileage as firm clientele and public confidence will go up since it will demonstrate that the firm will probably stay in business.

2.2 History of Credit Risk Modelling

The research on default and credit risk modeling has been on a positive growth with the adoption of better models over the years. Survival analysis in the field of economics during the period 1979-1984 was pioneered by Lane et al (1986) who identified variables significant to the bank failure in the USA, Altmans (2002) study of business study of

default risk. Other researchers have also assessed different sorts of risk models on cross sectional information sets include Maria Stepanova (2002), Allen and Rose (2006) and Babajide Abiola et.al (2014). All of these authors have made a significant contribution in the study of personal loans default. Subjective analysis or banker expert system was solely depended on amid the eighties by most financial institutions to survey the credit risk on loans a method which was mostly biased. Borrower reputation is a major character which the borrower use. A borrower capacity to repay a loan is also compared against his recurring debts. The nature of the borrowers investment will determine the lenders confidence in that if the borrower is making a large contribution to the investment the lender will be willing to lend as the borrower is less likely to default. For the investment of a large magnitude, the bank will offer guarantee in terms of margin and hold a part of borrower's savings until they can repay their loan. For other borrowers, collateral such as property of valuable assets helps to secure the loan.

2.3 Broader Credit Risk Issue

2.3.1 Credit risk

Credit risk can be defined as the vulnerability identified in the borrower inability to reimburse the loan. Accurate profiling of customers by banks and ensuring proper classification of loans to improve the likelihoods that the loan have value to the borrower as well as the bank has been the main challenge by most lenders. There are various classification of borrowers such as corporate, business or personal. Loans are also classified and issued as either long term also called corporate loans, short term also called business loans and other individual loans such as advances, overdrafts, fixed or reducing balance or as either as a combination depending on the institution and also the borrower rating. The amount being lent will also determine the type of loan security to be considered. Financial institutions have moved from the subjective frameworks utilized as a part of the eighties to the more unbiased and deductively based frameworks to decrease the credit risk. Such frameworks are measurable credit scoring models which attempt to anticipate the probability that a credit existing borrower will default over a particular period. Other methods which have been widely used over the years include discriminant analysis and logistic regression for creating scoring systems.

Credit risk can also be termed as default risk with both having a universal meaning as the level of value variance in debt instruments and derivatives as an after effect of changes in the essential credit nature of borrowers. It is a significant concern area to all the loaning organizations and accordingly credit scoring is significant for banks in determining borrower's creditworthiness. In modern finance credit risk determining is one of leading topics with various contributions from various authors. Banks have also had to reveal their customers information on credit history. Risk is a situation with probable exposure to adversity or a situation where there is a probability of deviation of an outcome from the expected outcome.

2.3.2 Default

Default is defined as the state entered when one cannot repay the amount in debt after three consecutive months. This period is not considered by some other authors. These state can be entered even before three months in some instances when a debtor becomes incapacitated either due so sickness or their investments are damaged. This default happens when an individual or a firm does not make their reimbursements as planned or abuses the stipulated contract which infers that only one non repayment results in default. In Kenya the loan undergoes various classifications before it is considered as a default. In personal secured loans once an individual does not pay for three consecutive times the grantor will be prompted to pay on their behalf and if they are not able to pay the loan will be considered as default and then legal action taken to recover the loan. These might include the auction of these security items among others.

Numerous definitions for default exist with all attempting to hint at a solid sign of financial distress. The most common definition for default is a state when an individual or a firm is legally declared unable to pay its creditors and have a negative equity. This is accounting view of default. Due to the last decade crisis in financial borrowings, numerous organizations have failed with some of them prompting budgetary disintegration of their financila conditions. There has also been a decreased in new loan facility conceded by banks as they have reduced the allowed credit levels. An individual and or firms rating and their relationship with a bank will influence some of the variables considered in a

loan when a new facility is requested.

A major determinant for the cost of loan is the classification of an individual or a firm which gives information on the probability to repay. Thus the perception and a good relationship with a bank has a decisive role in the determining the conditions (terms of) of a loan such as the principal lent interest lent and the repayment period. In addition individuals or firms with positive results in their investments are more likely to benefit as the banks will have more certainty in their ability to repay. More literature have shown many inconsistencies in the impact of these relations in the cost of credit.

Bigger firms gather credit from most banks and have a reduced credit costing as a result of having many relations with banks. Smaller firm in turn will gather loans from different banks even if they will have a higher cost of credit. Diamond (1984) argues that individuals and firms will benefit more from a bank with a unique banking relationship. This idea is developed in the believe that a unique relation with a bank will reduce the monitoring costs and in turn lead to less cost of credit. With the increase in creditors this may increase the spread charged and decrease the credit supply. Many individuals and firms will have multiple banking relations. Maintaining these many banking relations is more likely to occur in countries with a weak judiciary or legal systems and weak creditors rights.

2.4 Credit Risk Mitigation

Spreading of risk is vital for lending institutions. The riskier an individual or a firm is, the more the interest a bank will charge. In addition, it will give a shorter repayment period to the borrower. In some instances the bank will require a guarantee. Some studies have shown banks charging a higher interest rates on loans with collaterals that is on secured loans than on the unsecured loans. This may be proven by the fact that secured loans will be held by more risky borrowers and may even be of higher amounts. On the other hand unsecured loans will be of low amounts and mainly given to borrowers with a good repayment reputation who will be less likely to default.

Collaterals or guarantees are vital in spreading risk. A bank can also offer collateral or act as a guarantee to its customer. In this case the bank will charge a commission to the client and hold a part of his deposit as security or margin to ensure the client deliver. These scenario occurs mainly in contracting where the banks want to support its customer to get a contract. When demand for loans exceeds supply at the current interest rate the collaterals become even more important as a way of mitigating information asymmetries and in turn resolving the credit risk problem.

Once an individual default they may enter into an agreement with the bank on how to repay their debt in future. Some may require additional funding in order to recover. Firms can also be impaired resulting to low profitability, reduced sales and investments and dependency on borrowed funds. When a debtor goes into bankruptcy this stress not only affects them but also the creditors and other relations (other stakeholders) to the debtor which may include the employees, landlords, customers and suppliers. When an economy has distress some of these debtors will not recover due distress affecting even their competitors and due to revision in the economic worth of firm's assets. The severity of default will determine the debtor's recovery and a more severe default will have a more probability to the firm or the individual being extinct.

Shareholders equity is not guaranteed once a firm or individual struggle financially. In Kenya all commercial banks submit a daily, weekly and monthly report to the central bank on their financial position on lending and deposits. The government through central bank acts as a regulator and takes into account the interests of other stakeholders. This is mainly achieved by the government central role in creating consistent request and intermittent endeavours to answer how to predict default and get early warnings on failure. There are also set policies which gives investors' confidence to put their money and borrow from banks.

2.5 Advancements in Credit Risk Modeling

Various literatures have emerged in the recent years with the great need to mitigate risk in credit portfolio leading to growth in credit risk modeling. This process tries to find the parameters or variables which are highly associated to loans default. Some of these parameters incorporate the likelihood of default on individual advances or pool of exchanges (PD), estimation of losses given default (LGD) and the correlation between defaults. In the Global accord concurred in June 2004 (BIS, 2004) budgetary establishments are welcomed in the interior rating based (IRB) way to deal with appraisal for the one year probability of default and the LGD. Artificial neuro networks credit scoring models and rating systems are some of the traditional models used to estimate pools of transactions (PD). The most common of these and widely used is the credit scoring model. These models have a limitation in their ability to predict banks failure. Some of the questions which these models have not been able to answer though having been widely used are;

1. The optimal method to evaluate a customer
2. The variables which are most accurate to include in assessing their application,
3. What is the most appropriate data and information to look out for in decision making,
4. The best measure to determine whether a customer will default and the right information to use in classifying a customer as good or bad.

In any case, with the quick development in the credit business an enhanced administration of enormous advances credit scoring is viewed as the most critical in the banks for deciding credit value of people and firms. Its use although limited can lead to reduction of cost of credit associated with bad loans and lead to faster decision making in issuing loans. Consequently it is essentially utilized by financial institutions to enhance the procedure of credit accumulation and examinations which incorporate the lessening of credit investigator's expense. Monitoring of the existing customers can help a faster credit decision making process. In Kenya little literature is available on the use of this model in financial modeling. Information on the extent of use of credit scoring practices by banks in Kenya is virtually nonexistent.

As banks offer countless applications for their customized productss. Giving an extensive variety of new product channels utilizations of this model to banking sector have developed in the most recent decades. Consumer credit has been a major product for banks in the recent time. These applications have included various bank products including consumer loans which is one of the generally utilized as a part of the field card scoring. Other bank products include mortgages and overdrafts. Majority of researchers have concentrated more on the existing consumer loans rather than on new consumer loans.

2.6 Survival Models

Survival models are models which perform 'time to event' analysis on information. They are able to take into account censored observations. These censored observations are observations with incomplete information or did not encounter the event of interest over the span of the study. Thus, each subject in the study must have a defined beginning, end of the observation period, a time variable and an indicator of whether it experienced the event of interest or not.

The subjects in the study experience the event in our case default at various times, that is they are not tied, if time is measured continuously. However, since time is measured discretely in most cases, the subjects might default at the same time. When the subjects default at various times, the exact method is used when fitting the survival model. If there are some subjects who default at the same then, the Efron method or Breslow method is used when fitting the survival models. In this case there are ties.

2.7 Improved Cox Hazard Model

Survival analysis have been used mainly in the medical field and thus no much literature is available on financial field. In the recent developments, other models have however been used such as credit scoring models, logistic models and even Bayesian logistic regression. This project will focus on the cox hazard regression model.

The suggested ways of improving the cox proportional hazard model are as follows.

1. To select a few characteristic variables in the loans data and coarse-classify them and compare the results to the fine classification.

2. Use diagnostics mainly the residuals in order to test the sufficiency of credit risk. These residuals will be applied in full model and in testing for the individual covariates.

Chapter 3

Research Methodology

3.1 Introduction

Under this chapter, the research methods applied and the data used.

3.2 Research Question

The goal of this research is to examine the effects of the borrower's characteristics in loans default, volume of the loan granted and the lending fee.

The information on credits and advances is separated into various classifications such as borrower characteristics, loan duration, age, gender, loan price, loan status among others.

Question 1: examine the time to loan default
Question 2: how borrowers' characteristics affect loan default
Question 3: how can be cox model be improved in the loans default modeling.

3.3 Research Model

As discussed in the chapter 1, the objective of this research is use cox proportional hazard model to examine the co-integration of the borrowers' characteristics and banks loan granting decision.

3.4 Survival Analysis

Survival analysis is a unit of statistics that deals with the examination and analysis of real time data. The outcome factor of interest is time until an event occurs. Survival data is always paired (t, d) where t is the survival time and d the survival indicator. Survival models are capable of incorporating censored data. Censored data means that the event of interest or in this case default does not take place in the sampled period. In this project, cox proportional hazard model is used which is a well-known survival analysis model.

3.4.1 Basic Definitions

Event: This is referred as failure. It could be death, recovery among others.

Time: This is the survival time. It can be the actual time or censored time.

Subject: In survival analysis, these are participants in the study. In this case, these are loans. The event of interest is loan default which by either early repayment or violation of loan contract on repayment.

Risk set: the risk set at time t is a set of all subjects (loans) at the risk of defaulting at time t. That is, all the loans which did not default before time t.

Censoring: is a case where we have information about an individual survival time but we do not know their actual failure time.

Survival analysis methods use this information in fitting a model. There are other type of censoring which we will not consider in this study such as:

1. Left censoring
2. Right truncation
3. Left truncation
4. Interval censoring

Reasons for censoring
End of study- the study can end before the failure.

Loss of follow up- this can result due to lack of follow up or emigration.

Competing risk- where the event of interest occurs due to other reasons other than what is of interest.

In the event that explanation behind censoring is end of study we can reason that the controlling system is autonomous of time.

3.5 Cox Proportional Hazard Model

The model is expressed as

$$h(t,x) = h_0(t)exp\big(\sum_{i=1}^{p}\beta_i x_i\big); i = 1,2,...,p$$

Where $h(t,x)$ is the hazard function at time t for a subject with covariate values x_1, x_p, also referred to as the response or dependent variable as it depends on time point t and vector of covariates x.
$h_0(t)$ is the baseline hazard function, that is, the hazard function when all covariates are equal to zero. exp is the exponential function $exp(x) = e^x, x_i$, is the i^{th} covariate in the model, and β_i is the regression coefficient for the i_{th} covariate, x_i.
$h_0(t)$ depends only on t, time but not x. The exponent is also independent of time.

The cox proportional model is famous due to:

1. It is robustness and capability to take censoring information.
2. The estimated hazard are always non-negative.
3. $h(t,x)$ and $s(t,x)$ where $S(t)$ is the probability of surviving beyond some time t, can be estimated for a cox model using the minimum assumptions

4. The model has no intercept and the covariates are used to predict the hazard function and not the response $h(t, x)$.

The cox proportional hazards model is mostly used in medical field to assess the survival of patients given some covariates.

This model extends the concept of Kaplan Meier stated below and was first proposed in 1972 by incorporating covariates in the analysis of failure times.

The concept of Kaplan Meier

To decide the survival function without representing qualities of a particular loan, Kaplan Meier (1958) estimator is used. Kaplan Meier estimator assume the same probability of a loan default during its lifetime without taking into account the loan characteristics other than age. The Kaplan Meier starts by sorting the failure times $t_1 < t_2 < t_3 < ... < t_n$.

The Kaplan Meier model is expressed as follows:

$$\lambda(t|x) = \lambda_0(t).exp(\beta^T)X$$

This model is non parametric as it involves unspecified function in the form of an arbitrary baseline.

3.5.1 Basic Notations

$S(t)$- Survival function
This is the probability of surviving beyond some specified time t.

$$\begin{aligned} S(t) &= Pr[T > t] \\ &= 1 - Pr(T \leq t) \\ &= 1 - F(t) \end{aligned}$$

Where $F(t)$ is cumulative distribution function.
When T is continuous random variable,

$$F(t) = \int_{-\infty}^{t} f(u)\ du$$

Where $F(t)$ is a probability density function.
Thus

$$\begin{aligned} S(t) &= 1 - \int_{-\infty}^{t} f(u)\ du \\ &= \int_{t}^{\infty} f(u)\ du \end{aligned}$$

taking the derivative on both sides of the equation we have,

$$\begin{aligned} \frac{d}{dt} \int_{-\infty}^{t} f(u)\ du &= \frac{d}{dt}(1 - S(t)) \\ f(t) &= -\frac{d}{dt} S(t) \end{aligned}$$

$S(t)$ focuses on situations where the event does not occur upto and including time $T = t$ It is also largest for lower values of T and reduces towards zero. This is the left continuous property of $S(t)$

$H(t)$- Hazard function

This is the instantaneous or momentary failure rate. It is characterised as:

$$h(t) = \lim_{\Delta t \to 0} Pr \frac{[t < X \leq t + \Delta t | T > t]}{\Delta t}$$

This can be shown as follows: let $T = X$, then

$$h(t) = \lim_{\Delta t \to 0} Pr \frac{[t < X \leq t + \Delta t | X > t]}{\Delta t}$$
$$h(t) = \lim_{\Delta t \to 0} Pr \frac{[t < X \leq t + \Delta t, X > t]}{Pr(X > t)\Delta t}$$

That is $(t < X \leq t + \Delta t) \in X > t$. $Pr(X > t)$ is the survival function $S(t)$

$$h(t) = \lim_{\Delta t \to 0} Pr \frac{(t < X \leq t + \Delta t)}{Pr(X > t)\Delta t} \tag{3.1}$$

From the survival function,

$$\begin{aligned} Pr(X > t) &= S(t) \\ &= 1 - Pr(X \leq t) \\ &= 1 - F(t) \end{aligned} \tag{3.2}$$

Thus from equation 3.2 above, equation 3.1 becomes

$$h(t) = \lim_{\Delta t \to 0} Pr \frac{(t < X \leq t + \Delta t) - Pr(x \leq t)}{S(t)\Delta t}$$
$$h(t) = \lim_{\Delta t \to 0} \frac{(t + \Delta t) - F(t)}{S(t)\Delta t}$$
$$h(t) = \frac{f(t)}{S(t)}$$
$$= \frac{f(t)}{1 - F(t)}$$

$F(t)$-Cumulative density function

$$\begin{aligned} F(t) &= P(T \leq t) \\ &= \int_0^t f(z)dz \end{aligned}$$

$\lambda(t)$-Cumulative Hazard function

$$\begin{aligned}\lambda(t) &= \int_0^t h(z)dz \\ &= \int_0^t \frac{f(z)}{1-F(z)}dz \\ &= -l_n(1-F(z)) \\ &= -l_n S(t)\end{aligned}$$

3.5.2 Interpretation of the model

For a unit increase in X_i, the hazard rate is multiplied by a factor $e^{(\beta_i)}$ That is, the covariates have a multiplicative effect with respect to the hazard rate.

Consider a unit increase in one covariate for one individual

$$\begin{aligned}h_1(t) &= h_0(t).e^{\beta x_1} \\ h_2(t) &= h_0(t).e^{\beta(x_1+1)} \\ \frac{h_2(t)}{h_1(t)} &= e^{\beta(x_1+1-x_1)} \\ &= e^{\beta}\end{aligned}$$

Taking the log on both sides we have

$$log\left\{\frac{h_2(t)}{h_1(t)}\right\} = \beta$$

Thus β is the log of hazard ratio for a unit increase in x_j. But when the covariates increases by 1 unit, the risk (hazard ratio) actually increases by e^{β} units.

3.6 Theory of lifetime data analysis

Let T be the random variable representing time until repayment of a loan ceases due to either default or early repayment. Then, T has the following distribution

$$h(t) = \lim_{\Delta t \to 0} Pr\left\{ \frac{[t \leq T \leq t + \Delta t | T > t]}{\Delta t} \right\}$$

That is, an individual payoff or default at time t, conditional to his or her having stayed on the loan book up to that time.

When we have more information on the individual, covariates x_i, we want to determine the relationship between the distribution of failure time and these covariates.

Cox (1972) suggested the following model

$$h(t, x) = e^{(x\beta)} h_0(t)$$

Where β is a vector of unknown parameters and h_0 is an unknown function giving the hazard for the standard set of conditions when $x = 0$. There is an assumption that an individual with characteristic x is proportional to some unknown baseline hazard.

Cox showed that one can estimate β without any knowledge of h_0 using rank of failure and censored times. The likelihood function is given by

$$L(\beta) = \prod_{i=1}^{k} \frac{e^{x_i \beta}}{\sum_{l \in R_{t_i}} e^{\beta^T x_i}} ...iii$$

For $t_1 < t_2 < ... < t_k$ where $t_1 < t_2 < ... < t_k$ are k ordered failure times and R_{t_i} is the set of individuals at risk at time, t_i.

3.7 Maximum Likelihood Estimator

Cox showed that one can obtain consistent and highly efficient estimators of β by maximising the partial likelihood function

$$L(\beta) = \prod_{j=1}^{r} \frac{e^{x_j \beta^T}}{\sum_{l \in R_{t_j}} e^{\beta^T x_j}}$$

Where only the actual death times count

$$h_i(t) = \frac{h_0(t)}{4} e^{\frac{\beta^T x}{4}}$$

Using the logistic regression expressed as

$$\begin{aligned} log(\frac{p}{1-p}) &= \beta^T x \\ \frac{p}{1-p} &= e^{[\beta^T x]} \\ p &= e^{[\beta^T x]} - p.e^{[\beta^T x]} \\ p &= \frac{e^{[\beta^T x]}}{1 + e^{[\beta^T x]}} \end{aligned}$$

Incorporating the censored times to make it complete we let

$$\partial_i = \begin{cases} 1 & if the i^{th} individual got the event \\ 0 & if censored \end{cases}$$

Where $i = 1, 2, ..., n$ is the number of individuals and $j = 1, 2, ..., r$ is the number of event times with $r \leq n$
Then

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{e^{x_i \beta^T}}{\sum_{l \in R_{t_i}} e^{\beta^T x_i}} \right\}^{\partial_i}$$

The term is raised to the power of zero for censored event times and thus censoring still does not play a significant role.

$$\begin{aligned} log(L(\beta)) &= L(\beta) \\ &= \sum_{i=1}^{n} \partial_i \left[\beta^T x_i - log(\sum_{l \in R_{t_i}} e^{\beta^T x_i}) \right] \end{aligned}$$

Minimising $L(\beta)$ requires solving with respect to β, $\frac{\partial L(\beta)}{\partial \beta} = 0$ which is complex and the Newton Raphson procedure is often used.

Since credit performance data are normally recorded only monthly so that several failures at one time can be observed, that is tied failure times, and the likelihood function must be modified since it is unclear which individual to include at each failure time. This is a result of the cox ph model assumption that the hazard are continuous.

Approximation can be done as follows to the cox ph model using a discrete logistic model;

$$\frac{h(t,x)}{1-h(t,x)} = e^{(x\beta)}\frac{h_0 t}{1-h_0 t}$$

Where $h(t,x) = p(t \leq T < t+1 | T \geq t)$

When time is continuous then this equation reduces to

$$h(t,x)\Delta t = p(t \leq T < t+\Delta t | T \geq t)$$

Thus we have

$$\frac{h(t,x)\Delta t}{1-h(t,x)\Delta t} = e^{x\beta}\frac{h_0 t}{1-h_0 t}$$

As $\Delta t \to 0$ we have

$$h(t,x) = e^{x\beta}h_0 t$$

Using the credit performance data, the likelihood function *iii* becomes

$$L_{cox}(\beta) = \prod_{i=1}^{k} \frac{e^{(D_i \beta S_\prime)}}{\sum_{R \in R_{t_i,d_i}} e^{\beta S_{\prime R}}}$$

Where d_i denotes the number of failures at t_i, $R(t_i, d_i)$ denotes set of all subsets of d_i individuals taken from the risk set , $R(t_i)$. $R \in R(t_i, d_i)$ set of d_i individuals who might have failed at t_i.
Efron (1977) suggested an easier summation of the denominator as follows;

$$L_{cox}(\beta) = \prod_{i=1}^{k} \frac{e^{(S'^{D_i \beta})}}{\prod_{j-1}^{d_i}[y]}$$

where

$$y = \sum_{l \in R(t_i)} e^{x_i' \beta} - (\frac{j-1}{d_i}) \sum_{l \in (D_i)} e^{x_i' \beta}$$

3.8 Model Diagnostics

It is important to validate the cox model assumptions for proportionality for survival data.

Measures the disparity amongst fitted and predicted qualities.

In credit risk, we want to test if:

1. Do the cox ph model assumption hold
2. Do the covariates, x_i have to be transformed
3. Are there any outliers which might have a negative impact on parameter estimate

1. **Martingale Residuals**

   This is a transformation of the cox-Snell residual proposed by Therneau etal in 1990.

   These residuals can be viewed as the observed number of decisions either 0 or 1 for subject i between time 0 to t_i. The expected number is based on the fitted model. These residuals have a mean 0 and range between $-\infty$ and 1. For large samples, they are approximately uncorrelated. It is expressed as

   $$r_{m_i} = \partial_i - r_{c_i}$$

   where ∂_i is no. of observed events that occur at each failure time and r_{c_i} the systematic component.
   A plot of the individual covariates is used to determine the function form of a covariate.

r_{m_i} is plotted against plotted against the rank order of time and should not exhibit any pattern if the model is a good fit.

These residuals can also be expressed as

$$\hat{M}_i = M_i(\infty) - \int_0^{\infty} Y_i(t) e^{\{\hat{\beta}' z_i(t)\} d\hat{H}_0(t)}$$

For time independent covariates the model becomes

$$\begin{aligned} \hat{M}_i &= \partial_i - \hat{H}_0(t_i) e^{\{\hat{\beta}' z_i\}} \\ &= \partial_i - r_i \end{aligned}$$

$for i = 1, 2, ..., n.$
For large samples, $\hat{M}_i$ are uncorrelated and have mean 0. That is

$$\sum_{i=1}^{n} \hat{M}_i = 0$$

2. **Deviance Residuals**

   The deviance residuals is a improvement on the Martingale residuals. They help to solve the problem of the asymmetry in the Martingale residuals. For a subject i, it is defined as a function of the Martingale residual as follows:

$$\hat{D}_i = sign(M_i)\sqrt{-2[M_i + \partial_i log(\partial_i - M_i)]}$$

   M_i are the martingale residuals for the i^{th} individual and $sign(.)$ is the sign function. These are then plotted versus their individual covariates prognostic index.

3. **Schoenfeld Residuals**

   These residuals are defined for each independent variable in the model. Thus, the total number of the Schoenfeld residuals are always equal to the number of the significant independent variables. They depend on the contribution of each independent variable to the log partial likelihood. The Schoenfeld residuals can be thought of as the observed covariates minus expected covariates at each failure time. Their sum is always zero and are not defined for the censored times.

To check if the PH assumption holds, these residuals plot should exhibits a random that is unsystematic pattern at each failure time and would show that the covariates are not changing over time.

The partial likelihood equation for this residuals can be expressed as:

$$\sum_{\partial_i=1}\{Z_i(X_i) - \bar{Z}(X_i, \beta)\} = 0$$

where $Z_i(X_i)$ is the observed and $\bar{Z}(X_i, \beta)$ the expected variables.

4. **Score Residuals**

This is Plot of score residuals versus the model covariates of interest. They are useful in the identification of subjects that deviate significantly from the sample average.

These can be expressed as

$$Score_{jk} = Schoenfeld_{jk} + \sum_{t_i<t_j}[x_{jk} - \bar{x_k}(t_i)]exp(x_j\beta)[\hat{H}_0(t_i) - \hat{H}_0(t_i - 1)]$$

where $\bar{x_k}(t_i) = \frac{\sum_{i\in R}(t_i)x_{ik}e^{x\beta}}{\sum_{i\in R}(t_i)e^{x\beta}}$

For j_{th} subject and k_{th} covariate.

Chapter 4

Data analysis and Findings

4.1 Data set description

The research data is for loans up to 31 December, 2015 for the selected bank. Each loan entry for a customer is entered in it respective row in the excel data. This data has been issued on non-disclosure and thus the client name have been removed. The data has 21 variables which are easy to obtain for various banks.

The data have 1000 loan entries inclusive of loans, overdrafts, mortgages and advances. The performing loans are classified as normal while others which have not been repaid as per the loan contract have various classification such as substandard, doubtful, loss and watch. This have been coded to either 0 or 1 by a status variable, for normal which means non-defaulted loans and defaulted loans respectively.

4.2 Variables Segmentation

In the model fitted history of the previous loans repayment (Payment of previous credit), the sex and marital status (sex marital status) and the time in the current employment (length of current employment), are finely classified and the cox proportional hazard model fitted.

For instance the sex marital status categorical variable is classified into the following categories, male with the factors divorced, single, and married or widowed and female.

Coarse classified variables.
In the model fitted below, history of the previous loans repayment (Payment of previous credit), the sex and marital status (sex marital status) and the time in the current employment (length of current employment), are coarse classified and the cox proportional hazard model fitted.

For instance the sex marital status is classified as male or female.

The coarse classification typically uses simpler classification bands or features to achieve a better model than does the fine classification.
Dependent variables

Table 4.1: Dependent variable Description

Variable and Description	**Categories**	**Classification**
Status:	credit-worthy	0
	not credit-worthy	1

Independent variables

Table 4.2: Independent variable Description

Variable and Description	**Categories**	**Classification**
Loan Category	Individual	1
	Business	2
	Corporate	3
Amount of credit in '1000		
Current account balance	no balance or debit	2
	0 <= ... < 200 KES	3
	... >= 200 KES or checking account for at least 1 year	4
	no running account	1
Repayment period in months	<=6	10
	6 < .. <= 12	9
	12 < .. <= 18	8
	18 < .. <= 24	7
	24 < .. <= 30	6
	30 < .. <= 36	5
	36 < .. <= 42	4
	42 < .. <= 48	3
	48 < .. <= 54	2
	> 54	1

This table show some variables used in the models and their respective descriptions.

4.3 Descriptive Analysis

The average default rate for the loans is 30%. The default rates per each loan category are as shown below.

Loan classification Analysis

Table 4.3: Loan Classification Analysis

Loan Classification	**Non-Credit worth**	**Credit worth**	**% relative frequency for non-credit worth**	**% relative frequency for credit worth**
Individual	37	79	12.33	11.29
Business	185	511	61.67	73.00
Corporate	78	110	26.00	15.71

The output above is on the total counts for the different credit classes with their corresponding number of customers who either defaulted or payed back the loan within the term agreed. The largest group of customers are business followed by corporate and individual customers. The corporate customers have a higher default rate with 41% of the corporate loans defaulting. The business loans default at a rate of 27% which is the lowest with the individual loans defaulting at 32%.

4.4 COX PROPORTIONAL HAZARD MODEL

The proportional hazard model was proposed by Sir David Cox (1972). It depends on an established regression scheme. Partial likelihood is utilised as the maximum likelihood estimation for the estimation of the model. The model assumes that the proportional hazard hypothesis holds in the estimation of the model coefficients.

Under this section the following will be covered:

1. Cox Proportional Hazard Model fitting,
2. Residuals fitting in Cox model and

3. Assessment of Model Adequacy.

In the dataset, the duration of credit month variable is the time information; the controlling variable is the status variable (1 for default in loan repayment, 0 for censored). The covariates are the Loan Classification either as individual, business or corporate, amount in the account at the beginning of a loan (account balance), history of the previous loans repayment (Payment of previous credit), the destination of the loan granted (purpose), the amount granted (Credit amount), the value in the stock (value savings stock), the time one has been in the current employment (length of current employment), the interest rate (instalment percent), the sex and marital status (sex marital status), the type of security to the loan (guarantors), duration in the current residential place (duration in current address), the valuable assets held by the borrower (most valuable asset), the age of the borrower (age yrs), other loans held at the time of borrowing (concurrent credits), the type of ownership of the residential place (type of apartment), other loans held at this bank (no of credits at this bank), Occupation, the number of dependants (no of dependents), access to communication device (telephone), whether a local or foreigner (foreign worker).

4.4.1 Initial Cox Proportional Hazard Model

In the model fitted below, history of the previous loans repayment (Payment of previous credit), the sex and marital status (sex marital status) and the time in the current employment (length of current employment), are finely classified and the cox proportional hazard model fitted. The results are as displayed below.

Summary statistics (Events):

Table 4.4: Events Summary statistics

Total observed	Total failed	Total censored	Time steps
1000	300	700	33

From the above table, the number of observations is different from the number of observed times time steps. Thus there are tied observations. Breslow's technique for tie handling method which is the default in R-GUI is used. However, Efron's method can also be used.

Descriptive statistics (Explanatory variables):

Table 4.5: Explanatory Variables Analysis

Variable	Obs	Obs. with missing data	Obs. without missing data	Minimum	Maximum	Mean	Std. deviation
CREDIT AMOUNT	1000	0	1000	250.000	18424.000	3271.248	2822.752
AGE YEARS	1000	0	1000	19.000	75.000	35.542	11.353

These are the quantitative variables to be fitted in the model.

Table 4.6: Qualitative variables analysis

Variable	Categories	Frequencies	%
SEX_MARITAL_STATUS	1	50	5.000
	2	310	31.000
	3	548	54.800
	4	92	9.200
PAYMENT_STATUS_OF PREVIOUS_CREDIT	0	40	4.000
	1	49	4.900
	2	530	53.000
	3	88	8.800
	4	293	29.300

(Refer the appendices for the full table)
These are the quantitative variables in the model fit.

Goodness of fit statistics:

Table 4.7: Goodness of fit statistics

Statistic	Independent	Full
Observations	300	300
DF	0	4
-2 Log(Likelihood)	3479.522	3335.359
AIC	3479.522	3343.359
SBC	3479.522	3358.174
Iterations	1	2

Test of the null hypothesis H_0: beta=0; no impact of the covariates. These table will be used in the comparison for the two cox models fit.

Table 4.8: Test for the null hypothesis

Statistic	DF	Chi-square	Pr > Chi
-2 Log(Likelihood)	4	144.1633	< 0.0001
Score	4	125.2541	< 0.0001
Wald	4	114.3355	< 0.0001

The H_0 hypothesis corresponds to the independent model. We seek to check if the adjusted model is significantly better than this model.
The following tests are carried out which follow a chi-square distribution namely the likelihood ratio test (-2 Log(Likelihood)), the Score test and the Wald test. The wald test is used for the individual predictors and for the global or overall test the likelihood test is used. We can see that the Pr > Chi is <0.0001 and thus all this tests show significant for the predictors and the global model fit.

From the table above, the goodness of fit which is the quality indicator of the model is shown. The AIC value for the full model is 3443.359. This value has no meaning unless it is being used to compare between two models. The outcomes above are identical to the R^2 and to the analysis of variance table in linear regression and ANOVA. The probability

of Chi-square test, 144.1633, on the log ratio is less than 0.0001. This is proportional to the Fisher's F test, this value evaluates in the event that the variables have critical information by looking at the model as it is characterized with a simpler model with no effect of the covariates. For this situation, as the likelihood is lower than 0.0001, we can infer that significant information is brought by the variables.

Summary of the variables selection:

Table 4.9: Variables selection summary

No. of variables	Variables	Variable IN/OUT	-2 Log(Likelihood)	Pr > LR
1	ACCOUNT_BALANCE-4	IN	63.800	0.000
2	CREDIT_AMOUNT	IN	57.520	0.000
3	SEX_MARITAL_STATUS-3	IN	12.171	0.007
4	PURPOSE-3	IN	10.672	0.031

In the full model, we have 21 variables with some been categorised. Forward selection technique have been utilized. The forward selection procedure begins by including the variable with the biggest contribution to the model. On the off chance that a second variable is such that its entrance likelihood is more noteworthy than the entry threshold value, then it is added to the model.
This procedure is repeated until no new variable can be entered in the model. However, backward selection can also be used but this method is more suitable when there are few variables

Regression coefficients:

Table 4.10: Regression Coefficients

Variable	Value	Standard error	Wald Chi-Square	Pr > Chi	Hazard ratio	Hazard ratio Lower bound (95%)	Hazard ratio Upper bound (95%)
CREDIT_AMOUNT	0.000	0.000	54.362	< 0.0001	1.000	1.000	1.000
SEX_MARITAL STATUS-3	-0.432	0.118	13.497	0.000	0.649	0.515	0.817
ACCOUNT BALANCE-4	-1.065	0.149	51.226	< 0.0001	0.345	0.258	0.462
PURPOSE-3	-0.455	0.144	9.944	0.002	0.635	0.478	0.842

In the table above, the parameter estimate, comparing standard deviation, Wald's Chi-square, the relating p-value and the confidence interval are shown for every variable of the cox ph model. The hazard ratios for every variable with confidence intervals are likewise shown.

The outcomes demonstrates the impact of the different variables. From the outcome shown for the likelihood of the Chi-squares, the variable with the most impact on the survival time is credit amount and account balance.
This shows that the credited amount and account balance at issuing point of the loan significantly affects loans survival time. The hazard ratio is acquired as the exponential of the parameter estimate.

So $\bar{\beta}$ for sex_marital_status_3 =-0.432 with standard error 0.118. This means that compared to sex_marital_status_3, the other sex_marital_status levels will decrease the hazard of default at *any* time by 35% $(exp(\hat{beta}) - 1)$.

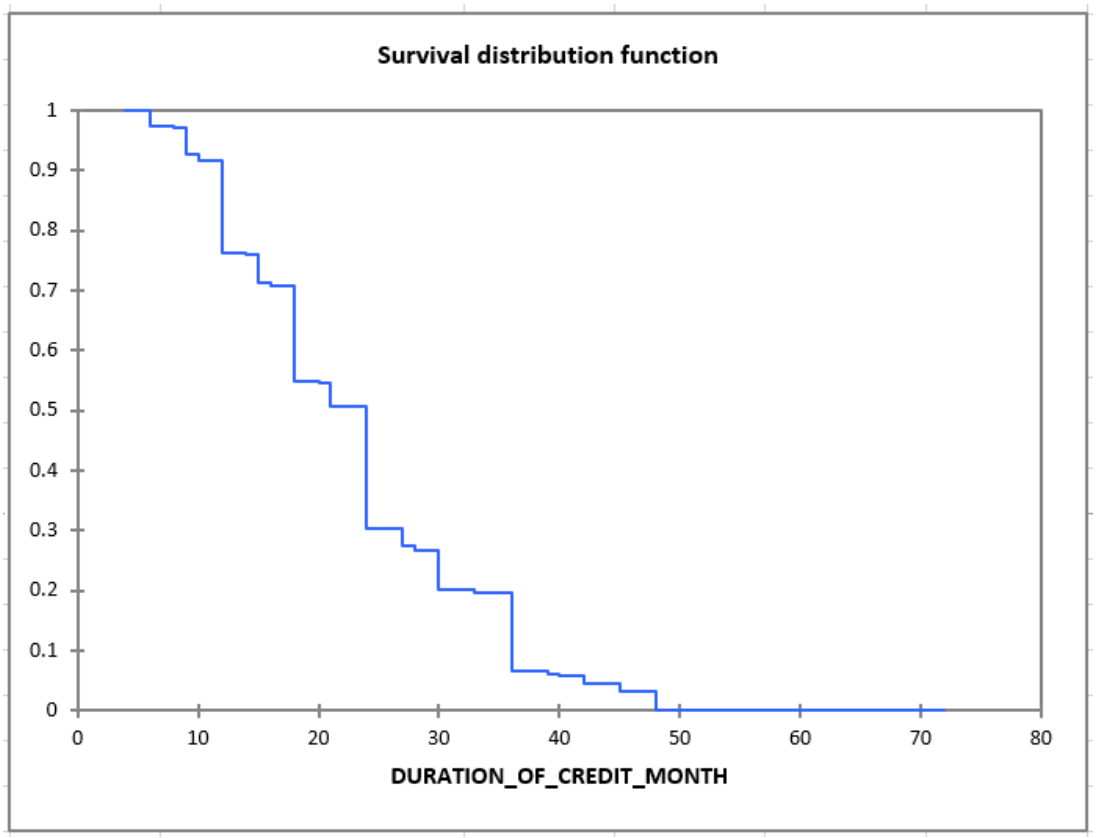

Figure 4.1: Survival distribution function

The chart above showcases the cumulative hazard function. This study has demonstrated that the main covariates with a critical effect is the credited sum and account balance. The coefficient being negative demonstrates that when a borrower has a low account balance his survival time is more prominent and the other way around. Alternate covariates don't significantly affect the survival time.

The average survival period for a loan is 21 months. At 72 months, the loan will have defaulted.

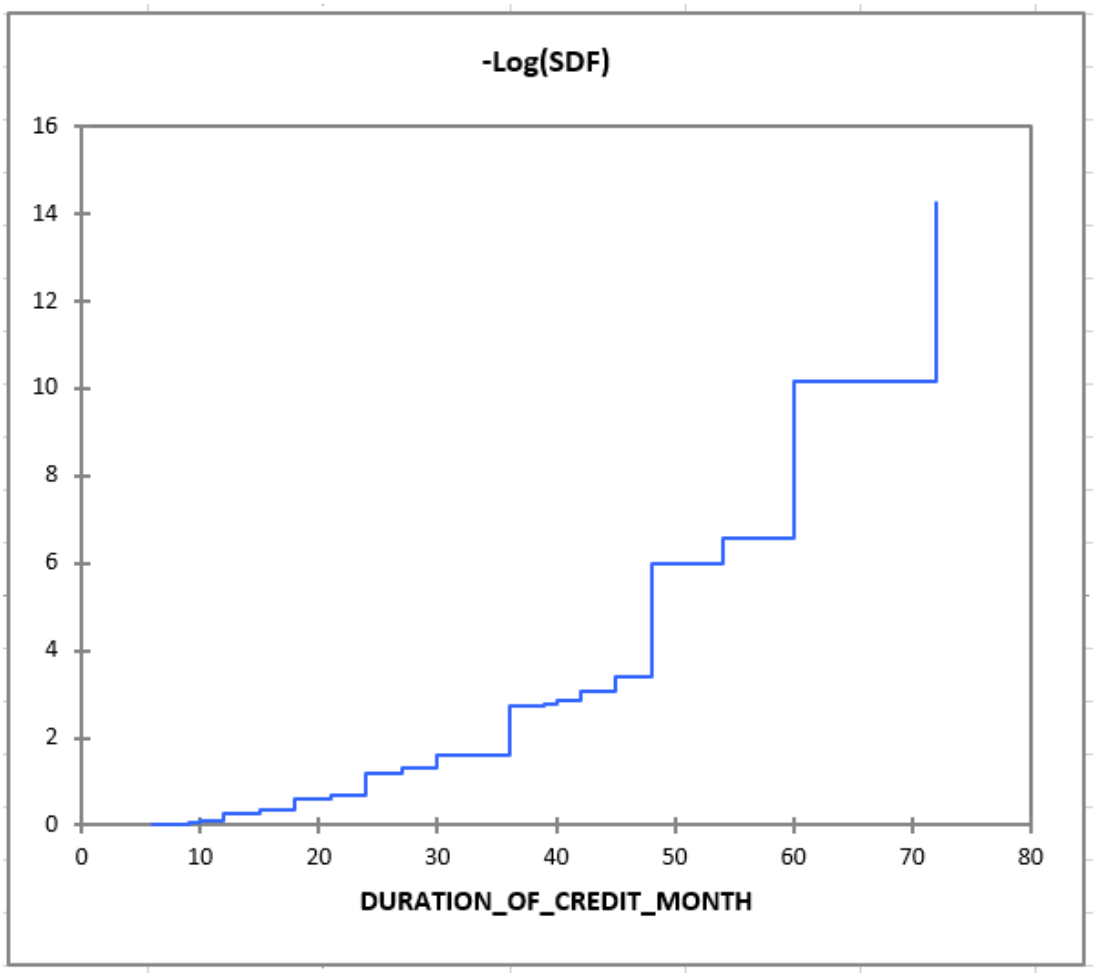

Figure 4.2: Negative log survival distribution function

This is the transformed survival function $S_i(T_i)$ for a survival function $S_i(t)$. The transformed random variable should have a uniform distribution on [0.1] and thus the plot for $-log[S_i(T_i)]$ will have a unit exponential distribution.

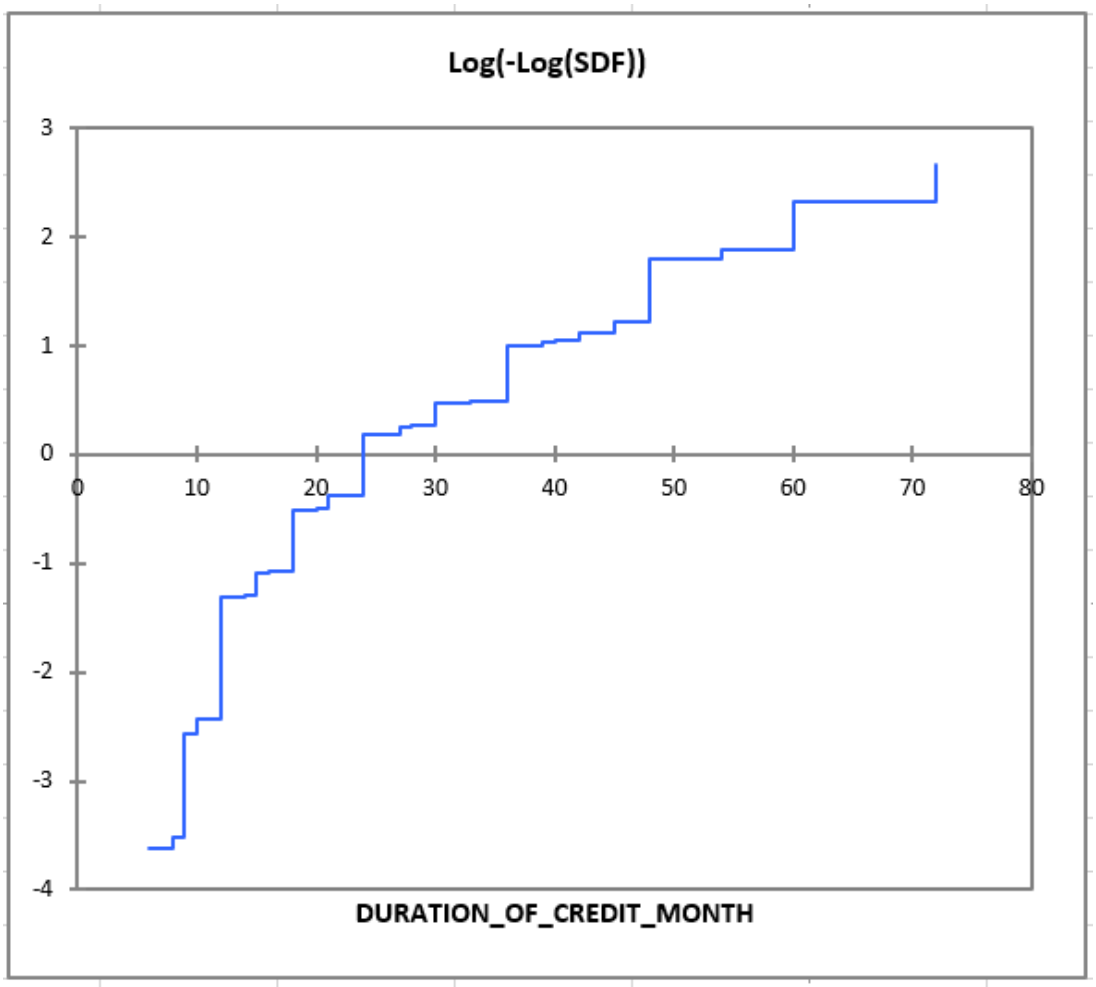

Figure 4.3: Log of Negative log survival distribution function

This fit is an adjustment of the -log(S(t)). It is useful in checking the proportionality assumptions of the hazards.

This plot demonstrates sensible fit to the PH assumption.

Residuals:
The residuals are shown, for each observation, the time variable, the censoring variable and the value of the residuals (deviance, martingale, Schoenfeld and score).

Martingale Residuals

These residuals are a linear transform of the Cox-Snell residuals defined as

1. **Cox-Snell Residuals**

   These were proposed in 1968 by cox and Snell and are defined as

$$r_{c_i} = e^{\hat{\beta} x_i . \hat{H}_0(t_i)}$$
$$\hat{H}_1(t_i) = -log\hat{S}_i(t_i)$$

   Where $\hat{H}_0(t_i)$ is the estimated cumulative baseline hazard, $\hat{H}_1(t_i)$ is the estimated cumulative hazard for the i^{th} individual at time t_i and $\hat{S}_i(t_i)$ is the estimated survival function of the i^{th} individual at time t_i.

   This residuals are used to estimate the overall fit of a cox ph model.

   The model can also be used as

$$r_j = \hat{H}_0(t_i) e^{\hat{\beta}_1 z_j}$$

   for $j = 1, 2, ..., n$ where r_j are censored test from a unit exponential distribution. We assume cox model holds and $\hat{\beta}_1$, $\hat{H}_0$ are close to the actual values of β_1 and H_0.

   If the cox model is a significant fit for the data, we expect a straight line through the origin with slope 1.

   They are useful to check functional form of covariates. For the point in the plot that appear skewed near 1 means the subjects contained "died too soon" while large negative "lived too long". These residuals range between -∞ to 1. They are calculated for the given subject, at the given timepoint t and interpreted as a difference between actual (observed) and expected (resulting from the model) number of events till the given timepoint t.

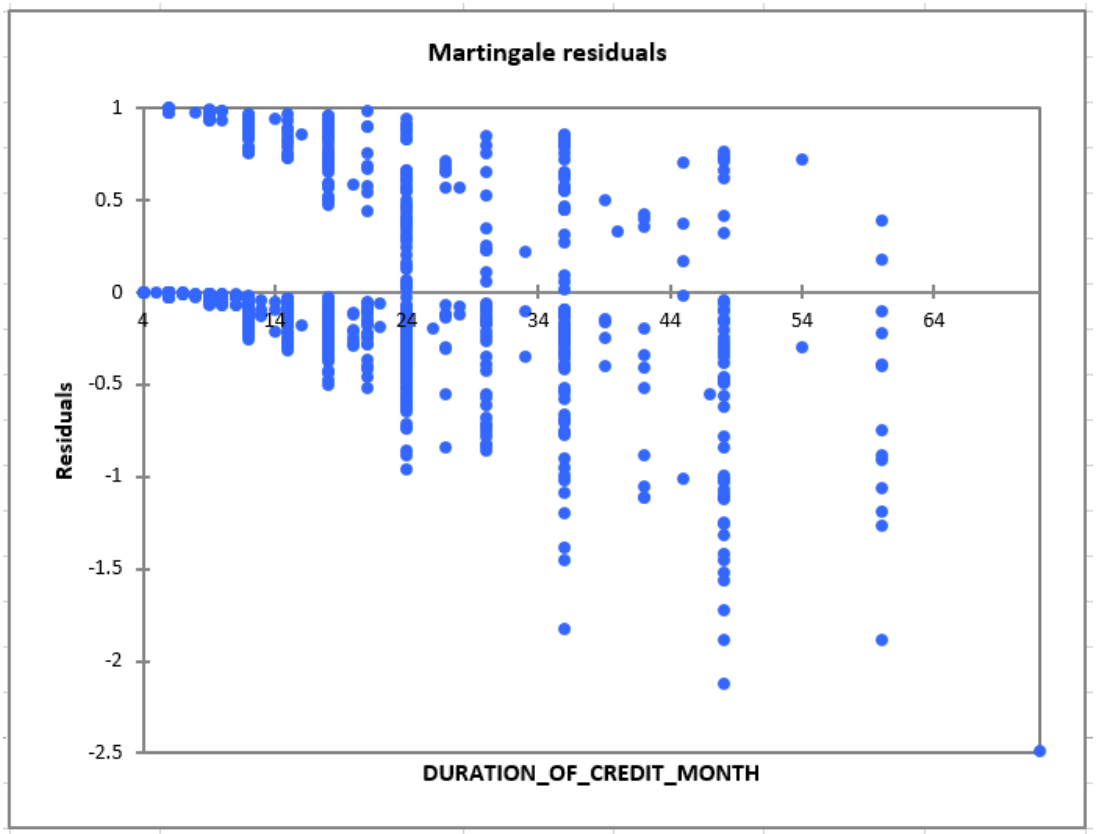

Figure 4.4: Martingale residuals

In the residual analysis for Martingale, there is an assumption of linearity which is fulfilled by the model residuals. The covariates of interest have a correct functional form and thus a good fit.

2. **Deviance Residual**

These can be thought as a change of Martingale residuals to make symmetric around zero. They are generally symmetrically disseminated around zero, with inexact standard deviation equivalent to 1. Expansive positive focuses in the models shows that the comparing subjects "died too early", while negative qualities demonstrate subjects "lived too long". Large or little values show there are exceptions and some

transformation or elimination for such variables would enhance the model. This is the main plot that is valuable for checking anomalies‘.

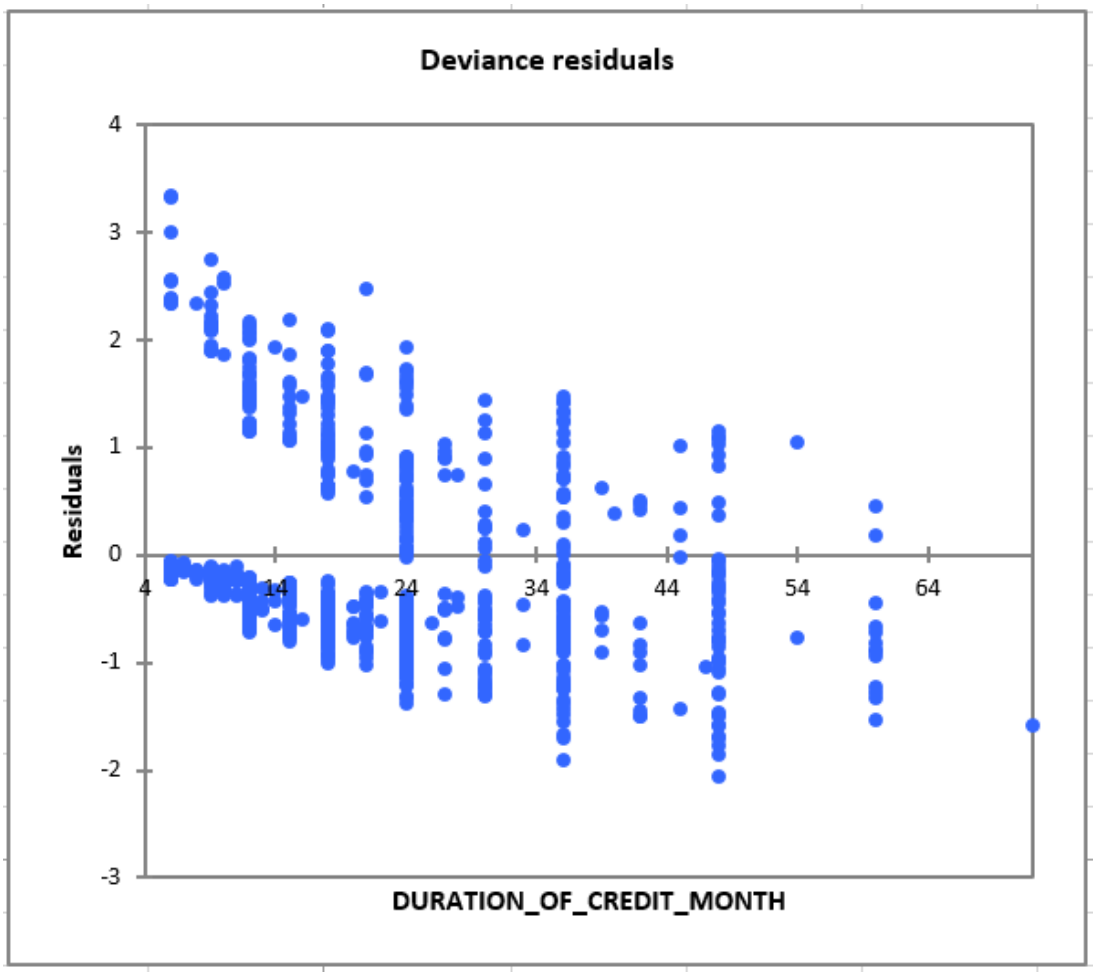

Figure 4.5: Deviance residuals

From the plot above, all the points are within the range and hence the model fit has a good fit and no transformations are required for this model.

3. **Schoenfeld Residuals**

In the consumer lending, it may not be appropriate to conclude that a particular characteristic has the same effect on the hazard rate during the life time of a loan. This is a proportional hazards assumption. The plots below shows the Schoenfeld residuals below are used to test this assumption.

The sum for these residuals is always zero. For a dichotomous (0,1) variable, Schoenfeld residuals will be between -1 and 1. The residual plot will have two bands, one above zero for *x=1*, and one below zero for *x=0*. They are useful to check for proportionality of hazards assumptions.

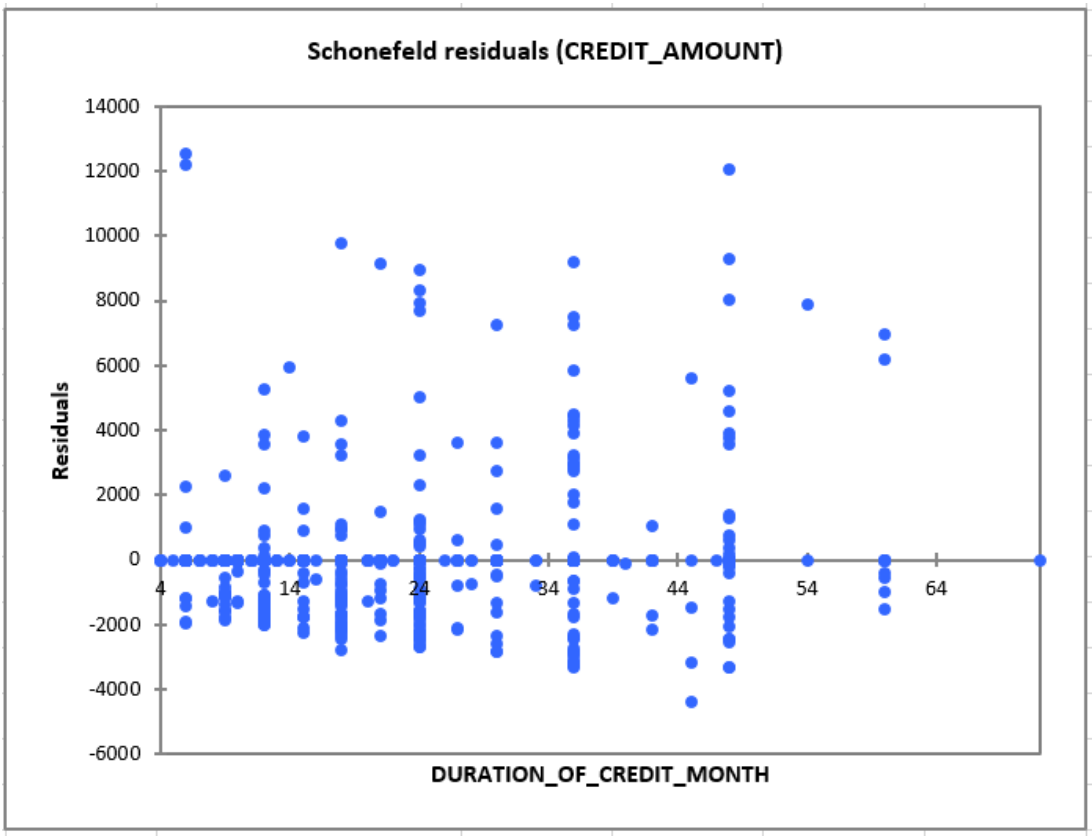

Figure 4.6: Schoenfeld residuals for credit amount

From the above plot, the Line on the plot is approximately horizontal which suggests that assumption of proportional hazard is satisfied and thus this model has a good fit for this data.

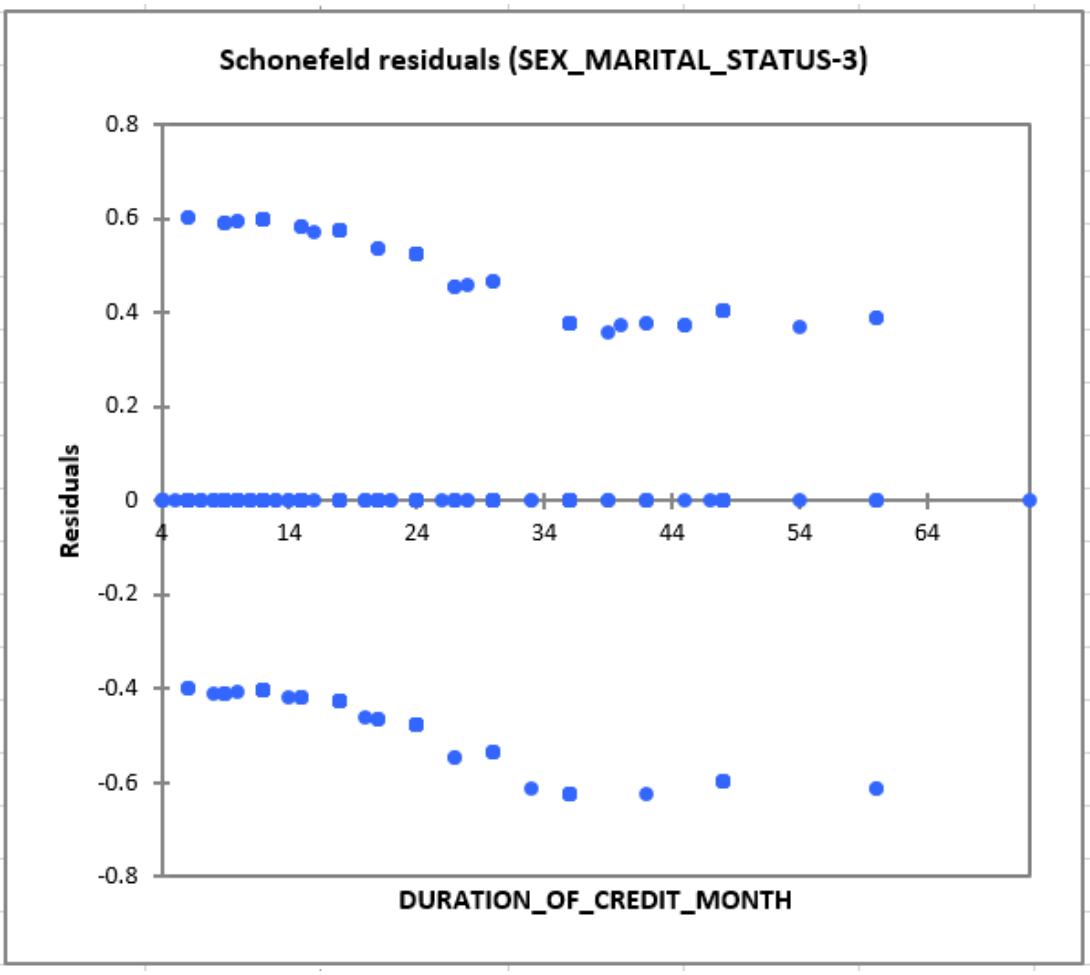

Figure 4.7: Schoenfeld residuals for sex_marital_status_3

From the figure above, the plot demonstrates a decreasing pattern, proposing a linear fit. The actual hazard ratio is linearly decreasing in log(t).

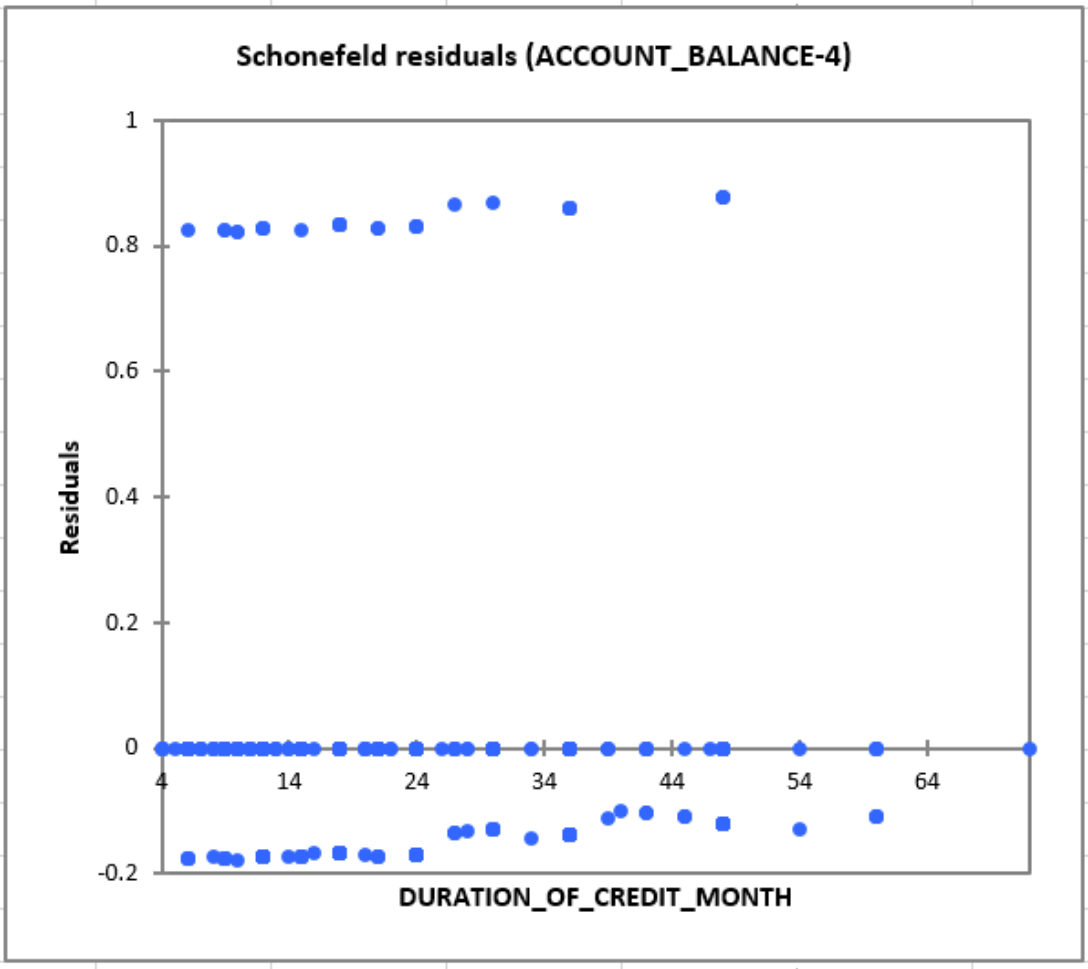

Figure 4.8: Schoenfeld residuals for account_balance_4

From the figure above, the plot displays an increasing pattern, suggesting linear fit. The actual hazard ratio is linearly increasing in log(t).

4. **Score Residuals**

These are calculated for the given subject, with respect to the given covariate and are interpreted as a weighted difference between value of the given covariate for the given subject and average value of this covariate in a risk set.

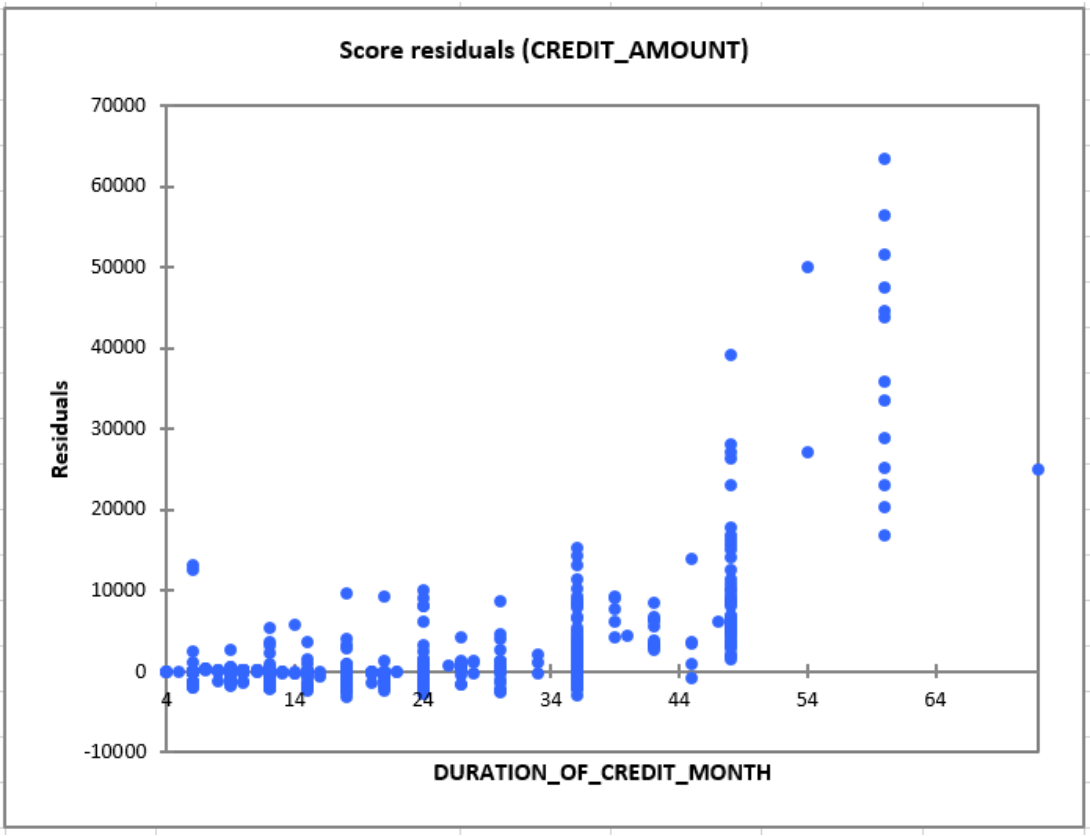

Figure 4.9: Score residuals for credit amount

The covariates in the figure above seem to lie within the sample average but a time goes by the covariates moves away from the sample average.

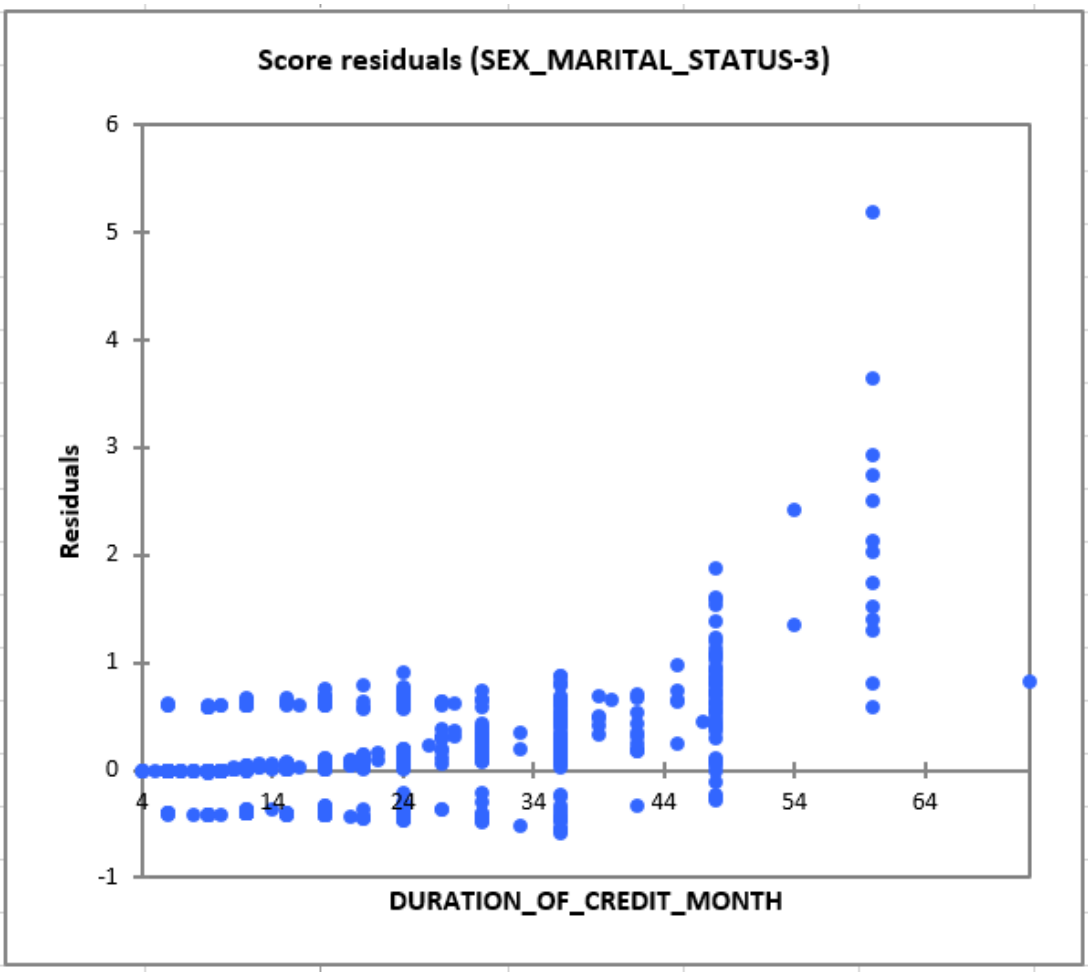

Figure 4.10: Score residuals for sex_marital_status_3

From the plot above, the covariates lie within the sample average but a time goes by the covariates moves slightly away from the sample average.

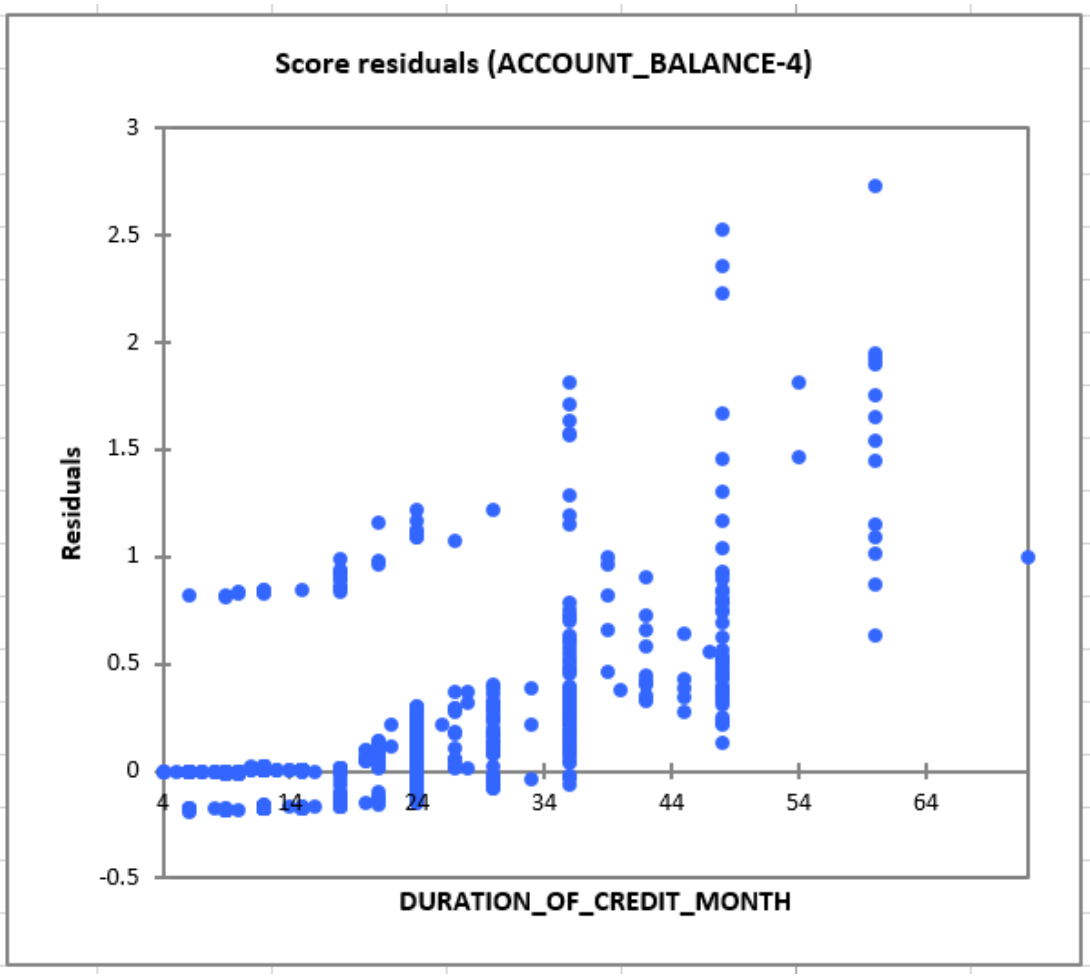

Figure 4.11: Score residuals for account_balance_4

From the figure above, most of the covariates seem to lie within the sample average but a time goes by some covariates deviates away from the sample average.

4.4.2 Improved Cox PH Model

In the model fitted below, history of the previous loans repayment (Payment of previous credit), the sex and marital status (sex marital status) and the time in the current employment (length of current employment), are coarse classified and the cox proportional hazard model fitted.

The coarse classification typically uses simpler classification bands or features to achieve a better model than does the fine classification. The results are as displayed below.

Summary statistics (Events):

Table 4.11: Events summary statistics

Total observed	Total failed	Total censored	Time steps
1000	300	700	33

From the above table, the number of observations is different from the number of observed times time steps. Thus there are tied observations. Breslow's technique for tie handling method which is the default in R-GUI is used. However, Efron's method can also be used.

Descriptive statistics (Explanatory variables):

Table 4.12: Explanatory variables descriptive statistics

Variable	Obs	Obs. with missing data	Obs. without missing data	Minimum	Maximum	Mean	Std. deviation
CREDIT AMOUNT	1000	0	1000	250.000	18424.000	3271.248	2822.752
AGE YEARS	1000	0	1000	19.000	75.000	35.542	11.353

These are the quantitative variables in the model.

Table 4.13: Quantitative variables summary statistics

Variable	Categories	Frequencies	%
PREVIOUS_CREDITS_PAYMENTS	1	89	8.900
	2	911	91.100
SEX	1	92	9.200
	2	908	90.800
LOAN_CLASSIFICATION	1	116	11.600
	2	696	69.600
	3	188	18.800

(Refer appendices for the complete table)
These are the quantitative variables under this model.

Goodness of fit statistics:

Table 4.14: Goodness of fit statistics

Statistic	Independent	Full
Observations	300	300
DF	0	5
-2 Log(Likelihood)	3479.522	3315.881
AIC	3479.522	3325.881
SBC	3479.522	3344.400
Iterations	1	2

Test of the null hypothesis H_0: beta=0

Table 4.15: Test for the null hypothesis

Statistic	DF	Chi-square	Pr > Chi
-2 Log(Likelihood)	5	163.6416	< 0.0001
Score	5	178.3635	< 0.0001
Wald	5	181.0917	< 0.0001

The H_0 hypothesis corresponds to the independent model. We seek to check if the adjusted model is significantly better than this model.
The following tests are carried out which follow a chi-square distribution namely the likelihood ratio test (-2 Log(Likelihood)), the Score test and the Wald test. All these tests indicate significant predictors and overall model fit as (Pr > Chi) < 0.0001. From table 4.12 above, the goodness of fit which is the quality indicator of the model is shown.
The AIC value for the full model is 3325.881. This value has no meaning unless it is being used to compare between two models. The results above are proportional to the R^2 and to the analysis of variance table in linear regression and ANOVA.
The probability of Chi-square test, 163.6416, on the log ratio is less than 0.0001. This is proportional to the Fisher's F test, this value evaluates in the event that the variables have critical information by looking at the model as it is characterized with a simpler model with no effect of the covariates. For this situation, as the likelihood is lower than 0.0001, we can infer that significant information is brought by the variables.

Summary of the variables selection:

Table 4.16: Variables selection summary

No. of variables	Variables	Variable IN/OUT	-2 Log(Likelihood)	Pr > LR
1	ACCOUNT_BALANCE-4	IN	63.800	0.000
2	CREDIT_AMOUNT	IN	57.520	0.000
3	LOAN_CLASSIFICATION-2	IN	9.931	0.019
4	LOAN_CLASSIFICATION-3	IN	23.070	0.000
5	VALUE_SAVINGS_STOCKS-5	IN	9.321	0.097

In the full model, we have 21 variables with some been categorised. Forward selection technique have been utilized. The forward selection procedure begins by including the variable with the biggest contribution to the model. On the off chance that a second variable is such that its entrance likelihood is more noteworthy than the entry threshold value, then it is added to the model.
This procedure is repeated until no new variable can be entered in the model. However, backward selection can also be used but this method is more suitable when there are few variables.

Regression coefficients:

Table 4.17: Regression Coefficients

Variable	Value	Standard error	Wald Chi-Square	Pr > Chi	Hazard ratio	Hazard ratio Lower bound (95%)	Hazard ratio Upper bound (95%)
CREDIT_AMOUNT	0.000	0.000	5.590	0.018	1.000	1.000	1.000
LOAN CLASSIFICATION-2	-1.537	0.171	80.580	< 0.0001	0.215	0.154	0.301
LOAN CLASSIFICATION-3	-1.937	0.323	36.080	< 0.0001	0.144	0.077	0.271
ACCOUNT BALANCE-4	-1.024	0.147	48.217	< 0.0001	0.359	0.269	0.479
VALUE_SAVINGS STOCKS-5	-0.500	0.173	8.330	0.004	0.607	0.432	0.852

In the table above, the parameter estimate, comparing standard deviation, Wald's Chi-square, the relating p-value and the confidence interval are shown for every variable of the cox ph model. The hazard ratios for every variable with confidence intervals are likewise shown.

The outcomes demonstrates the impact of the different variables. From the outcome shown for the likelihood of the Chi-squares, the variable with the most impact on the survival time is credit amount and account balance.
This shows that the credited amount and account balance at issuing point of the loan significantly affects loans survival time. The hazard ratio is acquired as the exponential of the parameter estimate.

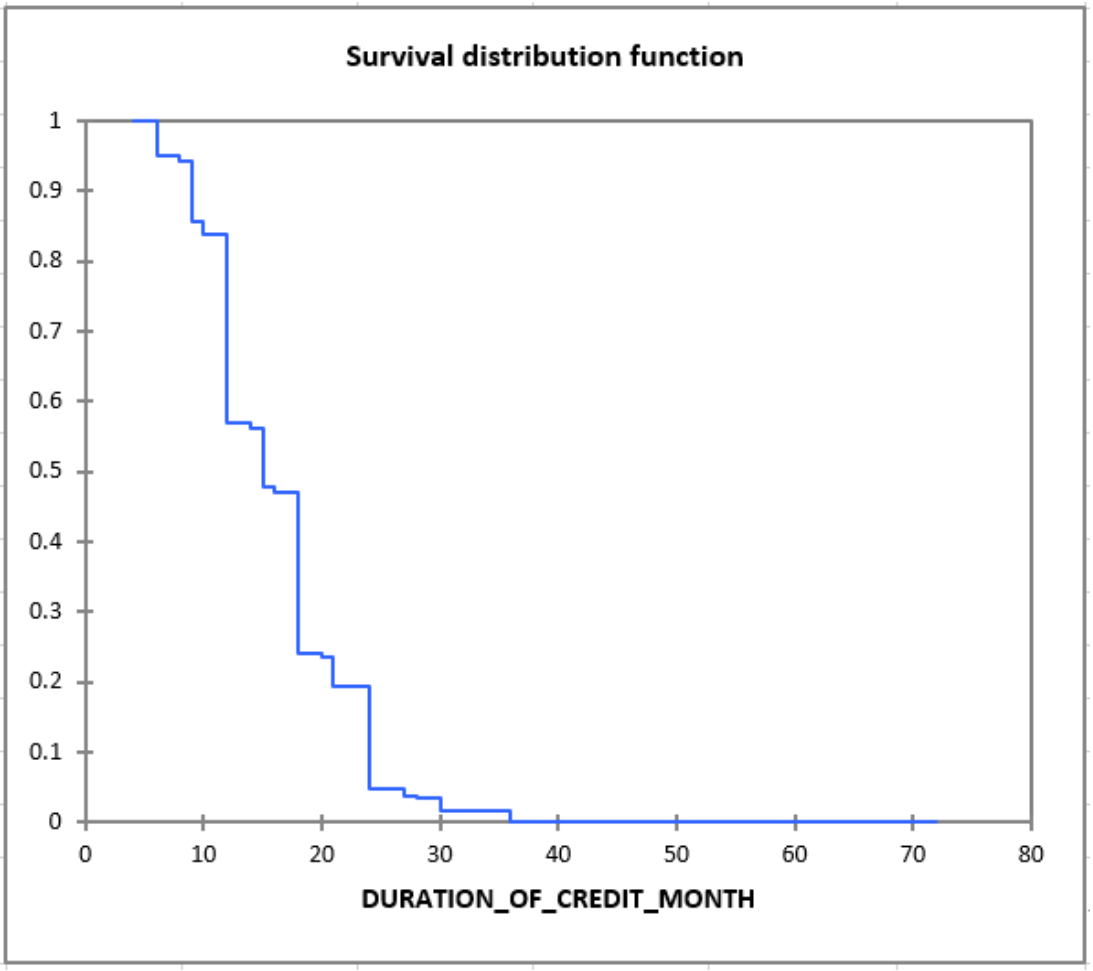

Figure 4.12: Survival distribution function

The chart above showcases the cumulative hazard function. This study has demonstrated that the main covariates with a critical effect is the credited sum and account balance. The coefficient being negative demonstrates that when a borrower has a low account balance his survival time is more prominent and the other way around. Alternate covariates don't significantly affect the survival time.

The average survival time for a loan is 16 months. By the time the loan is at 45 months it will have defaulted.

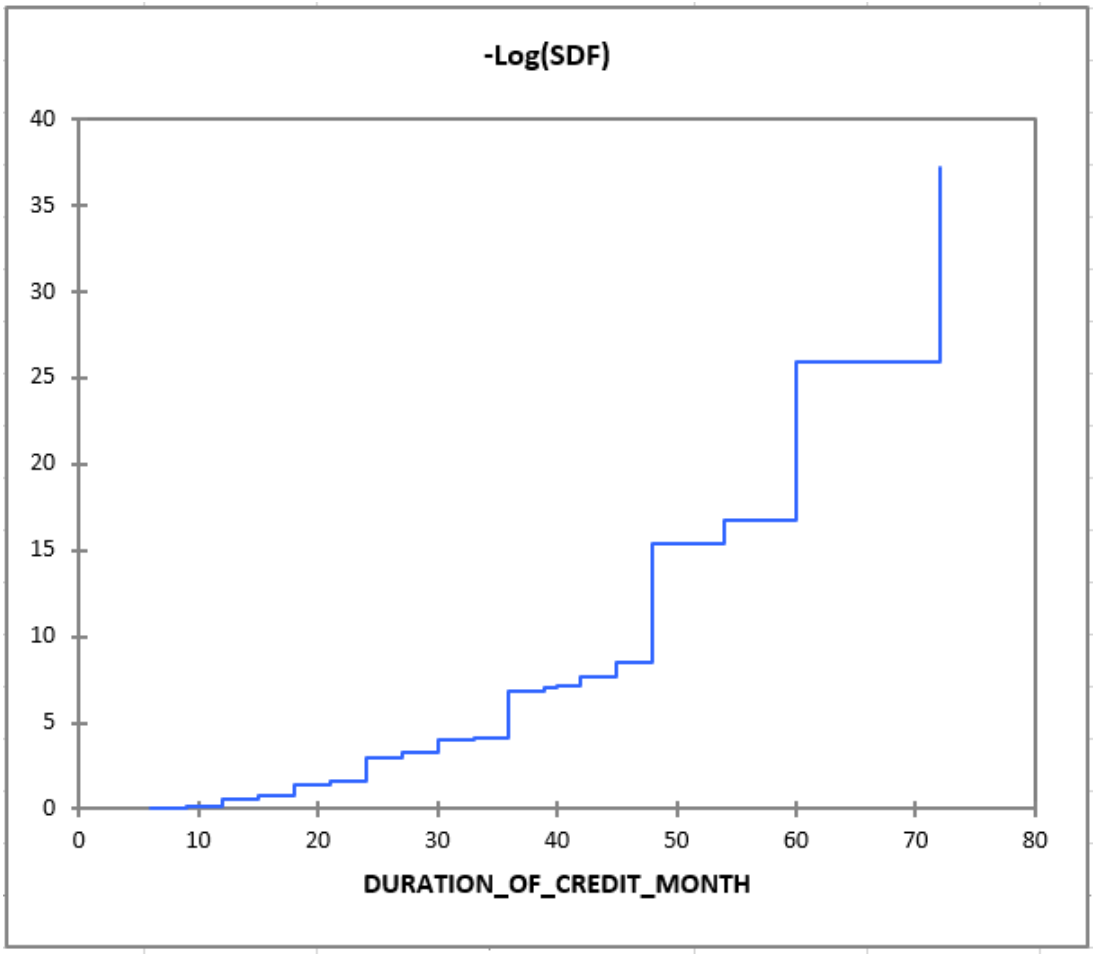

Figure 4.13: Negative log of survival distribution function

This fit will be adjustment of the -log(S(t)) and interpreted below. It is useful in checking the proportionality assumptions of the hazards.

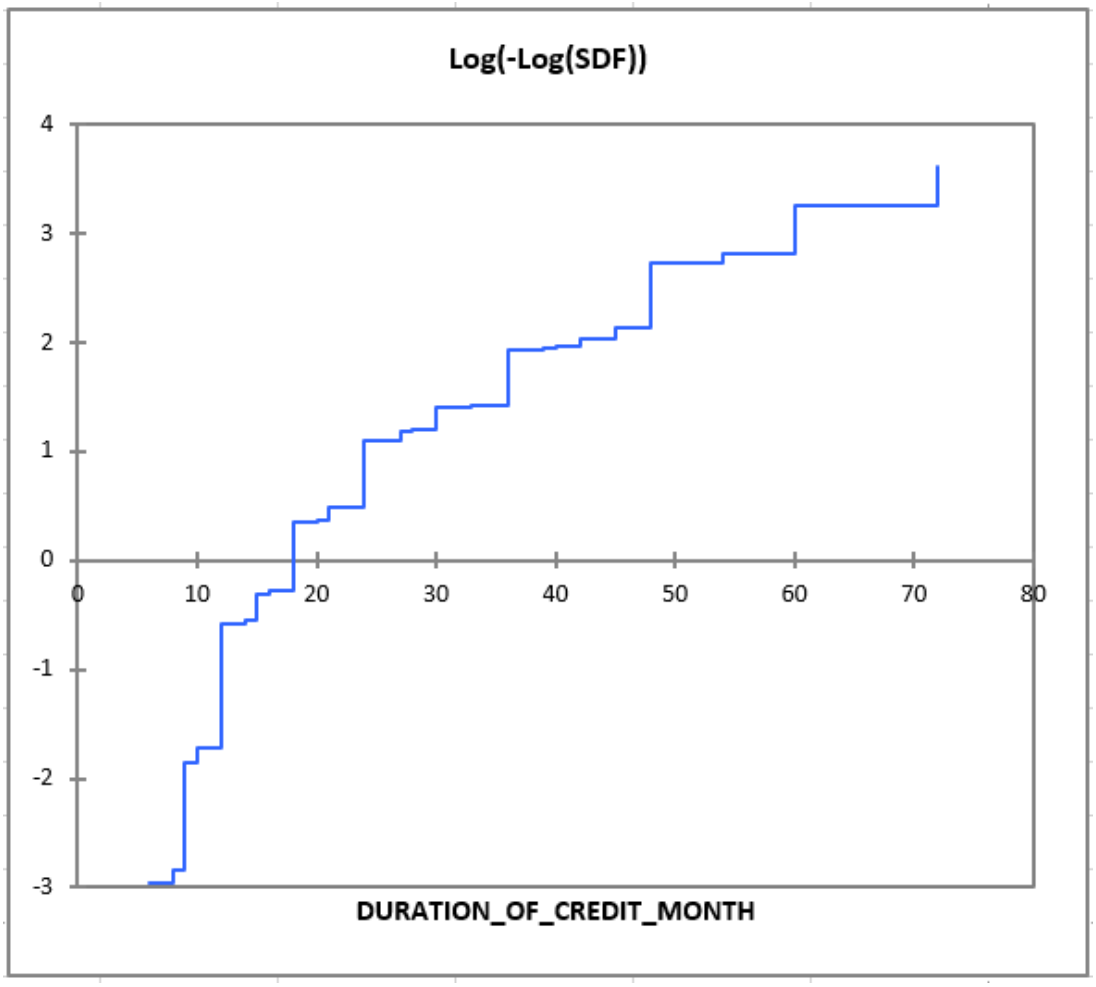

Figure 4.14: Log of negative log survival distribution function

This plot indicates sensible fit to the PH suspicion.

Residuals:

The residual tables below shows, for every observation, the time variable, the censoring variable and the estimation of the residuals (deviance, martingale, Schoenfeld and score).

1. **Martingale Residual**

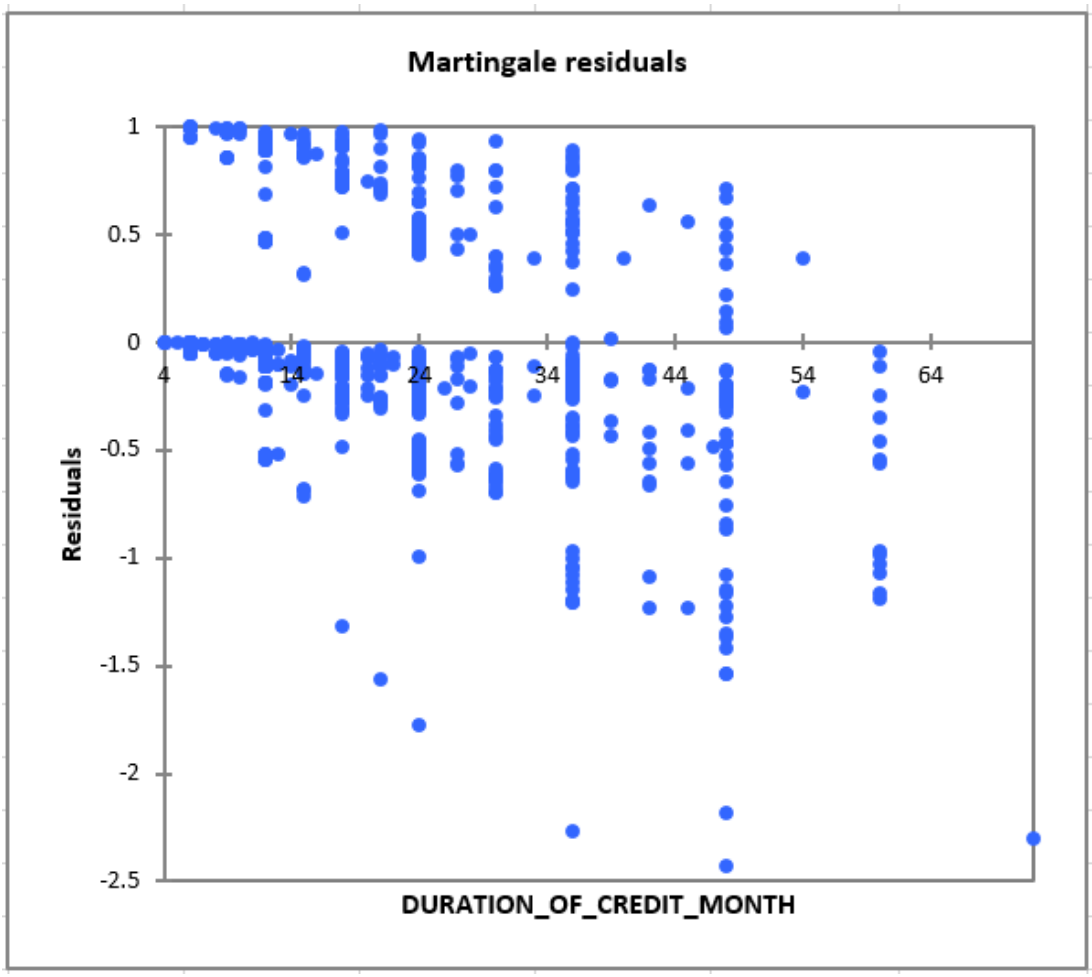

Figure 4.15: Martingale residuals

In the residual analysis for Martingale, there is an assumption of linearity which is fulfilled by the model residuals. The covariates of interest have a correct functional form and thus a good fit.

2. **Deviance residuals**

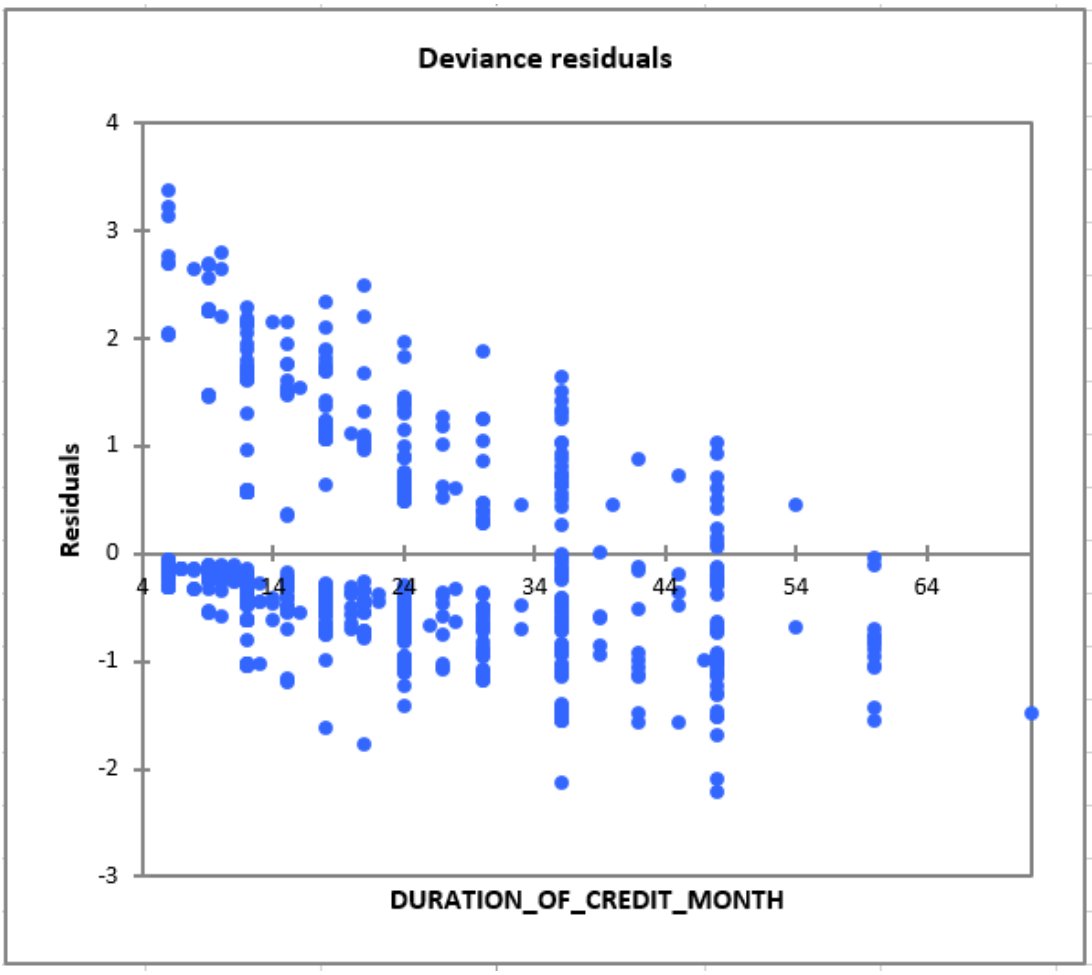

Figure 4.16: Deviance residuals

From the plot above, all the points are within the range and hence the model is a good fit and no transformations are required for this case.

3. **Schoenfeld residuals**

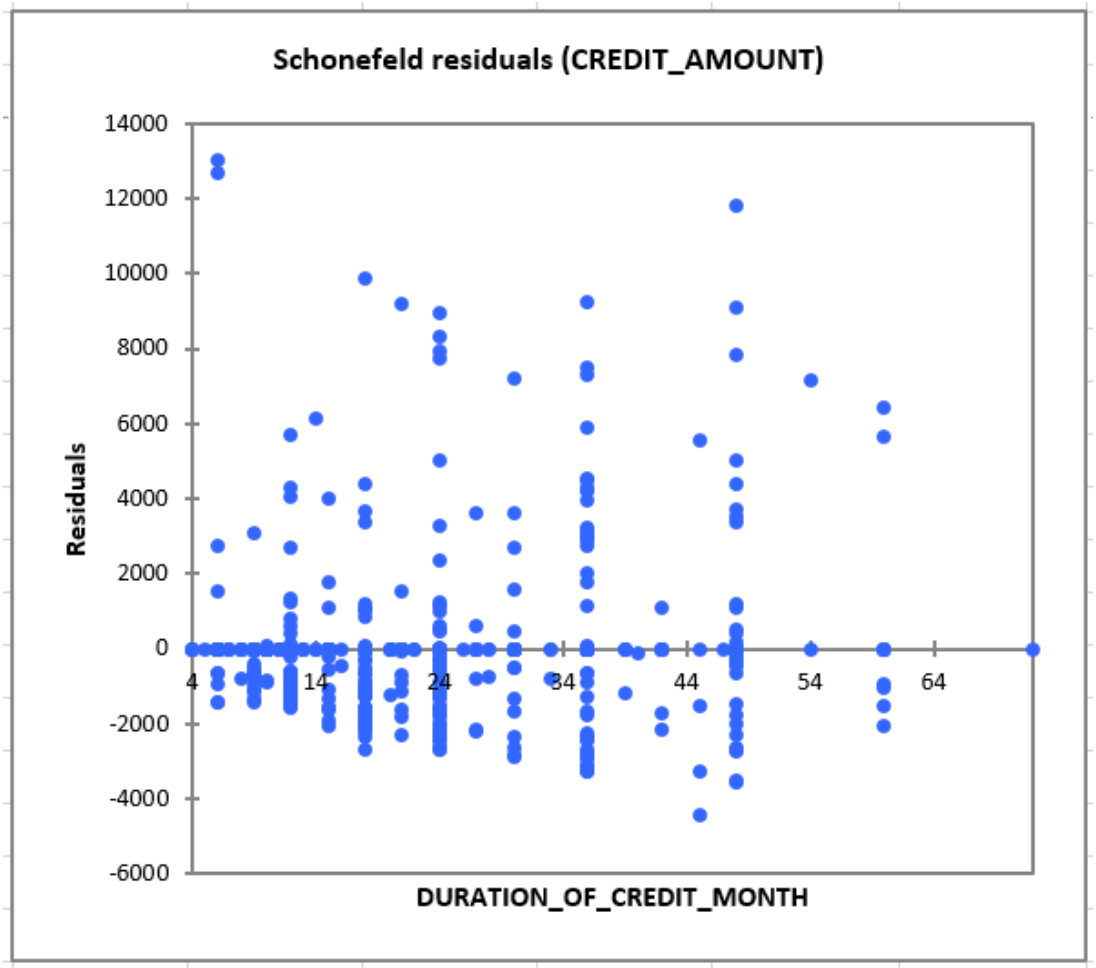

Figure 4.17: Shoenfeld residuals for credit amount

From the above plot, the Line on the plot is approximately horizontal which suggests that assumption of proportional hazard is satisfied and thus this model is a good fit.

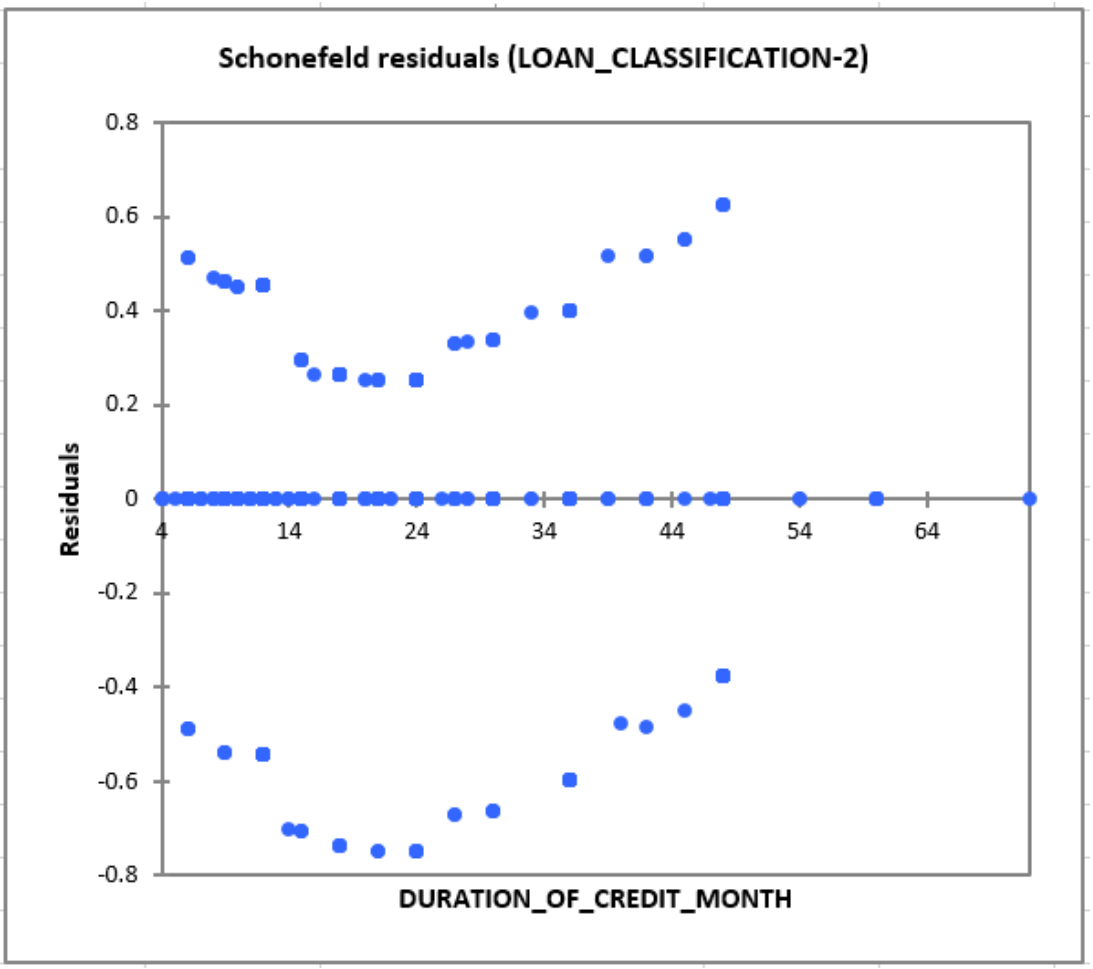

Figure 4.18: Shoenfeld residuals for loan_classification_2

From the figure above, the plot displays an increasing pattern, suggesting linear fit. The actual hazard ratio is linearly increasing in log(t).

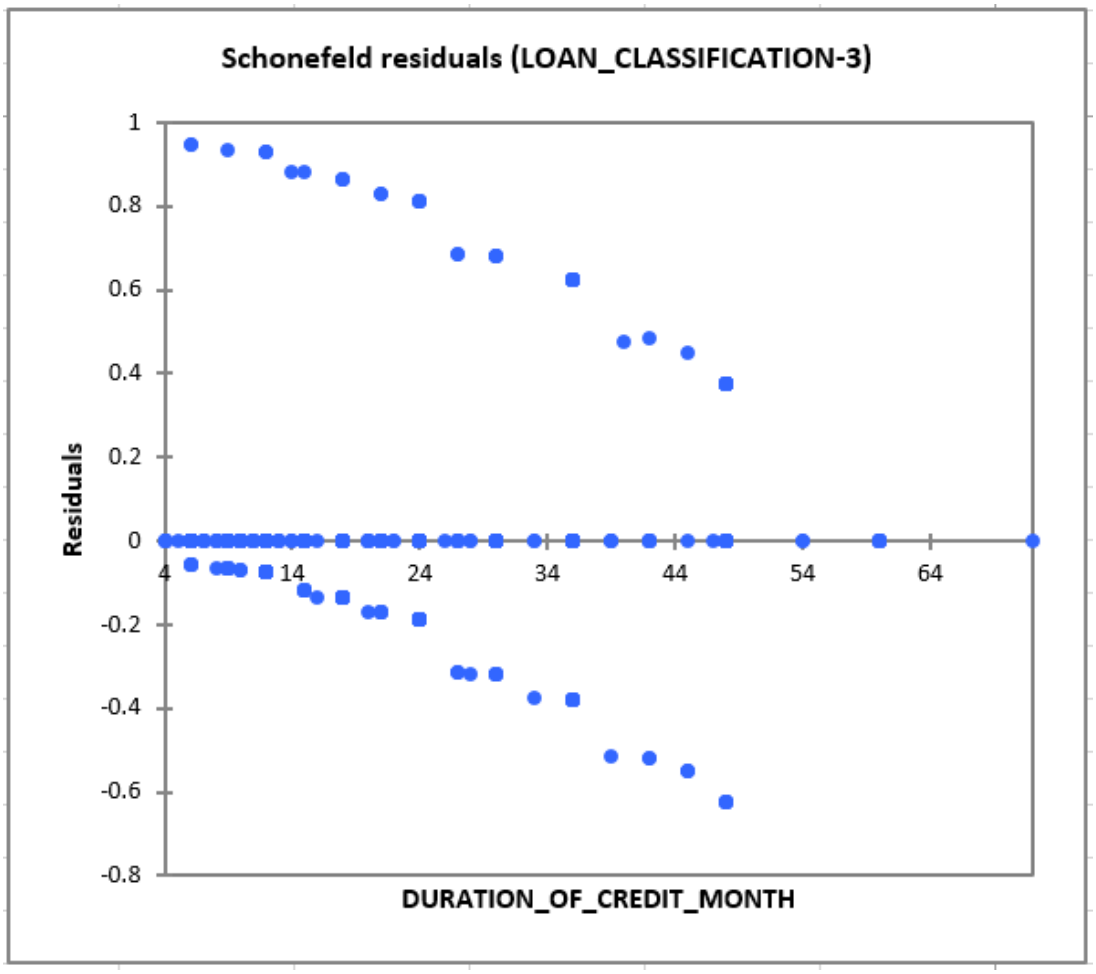

Figure 4.19: Shoenfeld residuals for loan_classification_3

From the figure above, the plot displays a decreasing pattern, suggesting linear fit. The actual hazard ratio is linearly decreasing in log(t).

4. **Score residuals**

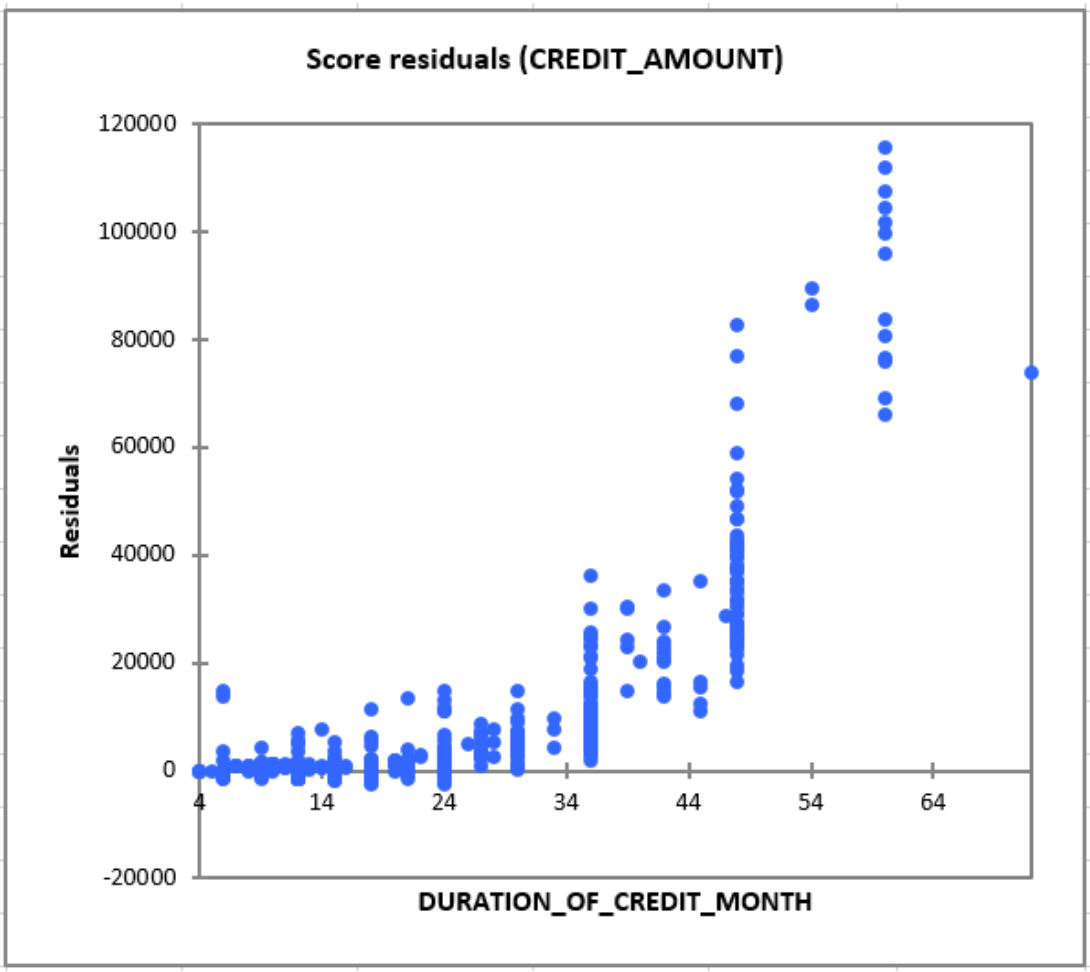

Figure 4.20: Score residuals for credit amount

In the figure above, covariates lie within the sample average but a time goes by the covariates moves away from the sample average towards the positive values.

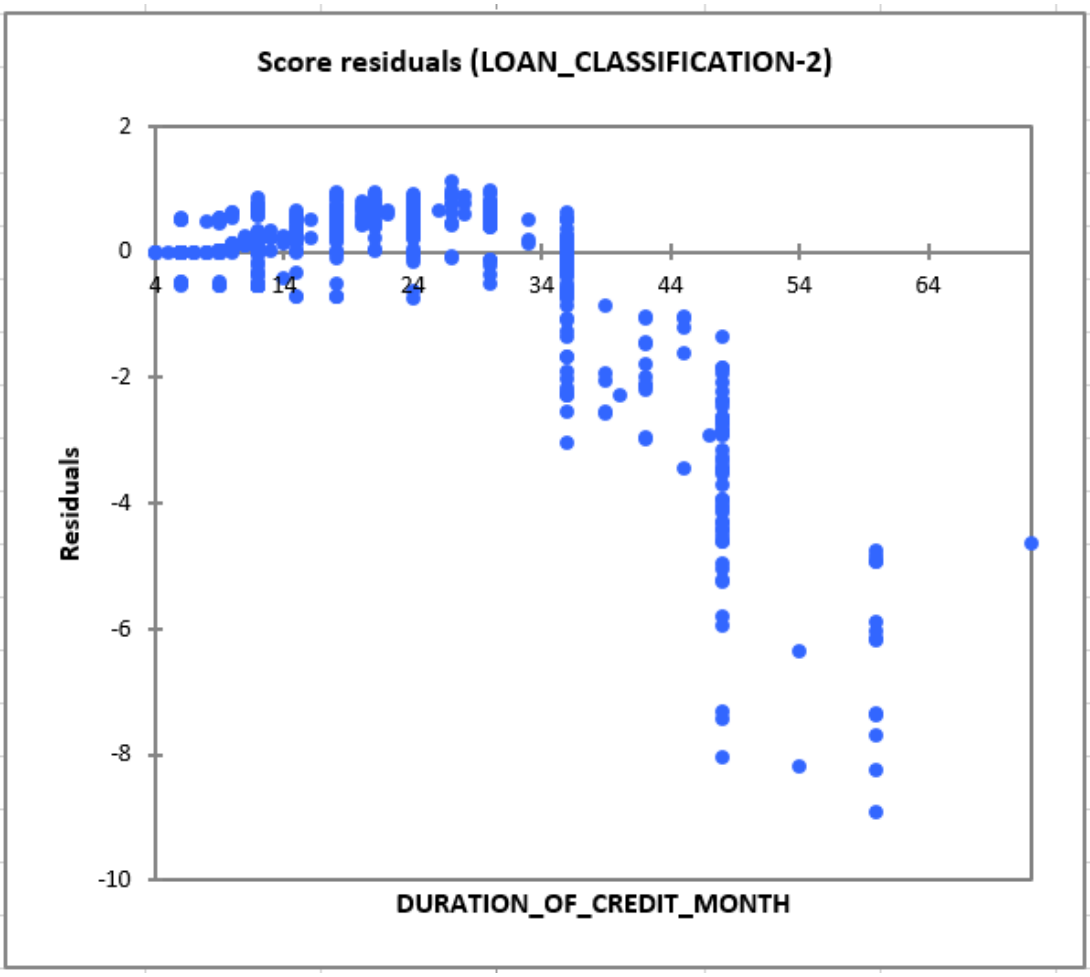

Figure 4.21: Score residuals for loan_classification_2

The covariates seem to lie within the sample average but a time goes by the covariates moves away from the sample average towards the negative values.

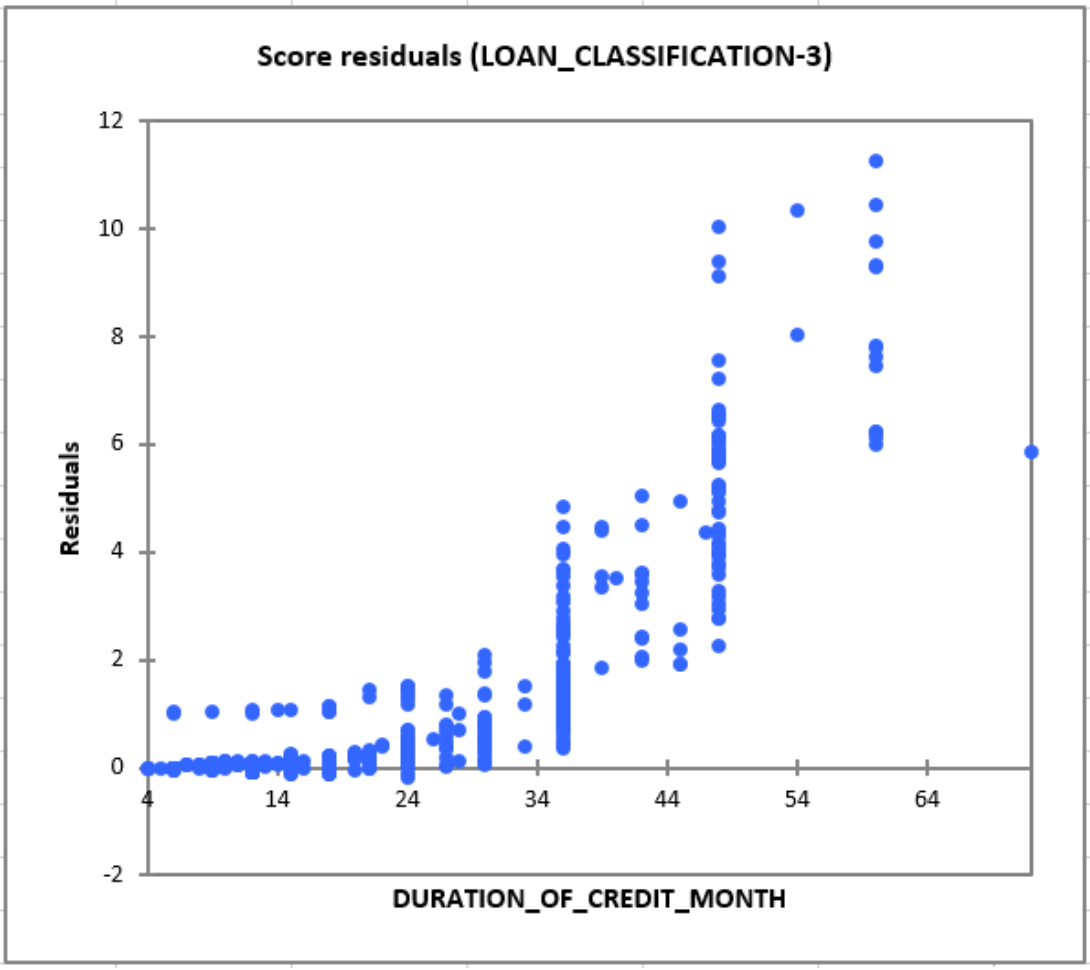

Figure 4.22: Score residuals for loan_classification_3

The covariates seem to lie within the sample average but a time goes by the covariates moves away from the sample average.

4.5 Models Selection

In the initial cox model, the AIC value is 3343.359 as compared to the improved cox model with an AIC value of 3325.881. This shows that the improved model is a better fit as it has a smaller AIC value.

Chapter 5

Conclusion and Recommendation

5.1 Conclusion

Survival analysis models have been used in financial risk management as alternative tools for financial institutions to calculate risks over time. Here we considered time to default of a loan portfolio data set. Residuals were used to check if the model satisfy the proportional hazard conditions expected in the cox PH modelling.

The study shows that there are numerous factors which influence non-repayment of a loan arising from the loan and the individual characteristic. These factors include the amount lent, loan classification, the value of the stock at the beginning of a loan, sex and the marital status of a borrower, purpose of the loan and the account balance at the start of a loan.

The improved cox ph model is better with an AIC value of 3325.881 than the initial cox ph model with AIC value of 3343.359 which has a fine classification of variables.

The study shows that the cox ph regression model can be adjusted and improved further. By coarse classification of variables, the model becomes more efficient. There are also various method for testing if the assumptions of the model hold using the residuals. This study has also shown how these methods can be used to increase the accuracy of the decision maker. The study has shown how the limitation of the Martingale residuals of being asymmetric can be overcome by using the deviance residuals.

5.2 Recommendations

The study recommends further studies on coarse classification schemes such as clustering. Coarse classification can also be used in the development of systems in computing which will have a better speed especially in computations or other processing to fine classification. In sampling, coarse classification can be applied on the individuals to select a sample quickly where the experiment will be carried out.

In addition, more improvement on the model can be studied on, which can be based on the bias reduction of the cox parameter which is biased away from 0. This can be done by combining with the proper weights of a generalized log rank and cox estimates to generate a new estimator which would be almost unbiased and better than the cox estimates.

More studies can be carried on the discretization of data and its effect on the partial likelihood under both proportional odds and proportional hazards.

Other improvements can be studied on including the smoothing of baseline hazard. The smoothing of the baseline hazard was proposed by Efron (1988) and Guillen et al (2007).

With more investment on the survival analysis studies, all stakeholders in the financial markets, medicine, engineering and other fields will be able to have model that suit their operations in the survival analysis.

Bibliography

[AR06] Allen L.N., Rose L.C. *Financial survival analysis of defaulted debtors, Journal of Research Society*, 2006.

[BOA14] Babajide Abiola A, Olokoyo Felicia O and Adegboye Folasade B *Predicting Bank Failure in Nigeria Using Survival Analysis Approach*, 2014.

[BJT99] Banasik, J., Crook, J. N., Thomas, L. C. *Not if but when will borrowers default. Journal of the Operational Research Society*, 1999.

[BCBS04] Basel Committee on Banking Supervision, https://www.bis.org/bcbs/. *International convergence of capital measurement and capital standards: a revised framework*,2004.

[BG52] Berkson, J., Gage, R. P..*Survival curve for cancer patients following treatment. Journal of the American Statistical Association 47 (259)*, 1952.

[BC13] Bonini, S., Caivano, G.. *The survival analysis approach in basel ii credit risk management: modeling danger rates in the loss given default parameter. Journal of Credit Risk Volume 9 (1)*, 2013.

[B75] Breslow, N. E.. *Analysis of survival data under the proportional hazards model. International Statistical Review/Revue Internationale de Statistique, 4557* 1975.

[CK] Carmen M. Reinhart, Kenneth S. Rogoff *Growth in a Time of Debt, Working paper 15639: Available on: http:*
www.nber.org
papers
w15639.pdf.

[C72] Cox, D. R.. *Regression models and life-tables. Journal of the Royal Statistical Society. Series B (Methodological), 187220*,1972.

[EA02] Edward I. Altman. *Revisiting Credit Scoring Models in a Basel 2 Environment* ,2002.

[EN08] Espen Gaarder Haug & Nassim Nicholas Taleb. *Why We Have Never Used the Black-Scholes-Merton Option Pricing Formula*,2008.

[IDT10] Isik, B, Deniz, S. and Taner, B. *Bayesian Credit Scoring Model with Integration of Expert Knowledge and Consumer Data*,2010.

[LLW86] Lane, W. R., Looney, S. W., Wansley, J. W..*An application of the cox proportional hazards model to bank failure. Journal of Banking and Finance 10 (4), 511531,* 1986.

[ST02] Stepanova,M., Thomas, L.*Survival analysis methods for personal loan data. Operations Research 50 (2), 277289*,2002.

[ST00] Sy, J. P., Taylor, J. M..*Estimation in a cox proportional hazards cure model. Biometrics 56 (1), 227236*,2000.

[TP00] Terry M. Therneau, Patricia M. Grambsch.*Modeling SurvivalData: Extending the CoxModel. Springer, New York.c 2014 NSP Natural Sciences Publishing Cor*,2000.

[T00] Thomas, L.C.*A Survey of Credit and Behavioural Scoring: Forecasting Financial Risk of Lending to Customers, International Journal of Forecasting, vol.16, no.2, pp.149*,2000.

[TPT90] Terry. M. Therneau, Patricia. M. Grambsch, and T. R. Fleming.*Martingale-based residuals for survival models, Biometrika. 77(1990), 147 160*,1990.

[W91] Whalen, G..*A proportional hazards model of bank failure: an examination of its usefulness as an early warning tool. Federal Reserve Bank of Cleveland Economic Review 27 (1), 2131,*1991

Appendices

Table 5.1: Quantitative variables summary statistics initial cox.

Begin of Table			
Variable	Categories	Frequencies	%
SEX_MARITAL_STATUS	1	50	5.000
	2	310	31.000
	3	548	54.800
	4	92	9.200
PAYMENT_STATUS_OF_PREVIOUS CREDIT	0	40	4.000
	1	49	4.900
	2	530	53.000
	3	88	8.800
	4	293	29.300
LENGTH_OF_CURRENT_EMPLOYMENT	1	62	6.200
	2	172	17.200
	3	339	33.900
	4	174	17.400
	5	253	25.300
LOAN_CLASSIFICATION	1	116	11.600
	2	696	69.600
	3	188	18.800
ACCOUNT_BALANCE	1	274	27.400
	2	269	26.900
	3	63	6.300
	4	394	39.400
PURPOSE	0	234	23.400
	1	103	10.300
	2	181	18.100
	3	280	28.000
	4	12	1.200
	5	22	2.200
	6	50	5.000

Continuation of Table 5.1			
Variable	Categories	Frequencies	%
	8	9	0.900
	9	97	9.700
	10	12	1.200
VALUE_SAVINGS_STOCKS	1	603	60.300
	2	103	10.300
	3	63	6.300
	4	48	4.800
	5	183	18.300
INSTALMENT_PER_CENT	1	136	13.600
	2	231	23.100
	3	157	15.700
	4	476	47.600
GUARANTORS	1	907	90.700
	2	41	4.100
	3	52	5.200
DURATION_IN_CURRENT_ADDRESS	1	130	13.000
	2	308	30.800
	3	149	14.900
	4	413	41.300
MOST_VALUABLE_AVAILABLE_ASSET	1	282	28.200
	2	232	23.200
	3	332	33.200
	4	154	15.400
CONCURRENT_CREDITS	1	139	13.900
	2	47	4.700
	3	814	81.400
TYPE_OF_APARTMENT	1	179	17.900
	2	714	71.400
	3	107	10.700
NO_OF_CREDITS_AT_THIS_BANK	1	633	63.300
	2	333	33.300

Continuation of Table 5.1			
Variable	Categories	Frequencies	%
	3	28	2.800
	4	6	0.600
OCCUPATION	1	22	2.200
	2	200	20.000
	3	630	63.000
	4	148	14.800
NO_OF_DEPENDENTS	1	845	84.500
	2	155	15.500
TELEPHONE	1	596	59.600
	2	404	40.400
FOREIGN_WORKER	1	963	96.300
	2	37	3.700
End of Table			

Shows the survival function at the mean of covariates.

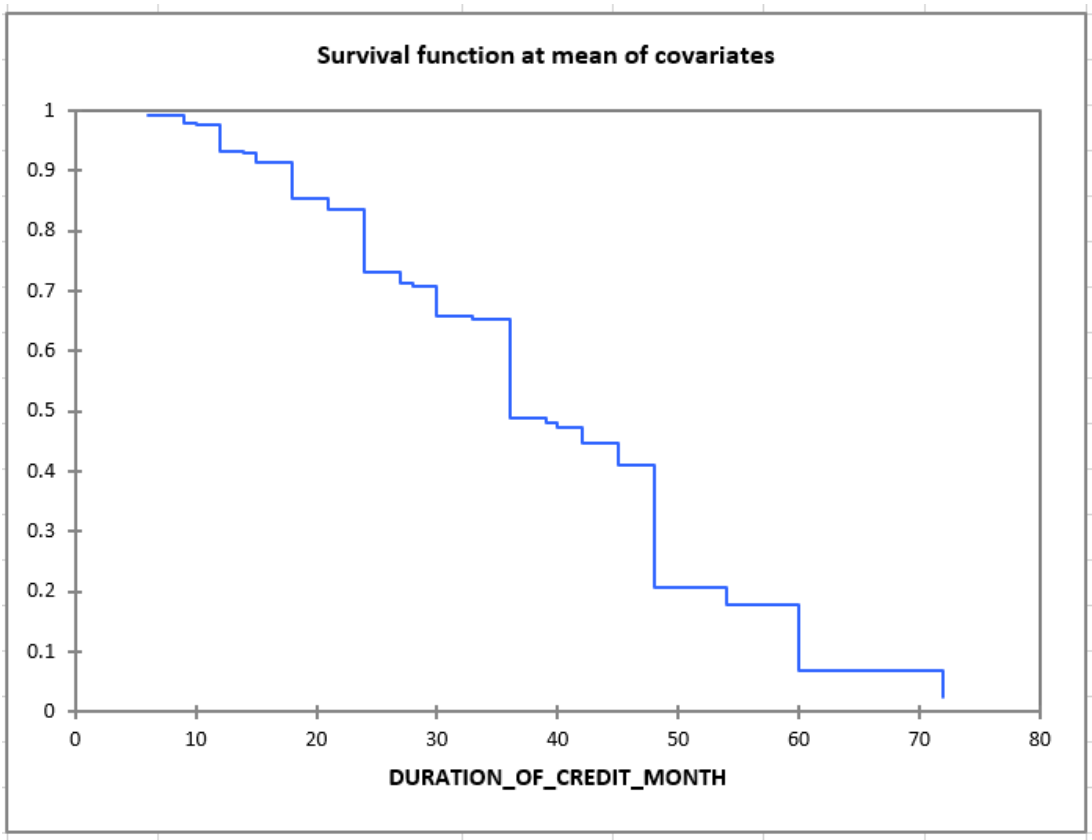

Figure 5.1: Survival function at mean of covariates

Shows the hazard function at the mean of covariates.

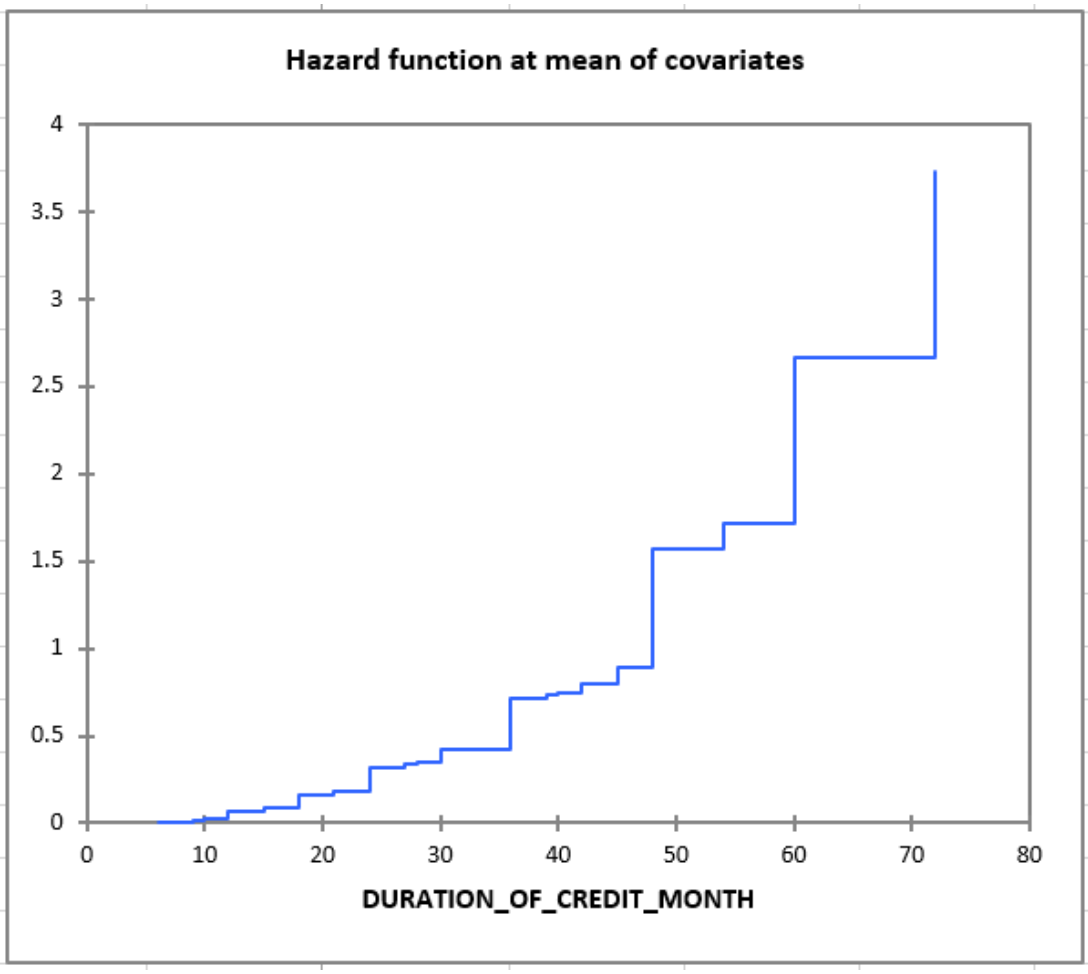

Figure 5.2: Hazard function at mean of covariates

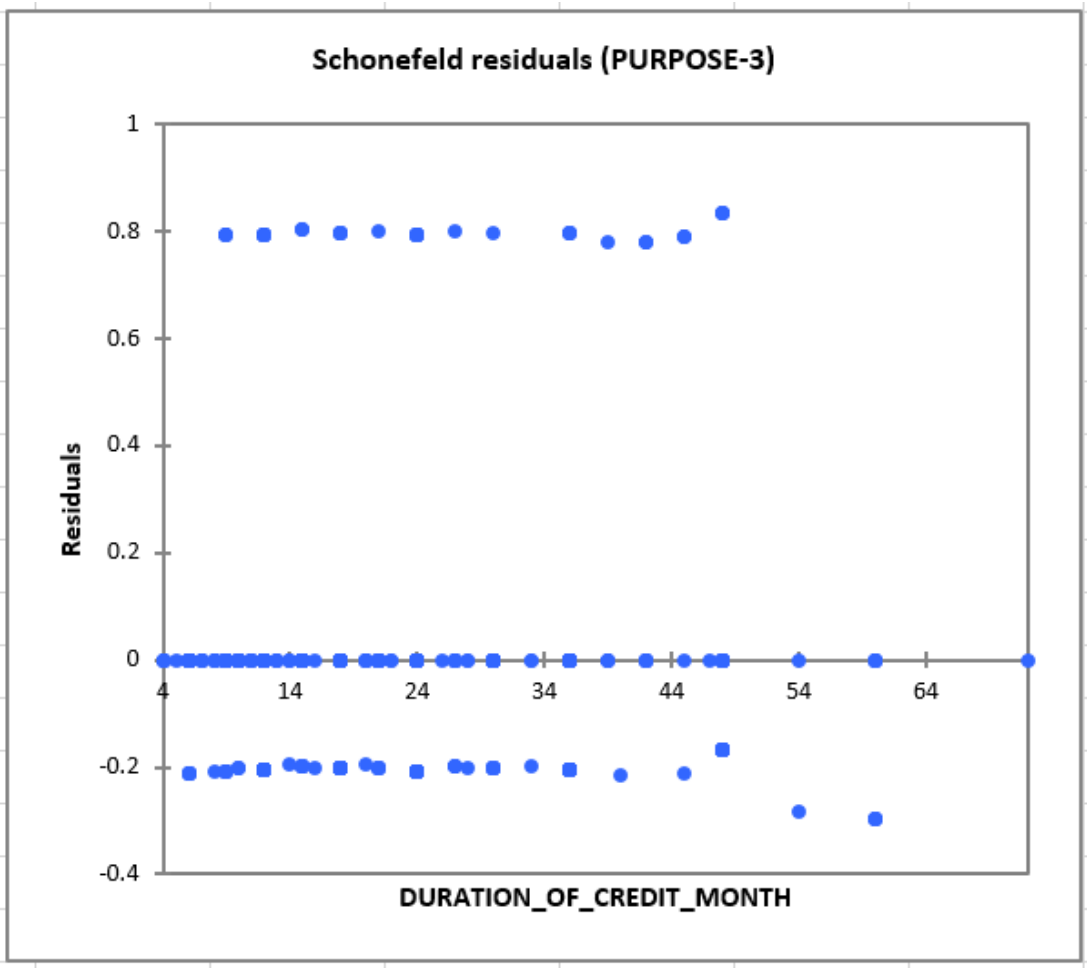

Figure 5.3: Schoenfeld residuals for purpose

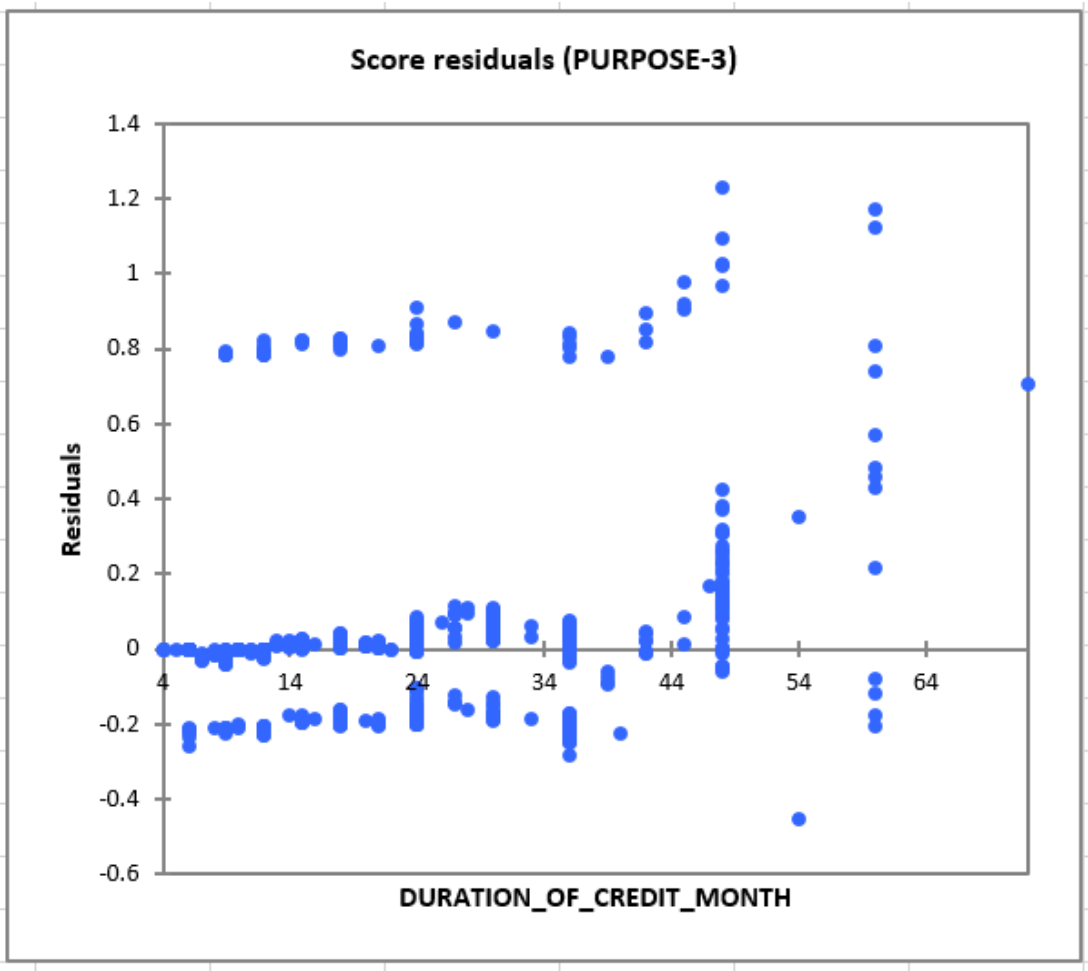

Figure 5.4: Score residuals for purpose

Table 5.2: Quantitative variables summary statistics improved cox.

Begin of Table			
Variable	Categories	Frequencies	%
PREVIOUS_CREDITS_PAYMENTS	1	89	8.900
	2	911	91.100
HAS_BEEN_EMPLOYED	1	62	6.200
	2	938	93.800
SEX	1	92	9.200

Continuation of Table 5.2			
Variable	Categories	Frequencies	%
	2	908	90.800
LOAN_CLASSIFICATION	1	116	11.600
	2	696	69.600
	3	188	18.800
ACCOUNT_BALANCE	1	274	27.400
	2	269	26.900
	3	63	6.300
	4	394	39.400
PURPOSE	0	234	23.400
	1	103	10.300
	2	181	18.100
	3	280	28.000
	4	12	1.200
	5	22	2.200
	6	50	5.000
	8	9	0.900
	9	97	9.700
	10	12	1.200
VALUE_SAVINGS_STOCKS	1	603	60.300
	2	103	10.300
	3	63	6.300
	4	48	4.800
	5	183	18.300
INSTALMENT_PER_CENT	1	136	13.600
	2	231	23.100
	3	157	15.700
	4	476	47.600
GUARANTORS	1	907	90.700
	2	41	4.100
	3	52	5.200
DURATION_IN_CURRENT_ADDRESS	1	130	13.000

Continuation of Table 5.2			
Variable	Categories	Frequencies	%
	2	308	30.800
	3	149	14.900
	4	413	41.300
MOST_VALUABLE_AVAILABLE_ASSET	1	282	28.200
	2	232	23.200
	3	332	33.200
	4	154	15.400
CONCURRENT_CREDITS	1	139	13.900
	2	47	4.700
	3	814	81.400
TYPE_OF_APARTMENT	1	179	17.900
	2	714	71.400
	3	107	10.700
NO_OF_CREDITS_AT_THIS_BANK	1	633	63.300
	2	333	33.300
	3	28	2.800
	4	6	0.600
OCCUPATION	1	22	2.200
	2	200	20.000
	3	630	63.000
	4	148	14.800
NO_OF_DEPENDENTS	1	845	84.500
	2	155	15.500
TELEPHONE	1	596	59.600
	2	404	40.400
FOREIGN_WORKER	1	963	96.300
	2	37	3.700
End of Table			

Shows the survival function at the mean of covariates.

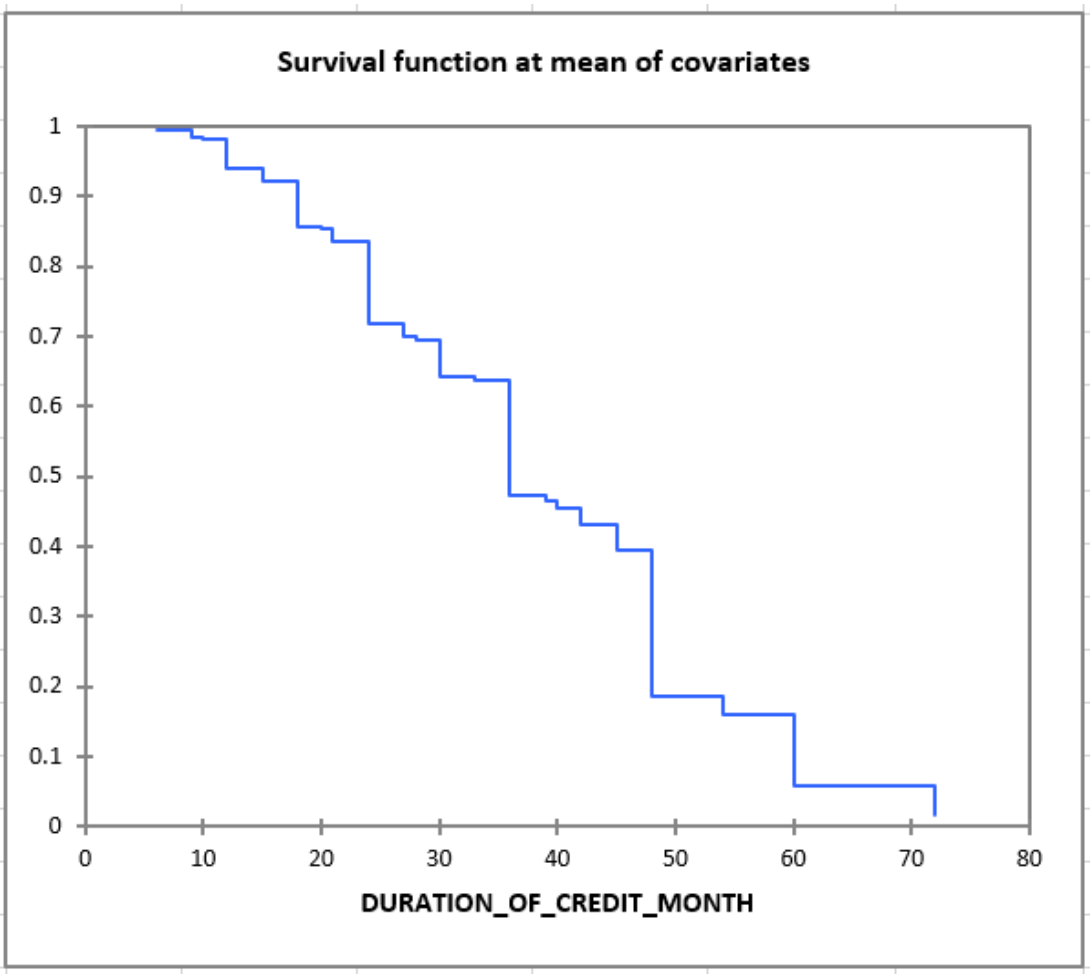

Figure 5.5: Survival function at the mean of covariates

Shows the hazard function at the mean of covariates

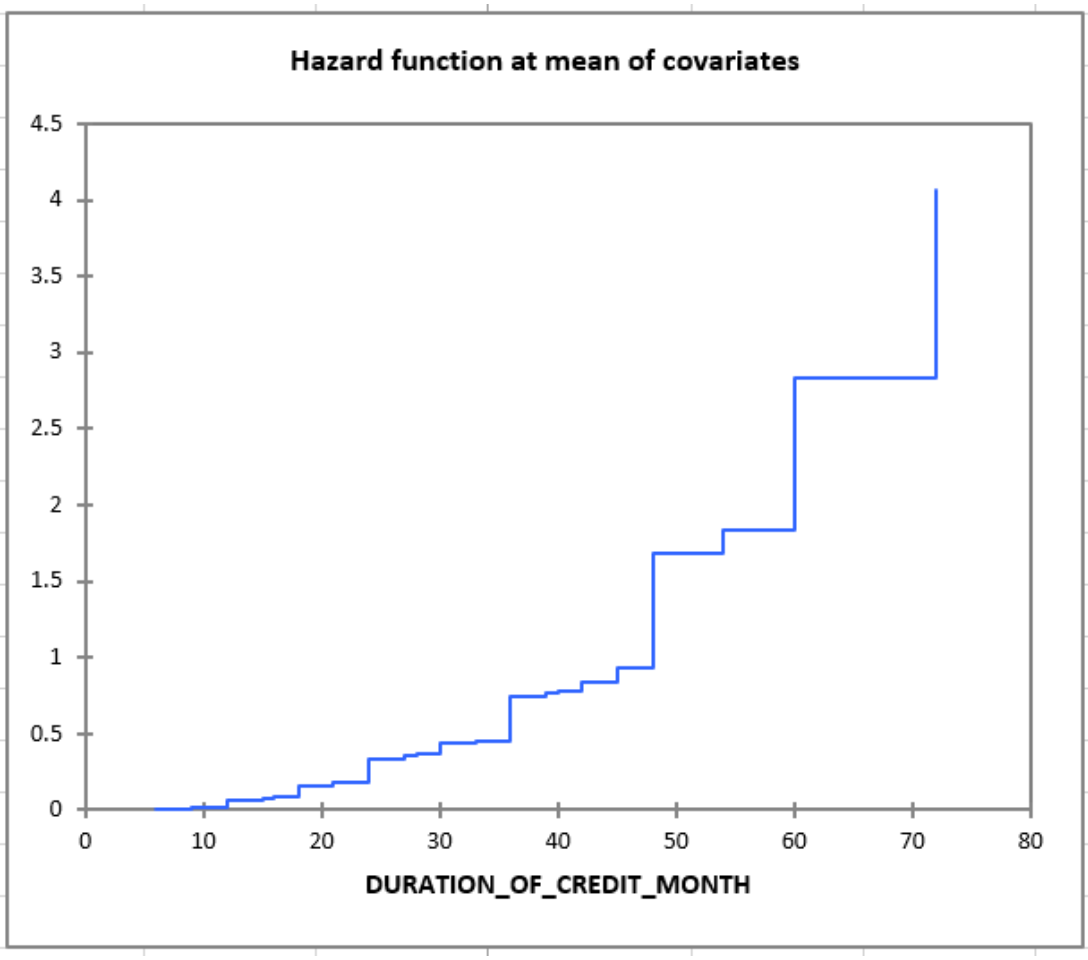

Figure 5.6: Hazard function at the mean of covariates

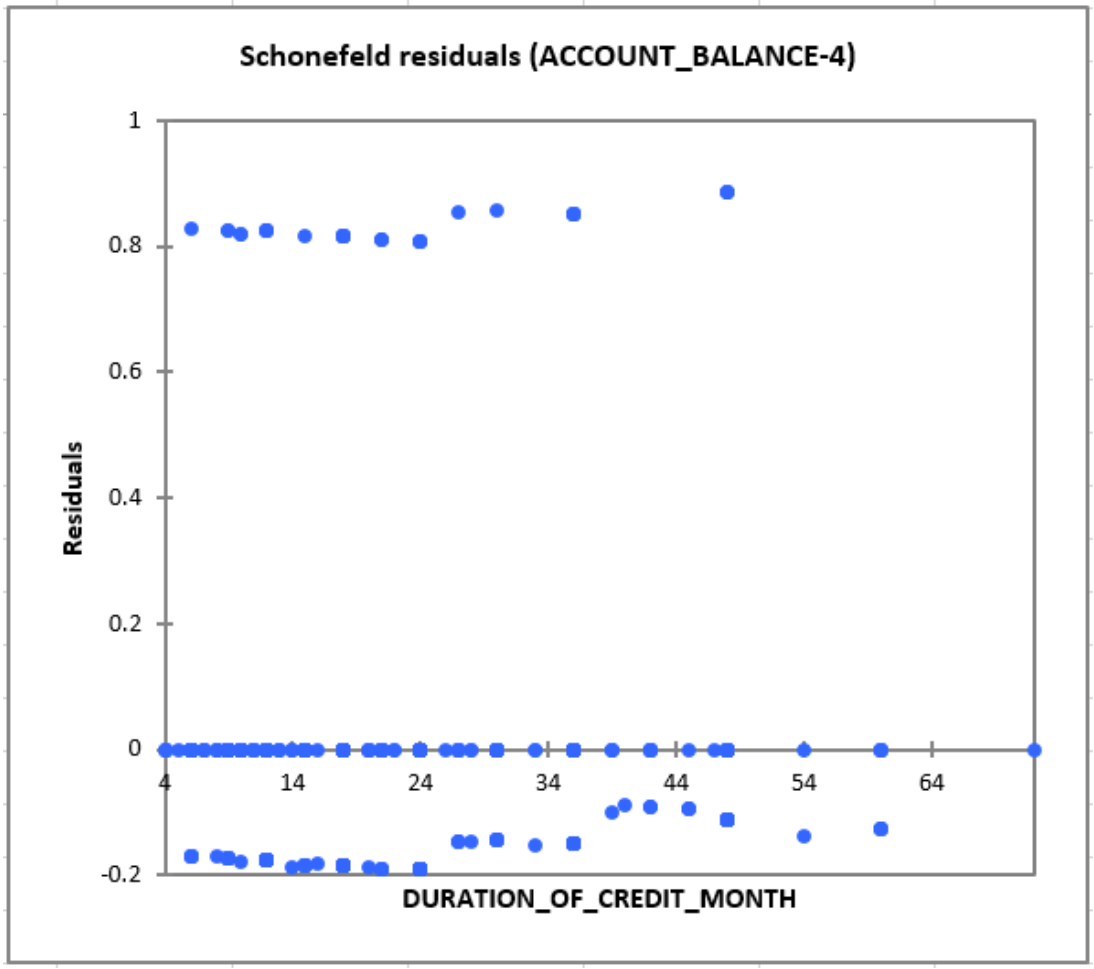

Figure 5.7: Schoenfeld residuals for account_balance_4

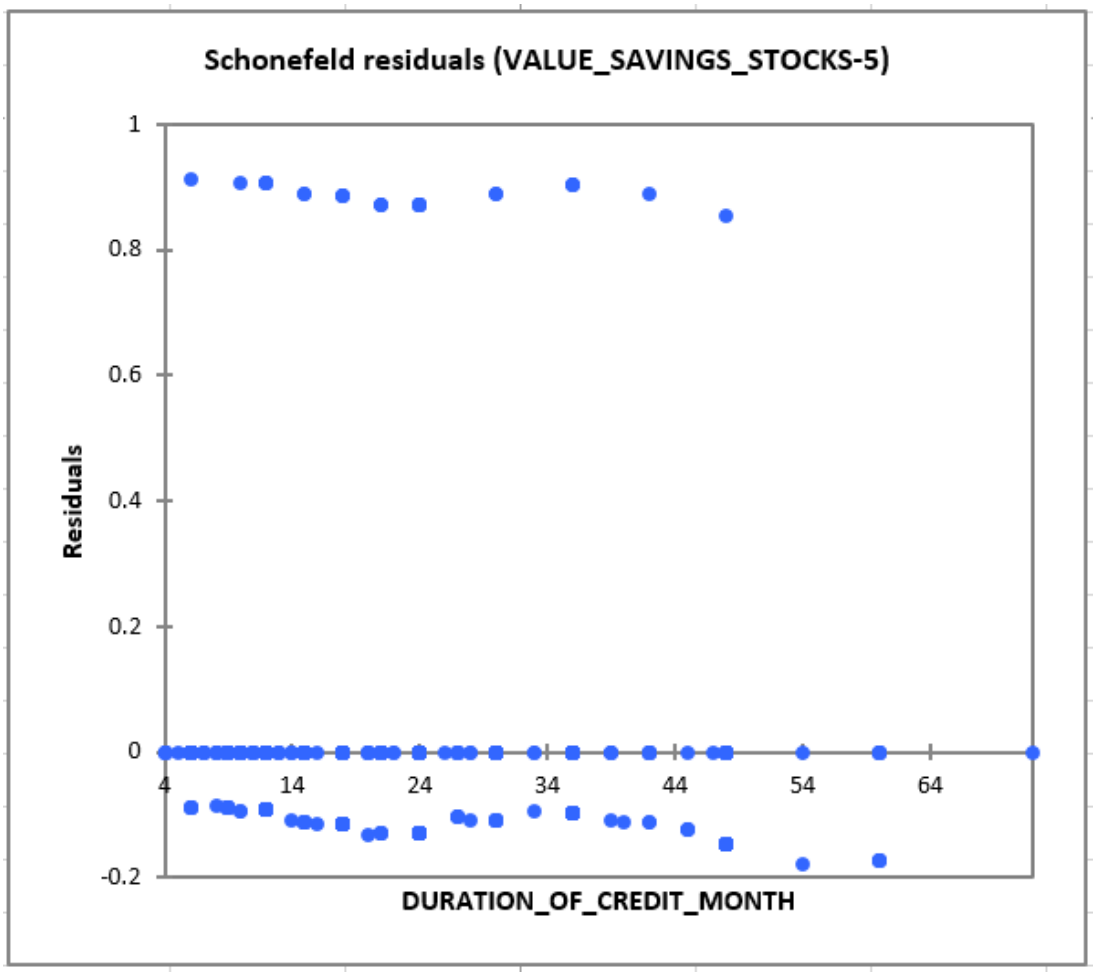

Figure 5.8: Schoenfeld residuals for value_saving_stocks_5

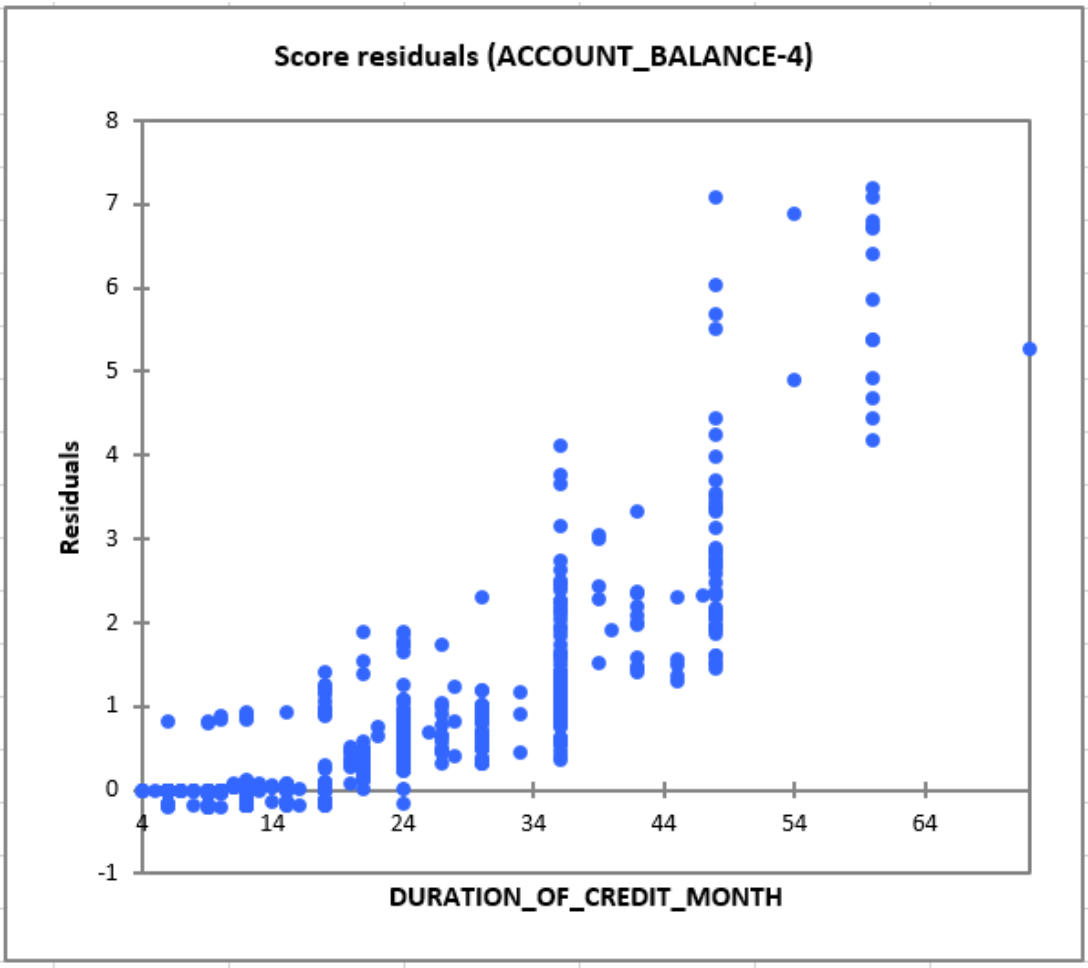

Figure 5.9: Schore residuals for account_balance_4

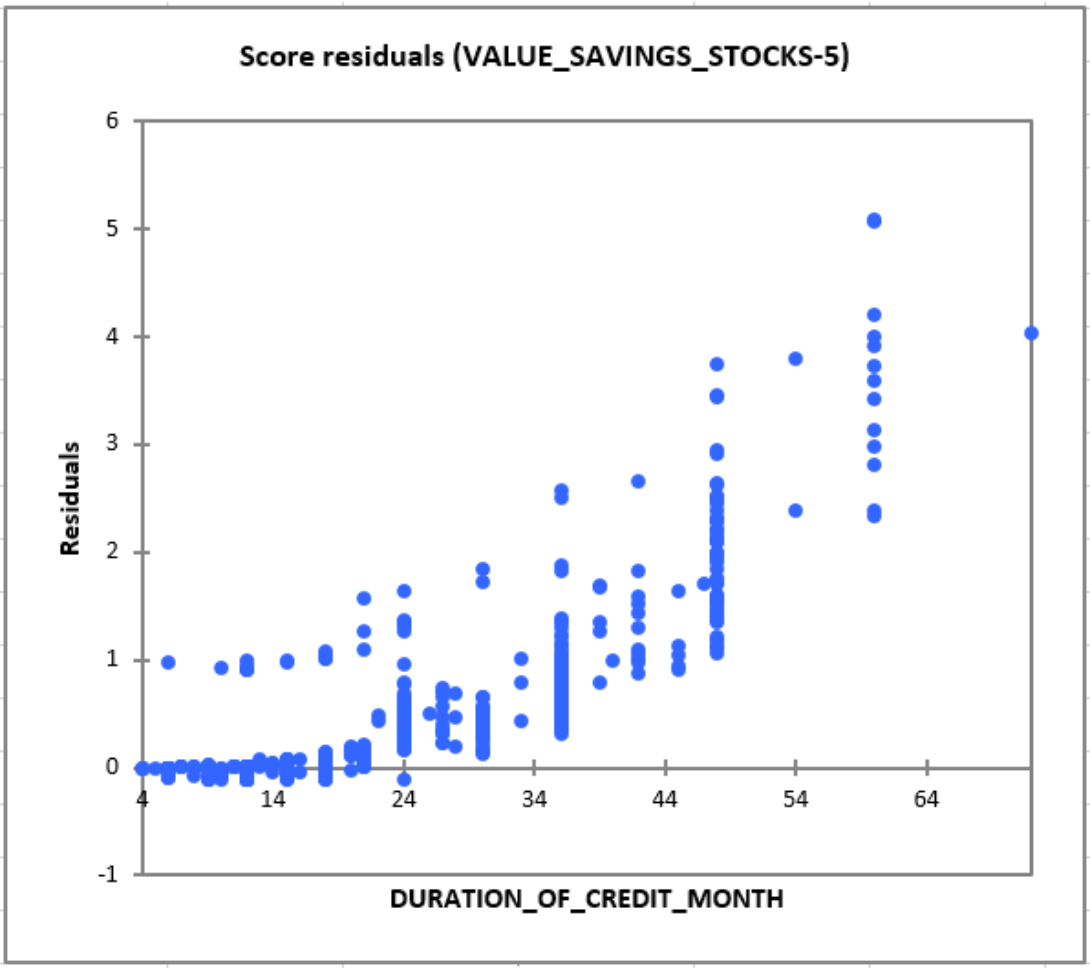

Figure 5.10: Score residuals for value_saving_stocks_5