THE UNIVERSITY OF NAIROBI

SCHOOL OF MATHEMATICS

MASTERS OF SCIENCE IN SOCIAL STATISTICS

OBUDA FELIX YALA

REG. NUMBER: I56/74741/2014

SUPERVISOR: DR NELSON OWUOR

Analysis of Credit Risk on Bank Loans Using Cox's Proportional Hazards Model

A research thesis submitted to the School of Mathematics in partial fulfilment of the requirements for the award of degree of Masters in Social Statistics

2016

0.1 DECLARATION

This thesis is my own personal work and has not been presented for award of any degree.

Signature.....Date.....

The thesis has been submitted for examination with my approval as university supervisor.

Signature.....Date.....

0.2 ACKNOWLEDGEMENT

My first and foremost appreciation goes to the Almighty God for enabling me to successfully undertake this thesis. It was not easy but His grace was always sufficient and His guidance enabled me to complete the work.

My special thanks to my family especially my mum Monica, wife Victoria and daughter Rael for their encouragements, emotional support and prayers that were always a constant reminder that this had to be done and that there was no room for failure.

My heartfelt appreciation to my supervisor Dr Nelson Owuor for his consistent guidance throughout this exercise and my classmates especially Victor Ouma, Simion Bichanga and Liz Mueni for their availability for consultation and discussions throughout the process.

It was not easy but your presence made it possible. May the Almighty God richly bless and do you good all the days of your lifes.

0.3 ABSTRACT

Credit business is the leading income generating activity for banks and other financial institutions. However, it involves huge risks to both the banks and the borrowers. The risk of a trading partner not fulfilling his or her obligation as per the agreement on due date or any time before the due date can greatly jeopardize the smooth functioning of a financial institution.

All credit scoring models contain similar fundamental concepts. Lately, survival analysis has been used to give time angle to the event of interest. It is the area of statistical applications that deals with the analysis of lifetime data. The variable of interest is the time to the occurrence of the event of interest. It is ordinarily utilized in medical studies to test the impact of a drug in the lifetime of a patient and in engineering to study reliability aspects of a certain machine.The advantage with survival analysis is that the model is capable of including censored and truncated data in the development sample. The most common form is right censoring which states that the event is not observed within the study period e.g a customer who does not default.

Credit risk is one of the greatest concerns to most lending institutions. This paper is aimed at coming up with a model that can be used by the banks and other credit advancing institutions to calculate the risk associated with credit advancement. Cox's proportional hazards modelling is applied in the generation of this model since its the most suitable for survival data when proportional hazards has been proved in various groups. Since the paper is interested in customers who default on their loans when they are due, those who pay in time and those who have paid in advance are censored to isolate them from analysis of those who receive events of interest.

The Cox's PH model has been preferred due to its flexibility and robustness in determining hazard risks. It is different from the normal PH model because no assumptions are made on its baseline hazard function . This is the non-parametric part of the model. It however makes assumptions on its parametric part, the part containing the effect of the predictor variables on the hazard. The model is therefore semi-parametric as it contains both non-parametric part and the parametric part and it estimates the relative risk rather than the absolute risk. It is also capable of handling discrete and continuous measures of event times and its possible to incorporate time-dependent covariates over the course of the observation period.

Credit scoring models calculate a persons credit score primarily from information contained in his application form including age, time with bank (months), gender, employer details, county of birth, basic salary, credit history (number of loans taken with the bank), type of employment (contract, permanent, casual), amount, repayment period time, time at current job (months), marriage status and purpose of loan. The persons payment history reflects the various accounts that he has, including credit cards, mortgage advances and deposit accounts. Collections, foreclosures and lawsuits are categorized as factors and given a weight (Credit Risk Scoring Analytics, Issue No: 0710511).

Using survival analysis procedures for building credit hazard models is not new. It started with the paper by Narain (1992) and was then developed by Carling et al. (1998), Stepanova (2002), Roszbach (2003), Glennon (2005), Allen (2006), Goko (2006), Malik (2006) and Djadja (2007). A typical feature in all these papers is that they utilize parametric or semiparametric regression strategies for analysing the time to default, such as exponential, Weibull and Cox's proportional hazards models, which are ma-

jorly used in this kind of study. The model established for the time to default is then used to analyze Probability to Default (PD).

This paper proposes a basic idea to estimate PD, which is performed using the Cox's proportional hazards model. Some random right censoring mechanisms appear in the model and so survival analysis techniques become natural tools for use.

The conditional survival function utilized to model credit risk opens a remarkable perspective to studying loan defaults. Instead of looking at whether loan defaults or not, the paper looks at the time to default, given credit information of clients (endogenous variables) and considering the indicators for the economic status (exogenous variables). Therefore, the default risk is measured using the conditional distribution of the random variable time to default, T, given a vector of covariates, X.

Keywords: Cox's Proportional Hazards Model, Survival analysis, Censoring, Credit scoring, Hazard function

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Background

Several factors led to the introduction of automated credit scoring in the 1940s as indicated by Durand (1941). Around 1945, there was broad interest in credit indicating that the subjective methods had become obsolete and could no longer scale well to substantial numbers of applicants. The credit explosion, stimulated by the introduction of credit cards, inspired lenders to automate the credit analysis process leading to introduction of objective credit scoring systems.

There is no specific instrument that can be utilized to foresee the future precisely, yet when analysing loans, the banks attempt to predict the result of that loan. They have to find the likelihood of a customer either defaulting on that loan or not. Like all obligation instruments, a credit involves the redistribution of monetary resources after some time, between the bank and the borrower. The borrower at first gets a measure of cash from the lender, which he pays back, however, not necessarily in equal measures.

Credit scoring utilizes quantitative measures of features and performance of previous loans to foretell the future performance of credits with similar characteristics (Caire & Kossman, 2003). Credit scoring is a logical method for evaluating the credit hazard related with new credit applications. Statistical models determine prescient connections between application data and the probability of satisfactory repayment. They are empirically outlined; that is, they are developed completely from information obtained through prior experience.

Credit scoring is therefore a target hazard appraisal apparatus, instead of subjective techniques that depend on loans officer's assessment. It is a risk administration tool preferred for managing risks associated with financial advances. Scoring frameworks help banks in ensuring consistency in underwriting activities and can also provide the executive management with an insightful measure of credit risk pertaining to specific credits as well as the general overview of their loan book.

Current credit scoring frameworks perform the same functions, though in an objective way. Suppose we classify a certain population of customers into two groups, good and bad, the information extracted from the customers application form when they apply for a loan is then used by financial institution to determine which group the customer should belong. Instead of being examined in a subjective way, the information is shaped to form quantitative variables that are then inserted into a statistical model. For an individual, suppose there are k covariates, they are outlined as a vector, to form the input to the model. The covariates can then be used to produce a

score to estimate the probability, p, of that individual belonging to either the good or bad group. The relationship between the covariates and the probability of default is found by fitting the variables to a predefined model.

Credit scoring techniques were initially developed to assist organisations in automating the credit analysis process. Thus, the primary aim of conventional credit scoring system was to classify potential customers as either good or bad, so as to enable appropriate decision to be made concerning their loan applications. A bad customer would be defined as one who fails to repay the loan in full within a certain period, but this definition can be expanded to cover a range of undesirable behaviour. Rosenberg (1994), Henley (1997), outlined the diverse systems that can be built from such models. The meaning of "bad" in this context can somehow be arbitrary and is often driven by regulatory guidelines. While the definition can incorporate early settlement or fraudulent activities, the most popular definition of "bad" is default. Default could be interpreted as one or two payments ignored, three consecutive unpaid payments, or even when the loan becomes unrecoverable. If the definition of "bad" is excessively stringent, or not sufficiently stringent, it may negatively affect the quality of the credit scorecard (Siddiqi, 2005).

Credit scoring cannot foretell individual credit loss; instead, it predicts the probability or chances of a bad outcome, as predefined by each financial institution; usually this is some average or total days in arrears that leads to the loans becoming unprofitable. A credit scoring system should be able to affirm or reject a loan application; rather the underwriter must make

the decision whether he or she will include the credit score into the loan appraisal.

Lastly, credit scoring is not intended to improve approval rates; rather, it advances consistency and efficiency while maintaining or reducing historic delinquency rates. It likewise permits the clients to focus their consideration and time on applications that are not clear approvals or obvious declines (Caire & Kossman, 2003). Hence the research aims at coming up with a model that can be used in calculating the risk associated with personal loans.

1.2 Statement of the Problem

Decisions on whom to grant credit, and how much, originally relied mainly on the personal judgement of a credit officer. The officer used his skills and experience, guided by attributes that affect the credit soundness of the applicant, he then would decide whether or not to grant a loan. The attributes are referred to as the five Cs of credit (Thomas et al., 2002). They are:

- Character - The eagerness to pay obligation. For instance, to what extent has the candidate been at their present place of employment?
- Capacity - The borrower's ability to pay the obligation. Compensation and other wages are significant determinants here.
- Collateral - Possessions that may be utilized to secure the obligation

are classed as guarantee. For a home loan, the home obtained is utilized as insurance.

- **Capital - A very much resourced individual will probably be conceded an advance.**
- **Conditions - Current and anticipated monetary conditions are additionally checked.**

Various factors have prompted the need to progress automated credit scoring particularly the explosion in the demand for credit and it turns out that the subjective strategies do not scale well to substantial numbers of loan applicants. Since the year 2000, the Kenyan financial market has encountered developing liquidity, which has prompted fast development of banks hence dynamic loan products. This has promoted the need to review the credit granting criteria to reflect the development in volumes of loan portfolio and to react to the current worldwide credit crunch. Surprisely, research on credit risk has not received significant attention from both bank experts and the general researchers in Kenya as it ought to receive.

Throughout the years, banks have perpetually used customary credit scoring procedures to rate loan applicants. Various studies have been done on the issue of credit risk analysis using diverse approaches. Besides, existing credit scoring models classify borrowers into various risk classifications but cannot predict when the borrower is likely to default. It is more useful for the lender not only to know the probability of defaulting but also when the default is likely to happen. This helps to analyse risks and improve the focus on optimized profitability. For instance, if the bank knows that some loan applicants are the bad type, instead of rejecting their loan applications, it

may grant loans to them at higher interest rates, as long as the term of the loan is shorter than the likely time to default. Thus some bad applicants can also be viewed as profitable propositions.

Therefore, credit risk is one of the major concerns to the financial institutions and regulators. Determining the probability of default, PD, in consumer credits has to be addressed by the banks and other financial institutions. This is the first step needed to compute the capital in risk of insolvency, when their clients do not pay their loans, a scenario referred to as default. The risk coming from this type of scenario is called credit risk, which is the subject of research for this study.

1.3 General Objective

The broad objective of this research is to use Cox's regression model to determine a combination of potential explanatory variables that influences the hazard function and evaluate the effects of these explanatory variables on the hazard function.

1.4 Specific Objectives

The specific objectives include:

- Develop a model for credit risk

- **Determine a combination of potential explanatory variables that influences the hazard function**
- **Evaluate effects of the explanatory variables on the hazard function**

1.5 Research Questions

- **What is the Coxs PH model for credit risk?**
- **Which are the potential explanatory variables that influence the hazard function?**
- **What are the effects of the explanatory variables on the hazard function?**

1.6 Justification of the Study

The timing of loan defaults has important inferences on loan recovery rates and lender profitability. This paper evaluates the effects of borrower characteristics on default probability changes over the lifespan of a loan. Standard observational models of loan performance examine whether borrower characteristics, such as credit score, increase or decrease the probability that a credit will default on average.

However, the effect of typical credit-worthiness indicators may change over the lifespan of a loan if, for example, defaulters behave in systematically different ways. Neglecting to model, these differences may lead researchers to conclude that there is no impact, when in fact, solid effects exist for

defaulters but have diverse signs.

1.7 Scope of the Study

The study intends to use a sample of 1,000 applicants in this study to achieve the objectives aforementioned. It will employ Coxs regression to estimate probability to default and also determine the survival time estimators.

The applicants who offset their loans fully before the loan period will be censored. Those whose loans run the full time and are paid in full will also be censored to enable the researcher concentrate on defaulters.

1.8 Limitations

The biggest challenge with this study was acquisition of data for study. Since bank records are deemed to be sensitive, it was a challenge convincing the concerned to provide data. Even when it was provided, there were a lot of constraints on what to report on and what not to report.

There has been a debate on when a loan actually becomes "bad" since the definition differs from one financial institution to another. However, the Central Bank of Kenya Prudentials defines "bad" loans as those unpaid for more than 90 days. The researcher used this provision in this research work.

Chapter 2

LITERATURE REVIEW

2.1 Introduction

A loan's time to default is important both to the lender and the whole financial industry. Numerous research works have been done on loans and the time they take to default including reasons why responsible and honest customers would decide to ignore their responsibilities in payments of loans.

It started with the introduction of computerized credit scoring in the 1940's and as indicated by Durand (1941), towards the end of World War II, there was an outburst in the demand for loans and it became clear that the subjective approaches did not scale well to the substantial numbers of applicants.

2.2 Modeling consumer credit risk via survival analysis by Ricardo Cao, Juan M Vilar & Andres Devia

This paper performs three different methods to estimate the probability to default of a borrower. The first one is built on Cox's proportional hazards

model; the second one utilizes generalized linear models, while the third one uses a random design nonparametric regression model. In all the cases, some random right censoring procedure appears in the model, making survival analysis techniques the natural tools to be used.

The conditional survival function utilized in modelling credit hazards opens an interesting perception to studying defaults. Instead of looking at default or not, we look at the time to default, given credit data of clients (endogenous variables) and considering the indicators for the economic status (exogenous variables). Therefore, the default hazard is measured using the conditional distribution of the random variable time to default, T, given a vector of variables, X. The variable T is not completely observable due to the censoring procedure.

Cox's proportional hazards methodology is used to analyze the conditional survival function $S(t|x)$. The crucial point in this method rests on the approximation of the cumulative conditional hazard function, $(t|x)$, using maximum likelihood. The main interest is to build a conditional model for the individual $PD(t|x)$, which is defined in terms of $(t|x)$. The use of Coxs proportional hazards model depends on Narain (1992) method for approximation of $S(t|x)$.

In the application of a generalized linear model, it is assumed that the lifetime distribution:
y **where** $\theta = (\theta_2, \theta_3, \cdots, \theta_{p+1})$ **is a p-dimensional vector and g is the link function, like the logistic or the probit function. Therefore, this model repre-**

sents the conditional distribution of the lifespan of a credit, T, in terms of the unknown parameters. Once these parameters are estimated, an estimator of the conditional distribution function is obtained and an estimator of PD can also be worked out.

In the implementation of a nonparametric conditional distribution estimator, the estimator proposed by Beran (1981) for conditional survival function is utilized. To be able to estimate the probability of default at time t given a covariate vector x, we switch the theoretical value of the conditional survival function with its estimator.

2.3 Censored Regression Techniques for Credit Scoring by Thandekile Hlongwane, Precious Mdlongwa, Hausitoe Nare & Isabel L Moyo

This paper uses linear regression models to generate a credit scoring framework, assuming a linear model where the probability p that an applicant will default is related proportionally to k explanatory variables:
$p = \beta^T x = \beta_1 x_1 + \cdots + \beta_k x_k$
where β is the vector of parameters $(\beta_1, \beta_2, \cdots, \beta_k)$.

The research was a case study of a commercial bank in Zimbabwe. Credit advances were categorised on a monthly basis as either good or bad where good referred to loans that were not 30 days behind in repayments. Bad referred to those that at the time prior to that month, had been more than

30 days behind in repayments.

The researcher used The Buckley James method to correct the bias present in linear regression with censored data by replacing censored points with their expected values. He then engaged Monte Carlo Simulation method to compare Linear and Buckley James regression and then select the best technique to use in calculating default risk in credit advancements.

The data collected included residential status, employment status, marital status, time at address, time at occupation, time at the bank, loan purpose, sex and age. The monthly performance data included whether the loan was still running or not, whether the loan was more than 30 days behind in repayment or if the loan had been fully repaid.

On analysis, scatter plot showed that there is a positive relationship between time to default and age. Older people have a higher default rate compared to younger people with the older being those over 50 years and younger being those below 50 years.

There was no relationship observed between time to default and sex but time to default and marital status were out rightly related. The singles had a higher default rate than the married, probably because the singles are mobile while the married are likely to settle in one place and could also assist each other to pay when one of them is faced with possibility of defaulting due to loss of job or business suddenly performing poorly.

There was no relationship between time to default and time with the bank, meaning that defaulting is not influenced by the time a customer has been with the bank. However, there was a positive correlation between time to default and time at current job. This means that the more one stays in one job, the more he gets stable financially and the less he is likely to default.

The researcher recommended that the bank should observe every credit issued and act as soon as it starts to go bad. The banks should put up a credit risk management team that should be liable for the following actions that will assist in reducing credit risk:

- **Reconstructing the credit score sheet and reassign scores to all the variables that affect defaulting and repayment.**

- **Implementing the Buckley James method, as it proved to be better performing.**

- **Reconsidering the minimum age for a loan applicant, as the study showed that 21 years is not valid for loan application.**

- **Reviewing the customers that fall under single and married in the credit score sheet as there are widows and widowers.**

- **Closely monitoring the loan performance of each customer taking survival analysis into consideration as well.**

2.4 Risk Factors for Consumer Loan Default: A Censored Quartile Regression Analysis by Sarah Miller

The author used Cox's Proportional Hazard Model, proposed by Cox (1974) to predict the effect of borrower characteristics on probability of default, or default risk. The model adopts a parametric time-constant form on the effects of observed characteristics. It required that the covariates shift the baseline hazard up or down monotonically and do not identify changes in the effects over the lifetime of a loan.
The researcher impressed the default probability at time $t, h(t)$, for borrower i with characteristics x_i as:

$h(t) = h_o(t)exp(x_i^T \beta)$ **and after log transformation,**

$log(h(t)) = \alpha(t) + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$

And here, the baseline hazard function,

$\alpha(t) = log(h_o(t))$

varies over time but is the same for all borrowers.

The model does not make any assumption on the baseline hazard function. However, it assumes a parametric time constant form on the properties of observed variables. It assumes that the influence of the covariates is de-

tachable from the temporal effect in $\alpha(t)$ so that the variable x_p shifts the respective default probability up or down subject to the sign of β_p.

Bassett (1978) utilized quartile regression, a technique for modeling conditional quantiles of a dependent variable instead of the conditional mean. In this research work, the researcher uses the same for survival analysis, letting the covariates to have a different impact on default. The author models quantiles of the log of default time, $log(T)$, for borrower i with characteristic x_i as:

$Q_{log_{(T)}}(\tau|x_i) = \alpha(\tau) + \beta_1(\tau)x_{1i} + \cdots + \beta_k(\tau)x_{ki}$ **for a wide range of quantiles** $\tau\epsilon(0,1)$

For example, if the predicted quantile for loan $i, log(Ti)$, is equal to 3 at $\tau = 0.05$ or equivalently, ${}^6T_i = e^3 \approx 20$, meaning loan i has a 5% chance to default by day 20.

2.5 Survival Analysis Methods for Personal Loan Data by Maria Stepanova & Lyn Thomas

In this paper, the authors applied survival analysis techniques to personal data from a major UK financial institution which consisted of 50,000 loan personal applicants with repayment period of 36months.

The authors used Cox-Snell residuals, Martingale residuals and Deviance residuals as diagnostic tools to examine the model fit and outline any discrepancies between the fitted and predicted values.

Cox-Snell residuals (Cox and Snell 1968) are defined as:

$r_{Ci} = exp(\hat{\beta}x_i)H_o(t_i) = H_i(t_i) = -logS_i(t_i)$ **where** $H_o(t_i)$ **is the estimated cumulative baseline hazard,** $H_i(t_i)$ **is the estimated cumulative hazard and** $S_i(t_i)$ **is the estimated survivor function for the** ith **individual at time** t_i**.**

To test that the residuals have exponential distribution, $log(-logS(r_{Ci}))$ **is plotted against** $log(r_{Ci})$ **and a straight line with zero intercept will indicate that the fitted model is correct.**

For these observations, the plotted lines were close to straight line and since they were few, the researcher concluded that repayment after one month is not necessarily a typical or normal feature of a personal loan portfolio. Having ignored the observations, it was concluded that the model fits the data well.

Martingale residual (Therneau el al. 1990) is a transformation of the Cox-Snell residuals and is defined as:

$r_{Mi} = \delta_i - r_{Ci}$

It can be described as the difference between the observed number of the failures and the expected number of failures. r_{Mi} **is plotted against the rank order of time and the residuals should not exhibit any pattern if the model is adequate.**

When Martingale residuals were plotted for these observations, the values appeared in two bands, one representing the uncensored observations and another censored observations. They are always expected to be negative for the censored.

Deviance residual (Therneau el al. 1990) is a transformation of the Martingale residual and is defined as:

$r_{Di} = sgn(r_{Mi})[-2\{r_{Mi} + \delta_i log(\delta_i - r_{Mi})\}]$**. These residuals are preferred where the number of observations are few such as the medical studies. They were therefore not useful for this research work.**

Chapter 3

COX'S PROPORTIONAL HAZARDS MODEL

3.1 Introduction

The Cox's Proportional Hazard Model is the most frequently used multivariate analysis method for analysing survival time data. It is a regression model which describes the relationship between an event occurrence, expressed as hazard function and a set of covariates. The hazard function is the probability that an individual will experience an event within a certain time interval given that the individual has survived to the beginning of that interval.

Cox's regression is similar to multiple regression except that in Cox's, the dependent variable is the hazard function at a given time. If several explanatory variables are involved, then the hazard or risk can be expressed as:

$h_i(t) = h_o(t)exp\{\beta' x_i\}$

The quantity $h_o(t)$ is the baseline hazard function and corresponds to the probability of reaching an event when all the explanatory variables are zero.

A prominent feature of the Cox's model is that the baseline hazard function is measured non-parametrically, and so unlike most other statistical models, the survival times are not assumed to follow a particular statistical distribution and the t in $h(t)$ indicates that the hazard function may vary over time.

Some of the features of the Cox's PH model includes:

- **It is a product of a function in t and a function in x**
- **x is time independent**
- **The baseline hazard $h_o(t)$ does not depend on x but only on t**
- **The exponential involves the $x's$ but not t**
- **It follows the proportional hazard assumption**
- **The estimated hazards are always non-negative**

3.2 Survival Function

This refers to the probability that a system or an event under investigation will last beyond a specified time. The survival function of the Cox's PH model is given by:

$S(t|x) = exp[-H_o(t)exp(\beta^{'}x)]$

where

$H_o(t) = \int_0^t h_o(u)du$

is the cummulative baseline hazard function.

The distribution function of the Cox model is:

$F(t|x) = 1 - exp[-H_o(t)exp(\beta)'x]$

Let Z be a random variable with distribution function F, then $G = F(Z)$ follows a uniform distribution on the interval from 0 to 1, abbreviated as $G \sim Uni[0,1]$. If $G \sim Uni[0,1]$, then also $(1-G) \sim Uni[0,1]$. Thus if we let T be the survival time of the Cox model in the equation above, then it follows that $G = exp[-H_o(T)exp(\beta'x)] \sim Uni[0,1]$

If $h_o(t)0$ for all t, then H_o can be inverted and the survival time T of Cox's PH model can be expressed as:

$T = H_o^{-1}[-log(G)exp(\beta'x)]$

Where G is a random variable with $G \sim Uni[0,1]$. The above equation is suitable for the generation of survival times with the random numbers following a uniform distribution.

3.3 Hazard Function

When $t \geq 0$, the Cox's model follows an exponential distribution with the hazard function given by:

$$h(t|x) = h_o(t)exp(\beta' x)$$

The baseline hazard function $h_o(t)$ is constant and the model generates exponentially distributed survival times with scale parameters dependent on the regression coefficients and the considered covariates.

3.4 Maximum Likelihood Estimation of the Cox's PH Model

Parameters estimates in this model are acquired by maximizing the partial likelihood as opposed to the likelihood itself.

Assuming that n is the number of individuals observed, r is the number of failure times (events of interest) with exactly one failure at each time; $t_1 < t_2 < \cdots < t_r$

The partial likelihood is given as:

$$L(\beta) = \prod_{i=1}^{r} \frac{exp(x_i\beta)}{\sum_{j=1}^{n} exp(x_j\beta)}$$

The log likelihood is given as:

$$l(\beta) = \sum_{i=1}^{r} \{x_i\beta - log[\sum_{j=1}^{n} exp(x_j\beta)]\}$$

It has been proven that this partial log likelihood can be treated as an ordinary log-likelihood to obtain partial maximum likelihood estimators of β. It will assist to derive consistent and efficient estimators of β.

Suppose, we want to incorporate the censored times to make it complete. Let $\delta_i = 1$ if i^{th} individual default on his loan and $\delta_i = 0$ if i^{th} individual pays his loan in full within the loan period and $i = 1, 2, \cdots, n$ (number of individuals), $j = 1, 2, \cdots, r$ (event times) and $r \leq n$, then

$$L(\beta) = \prod_{i=1}^{n} \{ \frac{exp(x_i \beta)}{\sum_{j=1}^{n} exp(x_j \beta)} \}^{\delta_i}$$

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

For partial log likelihood, the equation becomes:

$$l(\beta) = log(L(\beta)) = \sum_{i=1}^{n} \delta_i \{ x_i \beta - log[\sum_{j=1}^{n} exp(x_j \beta)] \}$$

We then maximize $l(\beta)$ with respect to β to obtain the partial maximum likelihood estimators of the β.

This is done by use of Newton Raphson algorithm since the equation generated is non-linear.

Newton Raphson algorithm was developed by Isaac Newton and Joseph

Newton to assist in finding successively better approximations to the roots of real valued functions.

It takes the form:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

It is iterated until a stable and desirable result is achieved.

3.5 Proportional Hazards Assumption in the Covariates

The explanatory variables multiplicatively act on the hazard at any point in time to provide the key assumption of the proportional hazard model that the hazard of the event in any group is a constant multiple of the hazard in any other group.

To test this, we use the Schoenfeld Residuals. The residuals are based on the individual contributions to the derivative of the log likelihood (Hosmer and Lemeshow 1999, 198)

Assume p covariates and n independent observations of time, covariates and censoring which are represented as (t_i, x_i, c_i) where $i = 1, 2, \cdot, n$ and $c = 1$ for uncensored observations and zero otherwise.

In deriving the Schoenfeld residuals, we get the derivative of the k^{th} covariate:

$\frac{dL_p}{d\beta_k} = \sum_{i=1}^{n} c_i\{x_{ik} - \frac{\sum_{j=1}^{r} x_{jk} e^{x'_j\beta}}{\sum_{j=1}^{r} e^{x'_j\beta}}\}$

The residual is the difference between the observed and the expected values of the covariates at each failure time.

$\frac{dL_p(\beta)}{d\beta_k} = \sum_{i=1}^{n} c_i\{x_{ik} - \bar{x}_{w_i k}\}$

According to Hosmer and Lemeshow (1999, 198), the estimator of the Schoenfeld residual for the i^{th} subject on the k^{th} covariate are then obtained by substituting the partial likelihood estimator of the coefficient, $\hat{\beta}$.

$\hat{r}S_i k = C_i(x_{ik} - \hat{x}w_i k)$

If the residual exhibits a random (unsystematic) pattern at each failure time, then this gives evidence that the covariate effect is not changing with respect to time. This is exactly the proportional hazard assumption.

If it is systematic, it suggests that as time passes, the covariate effect is changing, violating the proportional hazard assumption.

Chapter 4

DATA ANALYSIS AND INTERPRETATION

4.1 Data Simulation

We generated a data set of bank loans. The data set contains client description variables personal information about the client (sex, age, gender, county of birth, employer, type of employment, employer name, basic salary) and a description of the loan (amount, repayment period). The data used in the study was randomly generated using sampling technique for the variables of interest. The variables are as listed in the table below:

Table 4.1: Table of Variables

No.	Variable	Levels	Data Type	Factor or Integer
1	Amount	Open	Numeric	Integer
2	Age	Open	Numeric	Integer
3	Gender	0=Female, 1=Male	Categorical	Factor
4	County of Birth	0=Garissa, 1=Kisii, 2=Kakamega, 3=Meru, 4=Nyeri, 5=Machakos, 6=Kajiado, 7=Kericho, 8=Kisumu, 9=Mombasa	Categorical	Factor
5	Employer	0=Armed Forces, 1=Min of Health, 2=Min of Education, 3=County Govt, 4=National Govt (others)	Categorical	Factor
6	Marital Status	0=Divorced/Widowed, 1=Single, 2=Married	Categorical	Factor
7	Repayment Period	4 i.e 36, 48, 60 and 72	Numeric	Integer
8	Basic Salary	Open	Numeric	Integer
9	Type of Employment	0=Contract, 1=Permanent	Categorical	Factor
10	Time taken to clear loan	Open	Numeric	Integer

4.2 Assumptions of Cox's Proportional Hazard Model

The Cox model is a multiple linear regression of the logarithm of the hazard on the variables x_i, with the baseline hazard being an intercept term that varies with time. The explanatory variables multiplicatively act on the hazard at any point in time to provide the key assumption of the proportional hazard model that the hazard of the event in any group is a constant multiple of the hazard in any other group. It implies that the hazard curves for the groups should be proportional and cannot cross each other. The

proportionality assumption should always be tested to verify that it exists since it is important for the analysis of survival time data.

4.3 Proof of Proportional Hazards assumption

The proportionality assumption was tested by plotting hazard curves for County of Birth, Gender and marital status. The test was aimed at verifying that the curves do not cross each otherwise it would mean the hazards are unproportional and therefore not suitable for Coxs model analysis.

Figure 4.1: County of Birth Survival Curve

The lines in the curve do not cross each other, meaning the hazards for counties are proportional.

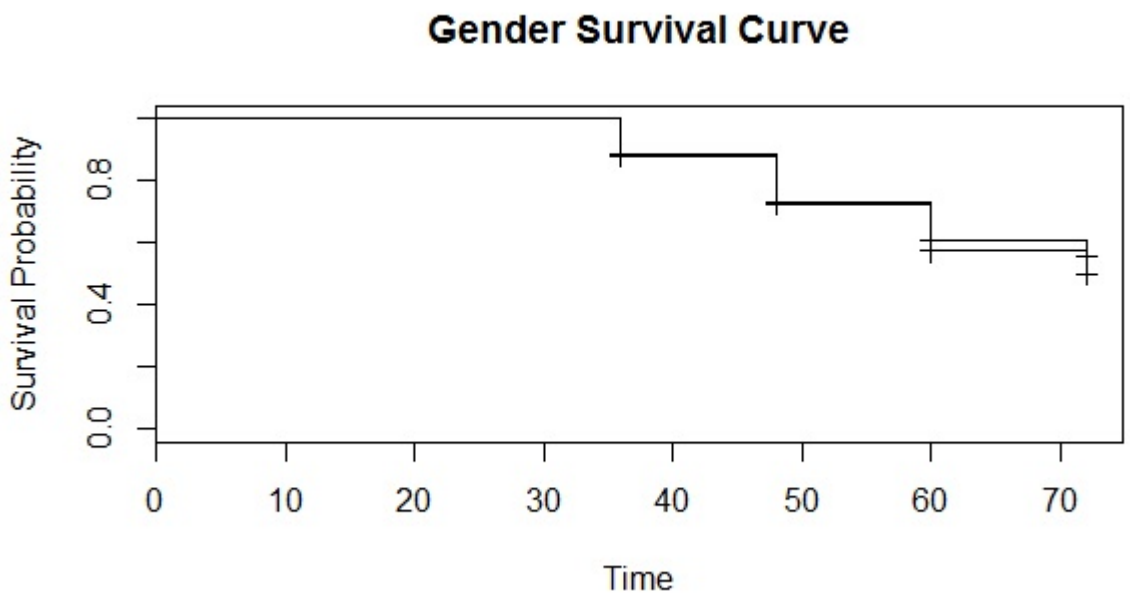

Figure 4.2: Gender Survival Curve

The lines in the curve do not cross each other, meaning the hazards for gender are proportional.

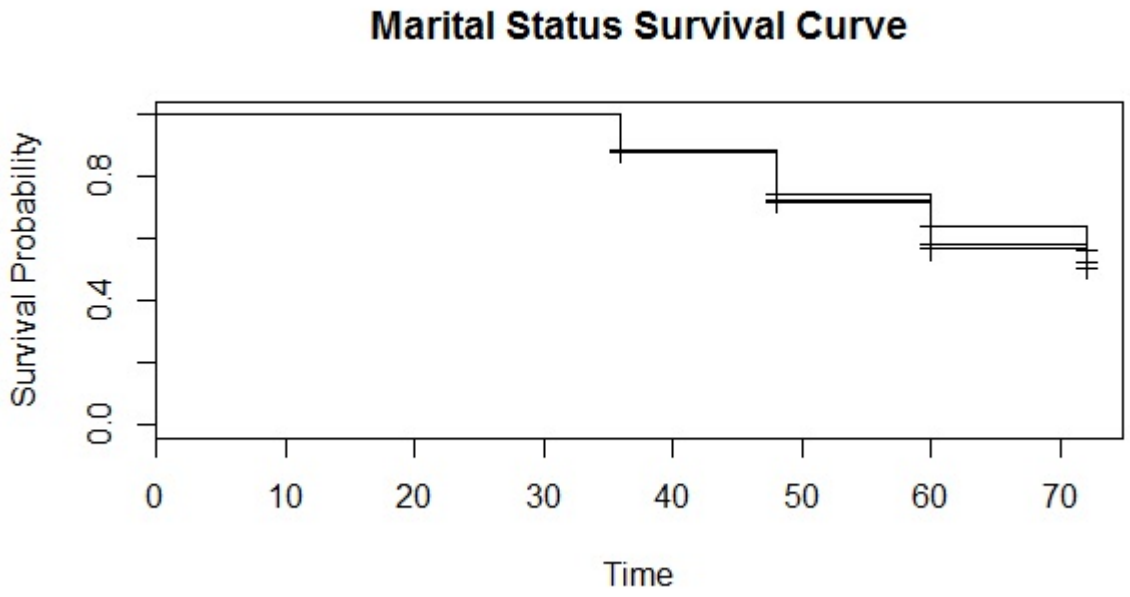

Figure 4.3: Marital Status Survival Curve

The lines in the curve do not cross each other, meaning the hazards for marital status are proportional.

This is sufficient proof that the observations can be analyzed by use of the Cox's Proportional Hazards Model.

4.4 Fitting the model

The model was fitted using all the variables collected from the loan forms which are compulsory to be filled by the loan borrowers, meaning the data did not have any cases of missing data.

With a total of 1000 entries, 376 were uncensored since these borrowers defaulted and form our events of interest. The rest either paid their loans

in time or well in advance before due date. Applying Cox's PH regression to the data, results were as follows:

The model had an R-square of 0.805 to mean that the variables in the model contributed 80.5% of the total variability. This was a good proportion and therefore the fitted model was considered good for this study.

4.5 Interpretation of the predictors

4.5.1 Age

Since $exp(\beta) < 0,100(1-0.9914) = 0.86\%$**; it means that an individual is 0.86% less likely to default with every unit increase in age.**

4.5.2 Type of employment

With $exp(\beta) < 0,100(1-0.174) = 82.6\%$**; it means that an individual is 82.6% less likely to default if employed permanently compared to when employed on contract.**

4.5.3 Loan Amount and Basic Salary

Both variables have their $exp(\beta) = 0$ **to mean their unit increase does not affect chances of default with other factors held constant.**

4.5.4 Marital status

With $exp(\beta) < 0,100(1-0.8907) = 10.92\%$**; it means that an individual is 10.92% less likely to default if married compared to if the applicant is divorced or widowed and 0.24% less likely to default if single compared to if the applicant is divorced or widowed.**

Table 4.2: Results

Variable	Coefficient (β_j)	exp (β_j)	SE	P Value	Likelihood of Default
Repaymentperiod	-1.749	0.174	82.25	0.983	82.6
countyofbirthMeru	-0.3556	0.7007	0.2147	0.0977	29.93
countyofbirthKisii	-0.3426	0.7099	0.2332	0.1418	29.01
countyofbithMombasa	-0.3419	0.7104	0.3555	0.3361	28.96
countyofbirthMachakos	-0.2277	0.7964	0.211	0.2805	20.36
countyofbirthKisumu	-0.201	0.8179	0.2642	0.4466	18.21
countyofbirthKajiado	-0.1697	0.8439	0.2203	0.441	15.61
countyofbirthKericho	-0.1644	0.8484	0.217	0.4488	15.16
countyofbirthKakamega	-0.08514	0.9184	0.2241	0.704	8.16
countyofbirthNyeri	-0.07267	0.9299	0.2209	0.7422	7.01
maritalstatusMarried	-0.1157	0.8908	0.1535	0.4512	10.92
maritalstatusSingle	-0.002372	0.9976	0.1333	0.9858	0.24
typeofemploymentPermanent	-0.08023	0.9229	0.1085	0.4595	7.71
employerMin of Health	-0.07258	0.93	0.1835	0.6924	7.0
employerNational Govt (others)	-0.01206	0.988	0.2133	0.9549	1.2
employerCounty Govt	0.04171	1.043	0.1906	0.8267	-4.3
employerMin of Education	0.04452	1.046	0.1897	0.8145	-4.6
genderMale	-0.04603	0.955	0.1075	0.6685	4.5
Age	-0.008653	0.9914	0.006373	0.1745	0.86
Amount	7.25E-08	1	1.24E-07	0.5595	0
Basicsalary	-1.80E-08	1	2.02E-07	0.9291	0
Timetakentoclearloan	0.08589	1.09	0.004047	<2e-16	-9

4.5.5 Employer

The analysis indicates that compared to those employed in the Armed Forces, those in the education sector are 4.6% more likely to default on their loans while those in the County governments are 4.3% more likely to default. Those who work in the national government are 1.2% less likely to default while Ministry of Health workers ratings stand at 7% less likely to default.

Ministry of Health workers are therefore the safest to offer loans. This is probably because the sector majorly contains professionals driven by passion for the job and not necessarily income.

4.5.6 County of birth

From the analysis, those born in Meru County are 29.93% less likely to default on their loans while those born in Nyeri County are 7.01% less likely to default. The comparisons have been done with those born in Garissa County. Those born in Kisii County are 29.01% less likely, Mombasa County are 28.96% less likely, Machakos County are 20.36% less likely, Kisumu County are 18.21% less likely, Kajiado County are 15.61% less likely, Kericho County are 15.16% less likely and Kakamega County are 8.16% less likely to default on their loans. This means that loans advanced to people born in Meru County are less risky compared to loans advanced to people born in other regions within Kenya while Nyeri leads in risk rating for loans when other factors are held constant.

4.5.7 Gender

The analysis indicates that the males are 4.5% less likely to default on their loans compared to the females.

4.5.8 Repayment Period

With all other factors held constant, we note that increasing repayment period by one unit reduces chances of defaulting by 82.6%.

Chapter 5

RECOMMENDATIONS

It is observed that a number of loans take some time before repayment begins to deteriorate and the borrowers ignore repaying altogether. It is important for the banks to keep on monitoring their loans from time to time and ensure they are in contact with their customers so that any character change or static status changes can be noted since these will reflect negatively on the loan repayment.

Since County of birth and employer greatly influence probability of default, it is advised that the banks should come up with sound policies on how applicants that have worst scenarios of joint probabilities of these attributes can be treated to ensure fairness as well as less risks.

The author recommends that the banks should assign weightings to customers descriptive data since they contribute differently to the credit risk and use the average weightings as a means to endorse whether an individual should be given loan or not.

ACRONYMS

PD - Probability to Default

PH Proportional Hazards

PG Prudential Guidelines

CBK - Central Bank of Kenya

APPENDIX

GENERATION OF CODES

```
remove(list=ls())
setwd("/Statistics/Sem 4/Project/codes")
library("survival")
library("leaps")

#Create data#
set.seed(2549)
amount<-round(runif(1000,500,2000),0)*1000
set.seed(8569)
age<-round(runif(1000,24,53),0)
set.seed(7163)
gend<-round(runif(1000, 1,2),0)
gend[gend==1]<-c("Male")
gend[gend==2]<-c("Female")
gender<-gend
set.seed(5364)
cob<-round(runif(1000, 1,10),0)
cob[cob==1]<-c("Kisumu")
cob[cob==2]<-c("Kisii")
cob[cob==3]<-c("Kakamega")
cob[cob==4]<-c("Meru")
cob[cob==5]<-c("Nyeri")
cob[cob==6]<-c("Machakos")
```

```
cob[cob==7]<-c("Kajiado")
cob[cob==8]<-c("Kericho")
cob[cob==9]<-c("Garissa")
cob[cob==10]<-c("Mombasa")
countyofbirth<-cob
set.seed(1637)
emp<-round(runif(1000, 1,5),0)
emp[emp==1]<-c("Armed Forces")
emp[emp==2]<-c("Min of Health")
emp[emp==3]<-c("Min of Education")
emp[emp==4]<-c("County Govt")
emp[emp==5]<-c("National Govt(Others)")
employer¡-emp
set.seed(8983)
ms<-round(runif(1000, 1,3),0)
ms[ms==1]<-c("Married")
ms[ms==2]<-c("Single")
ms[ms==3]<-c("Divorced/Widowed")
maritalstatus<-ms
set.seed(6769)
rp<-round(runif(1000, 1,4),0)
rp[rp==1]<-c(36)
rp[rp==2]<-c(48)
rp[rp==3]<-c(60)
rp[rp==4]<-c(72)
repaymentperiod<-rp
```

```
set.seed(3908)
basicsalary<-round(runif(1000,60,1000),0)*1000
set.seed(9836)
toe<-round(runif(1000, 1,2),0)
toe[toe==1]<-c("Permanent")
toe[toe==2]<-c("Contract")
typeofemployment<-toe
set.seed(7386)
timetakentoclearloan<-round(runif(1000,10,80),0)

#Generating time to event and status#
timetoevent<-repaymentperiod
status<-ifelse(repaymentperiod<timetakentoclearloan, status<-1,status<-0)

#Compile data in a table and view the table#
datacompile<-cbind(amount,age,gender,countyofbirth,employer,maritalstatus,rep

#View the table#
View(datacompile)

#Create a csv file and write the data on that file#
write.table(datacompile,file="datacompile.csv",sep=",")

#Reading data from the file#
readdata¡-read.csv(choose.files(),header=TRUE)
attach(readdata)
```

```
#Censoring those who paid their loans early and in time#
survcredit<-Surv(timetoevent,status==1)
survcredit

#Assumptions of the Cox's Proportional hazard model#
#Proportional hazards assumption#
plot(survfit(formula = Surv(timetoevent, status==1)  countyofbirth, data
= readdata, conf.type="none"),
lty=1, main="County of Birth Survival Curve", xlab="Time", ylab="Survival
Probability" )

plot(survfit(formula = Surv(timetoevent, status==1)  gender, data = read-
data, conf.type="none"),
lty=1, main="Gender Survival Curve", xlab="Time", ylab="Survival Prob-
ability" )

plot(survfit(formula = Surv(timetoevent, status==1)  maritalstatus, data
= readdata, conf.type="none"),
lty=1, main="Marital Status Survival Curve", xlab="Time", ylab="Survival
Probability" )

#Running the Cox Model#
coxmod4<-coxph(survcredit age+typeofemployment+amount+gender+basicsalar
summary(coxmod4)
```

REFERENCES

Allen, L. N. and Rose, L. C. (2006). Financial survival analysis of defaulted debtors, *Journal of Operational Research Society*, **57, 630-636.**

Baba, N. and Goko, H. (2006). Survival analysis of hedge funds, Bank of Japan,Working Papers Series No. 06-E-05.

Beran, J. and Djadja, A. K. (2007). Credit risk modeling based on survival analysis with inmunes, *Statistical Methodology*, **4, 251-276.**

Caire, D., & Kossmann, R. (2003). *Credit Scoring: Is It Right for Your Bank?* **Bannock Consulting.**

Cao, R. and Devia, A(2008). *Modelling consumer credit risk via survival analysis*

Glennon, D. and Nigro, P. (2005). Measuring the default risk of small business loans: a survival analysis approach, *Journal of Money, Credit, and Banking*, **37, 923-947.**

Hlongwane, T., Mdlongwa P., Nare H., and Moyo I.L (2014).Censored Regression Techniques for Credit scoring: *A Case Study for the Commercial Bank of Zimbabwe (Bulawayo)*

Malik, M. and Thomas L. (2006). Modelling credit risk of portfolio of consumer loans, University of Southampton, School of Management Working Paper Series No. CORMSIS-07-12.

Narain, B. (1992). Survival analysis and the credit granting decision. In: Thomas L., Crook, J. N. and Edelman, D. B. (eds.). *Credit Scoring and Credit Control*. **OUP: Oxford, 109-121.**

Roszbach, K. (2003). Bank lending policy, credit scoring and the survival of loans, Sverriges Riksbank Working Paper Series No. 154.

Stepanova, M. and Thomas, L. (2002). Survival analysis methods for personal loan data, *Operations Research*, **50, 277-289.**