UNIVERSITY OF CAPE COAST

A MATHEMATICAL MODEL FOR LENDING IN MICROFINANCE AND APPLICATIONS

BY

ABIGAIL BAIDOO

Thesis submitted to the Department of Mathematics of the School of Physical Sciences, College of Agriculture and Natural Sciences, University of Cape Coast, in partial fulfillment of the requirements for the award of Master of Philosophy degree in Mathematics

JULY 2019

DECLARATION

Candidate's Declaration

I hereby declare that this thesis is the result of my own original research and that no part of it has been presented for another degree in this university or elsewhere.

Candidate's Signature ..... Date .....

Name: Abigail Baidoo

Supervisors' Declaration

We hereby declare that the preparation and presentation of the thesis were supervised in accordance with the guidelines on supervision of thesis laid down by the University of Cape Coast.

Principal Supervisor's Signature ..... Date .....

Name: Prof. Olivier Menoukeu-Pamen

Co-Supervisor's Signature ..... Date: .....

Name: Prof. Ernest Yankson

ABSTRACT

Loan default is one of the major problems facing most financial institutions. The solution to this problem has been the use of a mathematical model to determine the probability of default of clients of these financial institutions. This study proposes a mathematical model for predicting the probability of default of clients from a microfinance institution. The logistic and survival analysis methods were used in building the model. The results from the logistic regression model showed that the variables Rate, Number of Repayment, Branch Name, Average Inflation Rate and Average Foreign Exchange Rate are significant in predicting the probability of default. The linearity test showed that Number of Repayment was nonlinear and was transformed using restricted cubic splines. The survival analysis model showed that the variables Rate, Product, Branch Name, Easter, New Year, Ramadan, Average Inflation Rate, Average Unemployment Rate, and Average Foreign Exchange Rate were significant in predicting the probability of default of clients. The variables Average Inflation Rate and Average Unemployment Rate were transformed using restricted cubic splines. There also existed interactions between Rate and Product, New Year and Ramadan, and Easter and Average Inflation Rate. The fitted models were evaluated and validated.

KEY WORDS

Counting Processes

Credit Scoring

Logistic Regression

Microfinance

Survival Analysis

ACKNOWLEDGEMENTS

I say a big thank you to my supervisors Prof. Olivier Menoukeu-Pamen and Prof. Ernest Yankson for their support and guidance.

I would like to express my appreciation to my family for their unending support and encouragement. I would like to thank the AIMS-Ghana Research Centre for their support and also the GNPC Foundation.

I express my gratitude to all friends, colleagues and senior colleagues who contributed in diverse ways to make this programme a success.

DEDICATION

I dedicate this work to my family and the (AIMS)-Ghana research center.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

Microfinance institutions provide credit facilities as part of their services to clients. The provision of loan is usually accompanied by a number of risks which, if not managed, increases the likelihood that a customer will not pay back a loan. These risks are usually known as credit risk. It therefore becomes necessary for these institutions to focus on determining the risk of potential loan customers who will not pay back their loans using credit risk models. According to Ojiako et al. (2013), credit risk models ensure that the number of customers who do not pay back their loans is reduced and the selection of customers who repay their loans increase. In this study, the main focus is on formulating a mathematical model for a microfinance institution to predict the probability of default of clients of this institution.

In this chapter, a brief introduction of this study is presented through the background of the study, problem statement and objectives the study.

Background to the Study

Financial services are mostly available to a small part of the population in most developing countries. As the economies of these countries keep growing, their financial sectors keep expanding, however, financial assets remain in the hands of a few people. Most of the people in developing countries do not have savings accounts and insurance policies, and do not receive credit from formal financial institutions. They seldom made or received payments through the financial institutions. The limited use of financial institutions in developing countries became an international policy concern (Fund, 2006). This necessitated the need to provide financial services to this part of the population who do not have access to the formal financial institutions, especially the poor. This was started through microfinance.

The modern concept of microfinance was started in the 1970's by Professor Muhammad Yunus, founder of the Grameen Bank. He started by lending

money from his own pocket to villagers who were not able to obtain credit at reasonable rates. After some years, the bank was set up to lend exclusively to groups of poor households (Morduch, 1999). This concept of microfinance has been adopted by most microfinance institutions in Ghana.

The microfinance sector in Ghana evolved into its current state through several financial sector policies and programs such as providing subsidized credits, establishing rural and community banks, liberalizing the financial sector, and the promulgation of the PNDC Law 328 of 1991 which allowed the establishment of non-bank financial institutions. The sector is made up of three broad categories: the formal, semi-formal and informal suppliers (Asiama & Osei, 2007). Currently, there are about 137 microfinance institutions that are operational in Ghana.

Microfinance has been defined in various ways based on its objectives and purposes. Otero (1999) defines microfinance as the provision of financial services to low-income, poor and very poor self-employed people. According to Asiama & Osei (2007), microfinance includes providing financial services and managing small amounts of money through a range of products and a system of intermediary functions targeted at low income clients. It is made up of loans, savings, insurance, transfer services and other financial products and services. Microfinance institutions are financial institutions that assist small enterprises, households and the poor who do not have access to the more institutionalized financial system, in mobilizing savings and obtaining access to financial services. These institutions differ in legal structure, mission and methodology (Blavy et al., 2004; Lafourcade et al., 2005).

Microfinance aims at development in various ways at the client, institutional and country levels. At the client level, microfinance aims at alleviating poverty by creating access to productive capital. It also creates private institutions that provide financial services at the institutional level. These institutions become distribution channels for providing services that respond to the material

capital needs of the poor which is part of the infrastructure of a country. At the country level, it becomes a regulated institution that is part of the financial system in a country (Otero, 1999). According to Badugu & Tripathi (2016), "microfinance is an economic development tool whose objective is to assist the poor to work their way out of poverty by providing permanent access to appropriate financial services such as insurance, savings, and fund transfer".

Microfinance institutions in Ghana however face some challenges that affect the delivery of services they provide. These challenges can be both external and internal. The external challenges of microfinance institutions include the near absence of basic infrastructure, lack of banking culture in the rural areas, peri-urban, and among the urban poor, the failure of other microfinance institutions and other rural and community banks, constant government policy changes, among others. Some of the internal challenges include high operational cost, limited support for human and institutional capacity building, fraud and theft, and loan default (Boateng, 2015).

According to the Basel Committee on Banking Supervision (2004), loan default occurs when one or both of the following happens: the bank considers that the debtor is unlikely to pay his or her credit obligations to the banking group in full or the debtor is past due more than 90 days on any credit obligation to the banking group. In this study, loan default is defined as the inability of a debtor to pay his or her credit obligations in full.

Statement of the Problem

Njeru Warue (2012) wrote that the most common and serious vulnerability in a microfinance institution is the chance that it may not receive its money back from borrowers. This is due to the fact that most microloans are not secured and default can spread from a handful of loans to a significant portion of the portfolio. Since the objective of every microfinance institution is to make profit in order to maintain its stability and improve growth and sustainability, the presence of high levels of loan default negatively affects the institution's level

of private investment and restrict the scope of credit to borrowers.

Loan default, therefore forms the basis of credit risk for microfinance institutions not only in Ghana but the world at large. This suggests that microfinance institutions need find a way of controlling credit risk. However, most microfinance institutions use subjective or qualitative judgment to assess the ability of clients to repay loans.

One of the major techniques employed in controlling credit risk is the development of credit scoring models. These models are used to assess the creditworthiness of clients of these microfinance institutions by predicting the probability of default of clients. There has been a lot of literature on developing credit scoring models for predicting the probability of default of clients of microfinance institutions. For instance Ofori et al. (2014); Kwofie et al. (2015), and Yeboah (2012) focused on predicting the probability of default of clients of some microfinance institutions in Ghana. These models mainly focused on the application variables such as the socio demographic information of clients, and the loan characteristics. Also, some macroeconomic variables are considered in some studies.

In this study, a new feature is the introduction of some variables linked to the Ghanaian environment in the computation of the probability of default of a customer. These variables include: Christmas, New Year, Easter, Ramadan and Academic Year.

Research Objectives

The objectives of this study have been classified into general and specific objectives as follows.

General objective

This study seeks to formulate a suitable mathematical model for predicting the probability of default of a microfinance institution using a given dataset.

Specific objectives

The main objective of the study is to develop a mathematical model for predicting the probability of default of clients of a microfinance institution. Specifically, this study seeks to:

1. identify the factors that influence the probability of default of clients of a microfinance institution based on a given dataset.

2. formulate a suitable mathematical model for predicting the probability of default based on the factors identified.

Significance of the Study

This study will be significant in the following ways. It will enable the microfinance institution to:

1. make faster and informed credit decisions.

2. reduce its losses and grant credit to more customers who will pay back in full.

3. reduce the cost involved in assessing credit risk.

4. objectively assess credit risk of clients.

Delimitation

This study is based on a dataset from a microfinance institution in Accra with five branches. The products offered include group loans and individual loans. The focus of this study was centered individual loans. That is, predicting the probability of default of customers who take individual loans.

Limitation

The major limitation of this study was availability of data. The data obtained for the study contained no information about the number of dependents, income level, occupation of clients and the number of times a clients has been in

arrears. Also, the data contained missing information about some clients which reduced the sample size used for the study.

Definition of Terms

In this section, we define some terms used in the thesis. These terms include concepts in probability theory, statistics, among others. Most of these definitions are borrowed from Castañeda et al. (2012); Fleming & Harrington (2011); Bojanov et al. (2013), and Heumann et al. (2016).

Probability Theory

In this part, we define some concepts in probability theory.

Definition 1

A *sample space* is the set of all possible results of a random experiment. We denote it by Ω.

Remark 1 (i) An element, $\omega \in \Omega$, is called an *outcome*.

(ii) A subset of the sample space is called an *event*.

(iii) Two events are said to be *independent* if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

(iv) Two events A and B are said to be *mutually exclusive* if $A \cap B = \emptyset$.

Definition 2

Let Ω be a non-empty set. A collection of subsets over Ω denoted by $\mathscr{F}$ is called a σ-*algebra* over Ω if the following conditions are satisfied:

(i) $\Omega \in \mathscr{F}$

(ii) $A \in \mathscr{F}$, and $A^c \in \mathscr{F}$

(iii) $A_1, A_2, \cdots \in \mathscr{F}$ and $\bigcup_{i=1}^{\infty} A_i \in \mathscr{F}$.

The elements of $\mathscr{F}$ are called events.

The following theorem holds for a collection of events.

Theorem 1 (Castañeda et al. (2012))

If $\Omega \neq \emptyset$ and $\mathscr{F}_1, \mathscr{F}_2, \cdots$ are σ-algebras, then $\bigcap_{i=1}^{\infty} \mathscr{F}_i$ is also a σ-algebra over Ω.

Definition 3

A family of sub-σ-algebras $\{\mathscr{F}_t : t > 0\}$ of a σ-algebra $\mathscr{F}$ is called increasing if $s \leq t$ implies $\mathscr{F}_s \subset \mathscr{F}_t$. An increasing family of sub-σ-algebras is called a *filtration.*

Remark 2

Let $\{\mathscr{F}_t : t > 0\}$ be a filtration, then

(a) the σ-algebra $\bigcap_{h>0} \mathscr{F}_{t+h}$ is denoted by $\mathscr{F}_{t+}$. The limit from the left, $\mathscr{F}_{t-}$, is the smallest σ-algebra containing all the sets in $\bigcup_{h>0} \mathscr{F}_{t-h}$.

(b) a right continuous filtration for any t is defined as $\mathscr{F}_{t+} = \mathscr{F}_t$.

In a random experiment, we are often interested in, at each time, the information generated by some events. Hence, we have a *generated σ-algebra.*

Definition 4

Let $\Omega \neq \emptyset$ and $\mathscr{A}$ be a collection of subsets of Ω. Let $\mathscr{M} = \{\mathscr{F} : \mathscr{F}$ is a σ-algebra over Ω containing $\mathscr{A}\}$. Then the σ-algebra generated by $\mathscr{A}, \sigma(\mathscr{A})$ is the smallest σ-algebra over Ω containing $\mathscr{A}$. That is $\sigma(\mathscr{A}) := \bigcap_{\mathscr{F} \in \mathscr{M}} \mathscr{F}$.

Definition 5

The smallest σ-algebra over $\mathbb{R}$ containing all intervals of the form $(-\infty, a]$ with $a \in \mathbb{R}$ is called the *Borel σ-algebra*, usually written as $\mathscr{B}(\mathbb{R})$.

Remark 3

If $A \subset \mathscr{B}(\mathbb{R})$, then A is called a Borel subset of $\mathbb{R}$.

The concept of random variables are usually defined with respect to a probability space, which is based on measurable space. Next, we define measurable space.

Definition 6

Let $\Omega \neq \emptyset$ and $\mathscr{F}$ be a σ-algebra over Ω. The couple $(\Omega, \mathscr{F})$ is called a *measurable space*.

Definition 7

Let $(\Omega, \mathscr{F})$ be a measurable space. A real valued function $\mathbb{P}$ defined over $\mathscr{F}$ which satisfies the following conditions:

(i) $\mathbb{P}(A) \geq 0$ for all $A \in \mathscr{F}$

(ii) $\mathbb{P}(\Omega) = 1$

(iii) if $A_1, A_2, \cdots$ are mutually exclusive events in $\mathscr{F}$, then $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$

is called a probability measure over $(\Omega, \mathscr{F})$.

The triple $(\Omega, \mathscr{F}, \mathbb{P})$ is called a *probability space*.

Definition 8

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. A random variable is a mapping $X : \Omega \to \mathbb{R}$ such that, for all $A \in \mathscr{B}(\mathbb{R}), X^{-1}(A) \in \mathscr{F}$. It is also defined in set notation form as: $\{X \in A\} := \{\omega \in \Omega : X(\omega) \in A\}$, with $A \in \mathscr{B}(\mathbb{R})$.

Remark 4

A random variable could be interpreted as the outcome of an experiment whose result cannot be determined beforehand or a random experiment.

The following theorem holds for distribution functions.

Theorem 2 (Castañeda et al. (2012))

Suppose that X is a random variable defined over the probability space $(\Omega, \mathscr{F}, \mathbb{P})$. The function $\mathbb{P}_X$ defined over the σ-algebra $\mathscr{B}(\mathbb{R})$ through $\mathbb{P}_X(B) := P(\{X \in B\})$ for all $B \in \mathscr{B}(\mathbb{R})$ is a probability measure over $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ called the distribution function of the random variable X.

Since every random variable has a distribution function, the concept of distribution function is defined as follows.

Definition 9

Let X be a random variable. The function F_X defined over $\mathbb{R}$ by:

$$\begin{aligned} F_X(x) :=& \mathbb{P}_X((-\infty, x]) \\ =& \mathbb{P}(X \leq x), \end{aligned}$$

is the *distribution (or cumulative distribution) function* of X.

The following theorem gives some properties of a distribution function.

Theorem 3 (Castañeda et al. (2012))

Let X be a random variable defined over $(\Omega, \mathscr{F}, \mathbb{P})$. The distribution function F_X satisfies the following:

(i) if $x < y$, then $F_X(x) \leq F_X(y)$

(ii) $F_X(x^+) := \lim_{h \downarrow 0^+} F_X(x+h) = F_X(x)$ for all $x \in \mathbb{R}$

(iii) $\lim\limits_{x \uparrow +\infty} = 1$ and $\lim\limits_{x \downarrow -\infty} = 0$

Remark 5

Random variables are classified based on their distribution functions. If the distribution of the random variable X is a step function, then X is said to be a discrete random variable. And if the distribution of the random variable X is absolutely continuous, then X is said to be a continuous random variable.

Theorem 4 (Castañeda et al. (2012))

Let X be a discrete real random variable defined over $(\Omega, \mathscr{F}, \mathbb{P})$ with distribution function F_X. Then the number of jumps of F_X is at most countable.

Definition 10

Let X be a random variable. X is said to be a *discrete random variable* if it takes values from some countable subset $\{x_1, x_2, \cdots\}$ of $\mathbb{R}$ only.

Remark 6

Let F_X be the distribution function of the discrete real random variable X. We

say that F_X presents a jump at $x \in \mathbb{R}$ if $F_X(x) - F_X(x-) \neq 0$. Then the magnitude of the jump is defined as: $P(X = x) = F_X(x) - F_X(x-)$.

Definition 11

Let X be a discrete random variable with values $x_1, x_2, \cdots$ (all different). The *probability mass function* of the random variable X is defined on $\mathbb{R}$ by the function p_X as:

$$p_X(x) = \begin{cases} \mathbb{P}(X = x_i), & \text{if } x = x_1, x_2, \cdots \\ 0, & \text{otherwise,} \end{cases}$$

with the following properties:

(i) $p(x_i) \geq 0 \; \forall \; i$ and

(ii) $\sum_i p(x_i) = 1$.

Definition 12

Let X be a real random variable defined over $(\Omega, \mathscr{F}, \mathbb{P})$. X is said to be *continuous* if and only if there exists a nonnegative and integrable real function f_X such that for all $x \in \mathbb{R}$ it satisfies:

$$F_X(x) = \int_{-\infty}^{x} f_X(t) \mathrm{d}t,$$

where f_X is called a probability density function.

Definition 13

A function $f(x)$ is called a *probability density function* if it satisfies the following properties:

(i) $f(x) \geq 0$ for all values of x

(ii) $\int_{-\infty}^{\infty} f(x) \mathrm{d}x = 1$.

Definition 14

Let X and Y be two real-valued random variables defined on the same probability

space. X and Y are said to be independent if for any pair of Borel sets, the σ-algebra generated by X denoted by $\mathscr{A} = \sigma(X)$ and the σ-algebra generated by Y denoted by $\mathscr{B} = \sigma(Y)$ are independent. That is,

$$\mathbb{P}(X \in \mathscr{A}, Y \in \mathscr{B}) = \mathbb{P}(X \in \mathscr{A})\, \mathbb{P}(Y \in \mathscr{B}).$$

Next we define some concepts of continuity.

Definition 15

A function f is said to be *right continuous* at x if and only if $\lim_{t \downarrow x} f(t)$ exists and

$$f(t+) = \lim_{t \downarrow x} f(t) = f(x)$$

Definition 16

A function f is said to be *left continuous* at x if and only if $\lim_{t \uparrow x} f(t)$ exists and

$$f(t-) = \lim_{t \uparrow x} f(t) = f(x)$$

Definition 17

A function is said to be *càdlàg* if the left limit exist and is right continuous.

Definition 18

A function is said to be *càglàd* if the right limit exist and is left continuous.

Definition 19

A *real valued stochastic process* is a collection of random variables $\{X(t); t \in T\}$ defined on a common probability space $(\Omega, \mathscr{F}, \mathbb{P})$ with values in $\mathbb{R}$. T is the index set of the process which is a subset of $\mathbb{R}$. The state space of the stochastic process, denoted by S is the set of values that the random variable $X(t)$ can take.

The sample path of the process is the mapping defined for each fixed $\omega \in \Omega$ as:

$$X(\omega) : \mathscr{T} \to S$$
$$t \to X(t, \omega).$$

Definition 20

A *stochastic process* X is :

(i) integrable if $\sup_{0 \le t < \infty} \mathbb{E} \mid X(t) \mid < \infty$

(ii) square integrable if $\sup_{0 \le t < \infty} \mathbb{E} \mid X(t) \mid^2 < \infty$

(iii) bounded if there exists a finite constant K such that

$$\mathbb{P}\{ \sup_{0 \le t < \infty} \mathbb{E} \mid X(t) \mid < K\} = 1.$$

Definition 21

A probability space $(\Omega, \mathscr{F}, \mathbb{P})$ which is equipped with a right continuous filtration is called a *stochastic basis*. It is denoted as $(\Omega, \mathscr{F}, \mathbb{F} = \{\mathscr{F}_t\}_{t \ge 0}, \mathbb{P})$.

Definition 22

A stochastic basis is *complete* if the $\mathscr{F}$ contains any subset of a $\mathbb{P}$-null set and if each $\mathscr{F}_t$ contains all $\mathbb{P}$-null sets of $\mathscr{F}$.

Remark 7

$\mathbb{P}$-null sets are sets with probability zero almost surely.

Given the information at time t, to determine whether the value of the process is known, we need the concept of adaptedness.

Definition 23

A stochastic process $\{X(t); t \in T\}$ is *adapted* to a filtration, $\mathbb{F} = \{\mathscr{F}_t\}_{t \ge 0}$ if, for every $t \ge 0$, $X(t)$ is $\mathscr{F}_t$-measurable, that is $\{\omega \in \Omega : X(t, \omega) \le a\} \in \mathscr{F}_t$ for all $a \in \mathbb{R}$.

In some cases, given the information just before time t, enables one to find the value of the process at time t. This notion is known as predictability.

Definition 24

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space with a filtration $\mathbb{F} = \{\mathscr{F}_t\}_{t\geq 0}$. A *predictable* σ*-algebra* for the filtration is the σ-algebra on $[0,\infty) \times \Omega$ by all sets of the form

$$\{0\} \times D,\ D \in \mathscr{F}_0 \text{ and } (a,b] \times D,\ 0 \leq a < b < \infty,\ D \in \mathscr{F}_d.$$

Definition 25

A process X is said to be *predictable* with respect to a filtration, $\mathbb{F} = \{\mathscr{F}_t\}_{t\geq 0}$, if as a mapping from $[0,\infty) \times \Omega$ to $\mathbb{R}$, it is measurable with respect to the predictable σ-algebra generated by that filtration. X is called an $\mathscr{F}_t$-predictable process.

Proposition 1

Let X be an $\mathscr{F}_t$-predictable process. Then, for any $t > 0$, $X(t)$ is $\mathscr{F}_{t-}$-measurable.

Most often, we observe the number of events that occur over time in survival analysis. The number of events that occur over time can be counted and modelled using counting processes.

Definition 26

A *counting process* is a stochastic process $\{N(t); t \geq 0\}$ adapted to a filtration $\mathbb{F} = \{\mathscr{F}_t\}_{t\geq 0}$ with $N(0) = 0$ and $N(t) < \infty$ almost surely, and whose paths are probability one right continuous, piecewise constant, and have jump discontinuities, with jumps of size $+1$.

Remark 8

In counting processes, $N(t) - N(s)$ is used to denote the number of events of a certain type that occurs in the interval $(s,t]$.

In probability theory, we are interested in a fair game. The concept of a martingale is used to describe a fair game.

Definition 27

Let $(\Omega, \mathscr{F}, \mathbb{F} = \{\mathscr{F}_t\}_{t\geq 0}, \mathbb{P})$ be a filtered probability space and $X = \{X(t) : t \geq 0\}$ be a càdlàg process. X is said to be a *martingale* with respect to $\{\mathbb{P}, \mathbb{F}\}$ if:

(i) X is adapted to $\mathbb{F} = \{\mathscr{F}_t\}_{t\geq 0}$.

(ii) $\mathbb{E} \mid X(t) \mid < \infty$ for all $t < \infty$

(iii) $\mathbb{E}[X(t+s) \mid \mathscr{F}_t] = X(t)$ almost surely for all $s \geq 0, t \geq 0$.

Remark 9

A martingale can also be submartingale or supermartingale:

1. X is said to be a submartingale if (iii) in Definition 27 is replaced by $\mathbb{E}[X(t+s) \mid \mathscr{F}_t] \geq X(t)$ almost surely.

2. X is said to be a supermartingale if (iii) in Definition 27 is replaced by $\mathbb{E}\{X(t+s) \mid \mathscr{F}_t\} \leq X(t)$ almost surely.

The properties of a martingale are as follows:

1. For any $k > 0, \mathbb{E}[X(t) \mid \mathscr{F}_{t-k}] = X(t-k)$, therefore it is expected that $\mathbb{E}[X(t) \mid \mathscr{F}_{t-}] = X(t-)$.

2. $\mathbb{E}[X(t) - X(t-h) \mid \mathscr{F}_{t-h}] = X(t-h) - X(t-h) = 0$ almost surely implies that $\mathbb{E}[\Delta X(t) \mid \mathscr{F}_{t-}] = 0$ for a right continuous martingale.

3. $\mathbb{E}[X(t) \mid \mathscr{F}_{t-}] = X(t-)$.

   An example of a martingale is Brownian motion.

Theorem 5 (Fleming & Harrington (2011))

Let X be a right-continuous nonnegative submartingale with respect to the stochastic basis $(\Omega, \mathscr{F}, \{\mathscr{F}_t : t > 0\}, \mathbb{P})$. Then, there exists a right-continuous martingale M and an increasing right-continuous predictable process A such that $\mathbb{E}\{A(t)\} < \infty$ and $X(t) = M(t) + A(t)$ almost surely for any $t \geq 0$. If $A(0) = 0$ almost surely, and if $X = M' + A'$ is another such decomposition with $A'(0) = 0$, then for any $t \geq 0$, $\mathbb{P}\{M'(t) \neq M(t)\} = 0 = \mathbb{P}\{A'(t) \neq A(t)\}$. If in addition X is bounded, then M is uniformly integrable and A is integrable.

Corollary 1

Let $\{N(t) : t \geq 0\}$ be a counting process adapted to a right-continuous filtration $\{\mathscr{F}(t) : t \geq 0\}$ with $\mathbb{E}[N(t) < \infty]$ for any t. Then, there exists a unique increasing right-continuous $\mathscr{F}_t$-predictable process A such that $A(0) = 0$ almost surely, $\mathbb{E}\{A(t)\} < \infty$ for any t, and $\{M(t) = N(t) - A(t) : t \geq 0\}$ is a right-continuous $\mathscr{F}_t$-martingale.

Spline Functions

In this part, we define some concepts related to spline functions. Splines are usually used in approximating functions.

Definition 28

The function $s(x)$ is called a *spline function* of degree k with knots $\{x_j\}_1^n$ if $-\infty =: x_0 < x_1 < \cdots < x_n < x_{n+1} := \infty$ and

(i) for each $j = 0, \cdots, n, s(x)$ coincides on (x_j, x_{j+1}) with a polynomial of degree not greater than k ;

(ii) $s(x), s'(x), \cdots, s^{k-1}(x)$ are continuous functions on $(-\infty, \infty)$.

Remark 10

From Definition 28, it can be concluded that:

(i) every polynomial is a spline function which does not have knots.

(ii) the kth derivative of a spline of degree k with knots $\{x_j\}_1^n$ is a piecewise constant function with breaks at $x_1, \cdots, x_n$.

Theorem 6 (Bojanov et al. (2013))

Let $S_k(x_1, \cdots, x_n)$ denote the class of all spline functions. The function f belongs to $S_k(x_1, \cdots, x_n)$ of degree k with knots at $x_1, \cdots, x_n$, if and only if it may be written in the form:

$$f(x) = \sum_{d=0}^{k} c_d x^d + \sum_{j=1}^{n} b_j (x - x_j)_+^k,$$

with some real coefficients $\{c_d\}$ and $\{b_j\}$. Moreover,

$$b_j = \frac{f^k(x_j+0) - f^k(x_j-0)}{k!},$$

for $j = 1, \cdots, n$.

Remark 11

From Theorem 6,

(i) $f(x+0) = \lim_{t \downarrow x} f(t)$ and $f(x-0) = \lim_{t \uparrow x} f(t)$

(i) $(x-x_j)_+^k$ is a truncated power function defined as:

$$(x-x_j)_+^k = \begin{cases} (x-x_j)^k & \text{for } x \geq x_j, \\ 0 & \text{for } x < x_j. \end{cases}$$

Corollary 2

Let $a < x_1 < \cdots < x_n < b$ be arbitrary fixed points. Then the functions

$$1, x, \cdots, x^k, (x-x_1)_+^k, \cdots, (x-x_n)_+^k,$$

form a basis of the space $S_k(x_1, \cdots, x_n)$ in $[a,b]$ and hence

$$\dim S_k(x_1, \cdots, x_n) = n+k+1.$$

Definition 29

The function $s(x)$ is said to be a spline function of degree k with knots $x_1, \cdots, x_n$ of multiplicities $z_1, \cdots, z_n$, respectively, if $-\infty =: x_0 < x_1 < \cdots < x_n < x_{n+1} := \infty$ and

(i) for each $j = 0, \cdots, n, s(x)$ coincides on (x_j, x_{j+1}) with a polynomial of degree k.

(ii) $s \in C^{k-z_j}(x_j, x_{j+1})$ for $j = 1, \cdots, n$.

Theorem 7 (Bojanov et al. (2013))

Let $S_k((x_1,z_1),\cdots,(x_n,z_n))$ denote the set of all splines with multiple knots. The function $f(x)$ belongs to the class $S_k((x_1,z_1),\cdots,(x_n,z_n))$ if and only if it can be written in the form:

$$f(x)=\sum_{d=0}^{k}c_d x^d+\sum_{j=1}^{n}\sum_{l=0}^{z_j-1}b_{jl}(x-x_j)_+^{k-l},$$

where c_d and b_{jl} are real constants. Moreover,

$$b_{jl}=\frac{f^{(k-l)}(x+0)-f^{(k-l)}(x-0)}{(k-l)!}.$$

Corollary 3

The functions

$$\{x_d\}_{d=0}^{k},\quad \{(x-x_k)_+^{k-l}\}_{k=1,l=0}^{n\ \ z_j-1},$$

are linearly independent on $(-\infty,\infty)$ and they constitute a basis in $S_k((x_1,z_1),\cdots,(x_n,z_n))$.

Definition 30

Let γ_{k-1} denote the class of all polynomials less than or equal to $k-1$. The spline function $s(x)$ of odd degree $2k-1$ with knots $x_1,\cdots,x_n$ is said to be a *natural spline function*, if the restriction of s over $(-\infty,x_1)$ and (x_n,∞) is a polynomial from γ_{k-1}.

Remark 12

The class of natural spline functions of degree $2k-1$ with knots $\{x_j\}_1^n$ is denoted by $N_{2k-1}(x_1,\cdots,x_n)$, which consists of all splines s from $S_{2k-1}(x_1,\cdots,x_n)$, satisfying the boundary conditions:

$$s^{(m)}(x_1)=s^{(m)}(x_n)=0,\quad m=k,\cdots,2k-1.$$

Lemma 1

The function $s(x)$ is a natural spline of degree $2k-1$ with knots $x_1, \cdots, x_n$ if and only if it can be written in the form

$$s(x) = p(x) + \sum_{j=1}^{n} q_j (x - x_j)_+^{2k-1},$$

where $p \in \gamma_{k-1}$ and the coefficients $\{q_j\}$ satisfies the conditions

$$\sum_{j=1}^{n} q_j x_j^m = 0 \quad \text{for } m = 0, \cdots, k-1.$$

In every regression analysis, some variables are used. In the next subsection, the concept of variables are discussed.

Definition 31

A *dependent variable* is the variable whose values are being determined or depend on other variables in a regression model. It is also called response variable or outcome.

Definition 32

An *independent variable* is the variable whose values do not depend on other variables or are used to determine the value of the response variable in a regression model. It is also called covariate or regressor.

Definition 33

Categorical variables are variables that take a finite number of values. That is any discrete, nominal, ordinal or qualitative variable can be regarded as a categorical variable.

Definition 34

Continuous variables are variables that take an infinite number of values. They are usually quantitative variables.

Organisation of the Study

This thesis is organised aside this chapter as follows. Chapter TWO gives

a review of related literature. The review is based on literature on credit scoring models and models for predicting the probability of default. Chapter THREE discusses the methods used in building a probability of default model for clients of a microfinance institution. The methods include logistic regression and survival analysis. Chapter FOUR presents the main results that were obtained in the study. First of all, a brief description of the dataset used for the study is given. Then the model for probability of default with respect to the logistic regression and survival analysis models based on the dataset are provided. Chapter FIVE discusses a summary of the results obtained in the study, conclusions to be drawn, and some recommendations.

Chapter Summary

In this chapter, an introduction of the study is given by a brief background and the problem to be examined in this study. The objectives, significance and delimitations of the study are also stated. Some terms that will be used in this study are also defined. Then the structure of the study, that is how the study is organised is outlined.

CHAPTER TWO

LITERATURE REVIEW

Introduction

In this chapter, a review of related literature is presented. It includes a discussion of literature on credit scoring models and models for predicting probability of default such as logistic regression and survival analysis models.

One of the functions of any financial institution is to grant loans (or credit) to individuals or firms. In financial terms, credit refers to an amount of money given to an individual or a firm by a financial institution which has to be repaid with interest in installments mostly at regular intervals (Hand & Henley, 1997). According to Kocenda & Vojtek (2009), one of the most profitable investments in lenders' asset portfolios is lending (at least in developed countries). However, an increase in the amount of loans results in an increase in the number of defaulted loans. Therefore, the lender faces the problem of differentiating between low and high risk clients before granting credit.

Fatemi & Fooladi (2006) argue that banks and other firms in the financial services industry must engage in credit risk management. Credit risk management has become very important because the types of counterparties are increasing (from individuals to governments) and the ever-expanding diversity in the forms of obligations. They defined credit risk as "uncertainty in a given counterparty's ability to meet its obligations". Credit risk analysis can be qualitative or quantitative. Qualitative analysis involves interviewing the potential borrower, going through the business plan and reviewing past financial history (if any). Quantitative analysis involves quantifying the predicted risk of a potential borrower through the use of a credit score. This requires keeping track of loan data (Coravos, 2010).

Lenders must engage in risk analysis before granting loans since the granting of credit is associated with risk. Rychnovskỳ (2018) explained that the loan approval process involves evaluating the client's ability to repay the loan and

verifying the income and other information provided. The ability to repay can be assessed by checking the sufficiency and stability of the income to cover all the expenses and also evaluating the riskiness of the client by estimating the probability of default. Yeh & Lien (2009) also emphasized that the main purpose of risk prediction is using financial information, such as business financial statement or customer transaction and repayment methods to predict business performance or individual customers' credit risk. They stated that "from the perspective of risk control, estimating the probability of default will be more meaningful than classifying customers into risky and non-risky".

In the section that follows, works on credit scoring models are presented.

Credit Scoring Models

Currently, credit scoring models are used to examine the credit worthiness of customers of financial institutions before granting credit. There are a lot of literature on credit scoring models and the techniques used in building such models.

Thomas et al. (2002) defined credit scoring as "the set of decision models and their underlying techniques that aid lenders in the granting of consumer credit. These techniques decide who will get credit, how much credit they should get, and what operational strategies will enhance the profitability of the borrowers to the lenders".

Coravos (2010) explained that the credit score can have two different meanings: a personal credit score, which measures an individual's consumer credit history, and an in-house credit scoring model, which is a combination of the personal credit score and other variables such as the business' cashflow.

According to Thomas (2000), credit scoring methods can be grouped into judgemental, statistical and non parametric statistical and artificial intelligence (AI) modelling approaches. The judgemental approach was based on answers to the 3C's or 4C's or 5C's. These C's are character of the person, capital, collateral, capacity and condition (in the market). The statistical based methods

for credit scoring models include the linear discriminant analysis model, logistic regression model and classification trees. The non parametric or AI modelling approaches include neural networks, expert systems, genetic algorithms and nearest neighbour methods.

Statistical models used in credit scoring quantifies risk and has a lot of advantages over judgemental scoring. They include consistency, explicitness, quantification of risk as a probability, ability to account for a wide range of risk factors, ability to be tested before use, among others. However, they also found that the credit scoring models have some disadvantages. They include susceptibility to abuse, assuming the future is the same as the past, requiring high quality data, assuming that a large proportion of risk is linked with characteristics found in the database, just to mention a few (Schreiner, 2004a).

According to Malik & Thomas (2010), credit scoring can be classified into application and behavioural scorings. Application scoring involves predicting customers' default risk at the time an application is made for the loan. Behavioural scoring is similar to application scoring but involves the observation of the recent payment and purchase behaviour of customers who have been granted loans. Credit scoring models can be used for two purposes: to predict the probability of default and provide a classification table to evaluate the predictive power of the model (Viswanathan & Shanthi, 2017). Thomas (2000) emphasized that credit scoring turns to be a classification problem when the input are answers to questions on the application form and the results of checks with a credit reference bureau and the output is grouping customers into good and bad.

Some of the major techniques used in building probability of default models are logistic regression and survival analysis models. In the next section, a review of literature on logistic regression models is presented.

Logistic Regression

Logistic regression models are one of the traditional models for predicting

the probability of defaults. Ofori et al. (2014) studied the determinants of credit default in microfinance institutions in Ghana using logistic regression model to predict the probability of default. The data sample used for the study was 2631 applicants. They found that the variables age, gender, marital status, income level, residential status, number of dependents, loan amount and tenure were significant in determining loan default. They recommended that in order to reduce the level of default, this model should be used by microfinance institutions to screen new loan applicants.

Yeboah (2012) also used the logistic regression analysis to study the factors that contribute to predicting microfinance credit default. A sample of 409 clients of a rural bank in Northern Ghana were used. The results showed that educational level, number of dependents, type of loan, adequacy of loan facility, duration of repayment of loan, number of years in business, cost of capital and period within year loan was advanced were significant in predicting the probability of clients. They recommended that microfinance institutions should adopt group loan policies.

Abid et al. (2016) used logistic regression and discriminant analysis to develop predictive models which distinguishes between good and bad customers. The data was collected from a Tunisian commercial bank from 2010 to 2012. The data consisted of four variables: applicant's age, loan amount, outstanding credit and occupational category. The results indicated that the logistic regression model has the lowest classification error rate.

Westgaard & Van der Wijst (2001) estimated the expected default frequency clients in a corporate bank portfolio using the logistic regression model. They used Norwegian default and accounting data from Dun & Bradstreet from 1995 to 1999. They examined the impact of financial variables and other firm characteristics such as age, size and industry classification. They concluded that "expected default frequency decreases as a function of the ratios of cash flow to debt, liquidity, solidity and financial coverage, as well as with size and age."

Viswanathan & Shanthi (2017) in evaluating the accuracy of forecast using the credit scoring methodology used the logistic regression and neural network models to build credit scoring models for a micro finance firm in India.They used primary data collected from ABC Ltd with a sample size of 640 customers. They found that the neural network model perform better than the logistic regression. However, they resorted to the logistic regression model to predict the probability of default because of the inability to express the neural network model in a precise form of equation.

Coravos (2010) identified the characteristics of a risky loan, using small business loan portfolio data from a national community development financial institutions (CDFI), for a CDFI borrower population. The characteristics include borrower-specific, loan-specific, lender specific characteristics and macroeconomic variables. They added interaction terms in the regression models. This was done because she claimed that the types of loans change with the economic health of the population. The models used for the analysis were ordinary least squares, logistic and multinomial logistic regression models.

However, it is argued that the logistic regression, discriminant analysis and other traditional models used in predicting the probability that an applicant for credit will default at a given time in the future are static models. That is, they concentrate only on the status of an applicant after a fixed period of time in their credit history (Banasik et al., 1999).

In the section that follows, literature that focuses on using survival analysis models to predict probability of default are reviewed.

Survival Analysis

Survival analysis models have been used as alternative models for predicting probability of default. Noh et al. (2005) examined whether the survival analysis model is a useful alternative and/or complement when compared to the traditional binary classifying approaches used in credit scoring. They used a Korean personal credit card dataset, consisting of $11,853$ customers, which

contained customers' card usage histories and repayment status for 13 months, and demographic information. They found that the overall classification power of the models is similar. However, in making comparisons at the same cut-off point, the logistic regression and neural network models showed better results in predicting 'good' customers while the survival analysis model showed better results in predicting 'bad' customers.

Hassan et al. (2018) examined the factors that affect credit risk using survival analysis models such as the accelerated failure time model and the Cox proportional hazard model. The factors included information from customers' credit application, their credit report and macroeconomic factors. The data used for the study was a proprietary data of customers of a Southern Louisiania credit union. They found that factors specific to customers and macroeconomic factors play an important role in the duration of a loan.

Banasik et al. (1999) applied three types of proportional hazards and accelerated failure time models to loan data and compared the results with logistic regression model. The data was made up of application information of $50,000$ loans that were accepted between June 1994 and March 1997 together with their monthly performance description for the period up to July 1997. They found that the proportional hazards models are competitive with the logistic regression model in identifying clients who default and superior to the logistic regression model in identifying those who will pay-off early in the first year.

Stepanova & Thomas (2002) also applied survival analysis models to personal loan data using the Cox proportional hazards model were used. The dataset was obtained from a major U.K. financial institution which consisted of application information of $50,000$ personal loans, together with the repayment status for each month of the observation of 36 months. They identified three developments that improved the application the Cox proportional hazards model to building credit scoring models: coarse classification of characteristic variables used in credit scoring, use of diagnostics to test the adequacy of the credit risk

model and the use of time dependent models to overcome the restriction of the proportional hazards.

Rychnovskỳ (2018) examined a new performance criteria which focuses on the predictive power of models and compared logistic regression model with the survival-based Cox proportional hazards model. The data used was from a Czech bank and consisted of 19, 139 clients. He found that the Cox proportional hazards model outperforms the logistic regression model based on the new performance criteria.

Bellotti & Crook (2009) studied the effects of the general economic conditions measured by macroeconomic variables on the probability of default of clients using the Cox proportional hazards model, and also compared the predictive performance of the Cox proportional hazards model with time varying covariates with logistic regression models. They used sample application and monthly performance data for a single UK credit card product provided by a UK bank which spanned a period of credit card accounts opened from 1997 to mid-2005. The results showed that the Cox proportional hazards model is competitive in comparison with the logistic regression model as a credit scoring model for prediction and the inclusion of macroeconomic variables showed a statistically significant improvement in predictive performance.

Dirick et al. (2017) studied the performance of several survival analysis techniques used in credit scoring. The datasets used were from five financial institutions in U.K. and Belgium, and consisted mainly of personal loans and loans of small enterprises with varying loan terms. They assessed model performance using three evaluation measures: area under the curve (AUC) in the receiver operating characteristic curve, default time prediction differences and future loan value estimation. The results showed that the Cox proportional hazards model in combination with penalised splines for continuous covariates works better than the other survival models.

The following section gives a summary of what has been discussed in this

chapter.

Chapter Summary

In summary, this chapter focused on a review of related literature on credit scoring models and some models used in predicting probability of default such as logistic regression and survival analysis models. The review showed that the logistic regression and survival analysis models have high predictive power. However, the survival analysis model had the ability to incorporate more information and variables in the model than the logistic regression model. Also, the survival analysis model can provide time based estimates of the probability of default.

There are a lot of assumptions for the logistic regression and survival analysis models such as the linearity assumption. That is in most cases, it is assumed that the covariates affect the response variable in a linear relationship. However, this assumption is not always true. In the review presented above, none of the models presented a test for the linearity assumption. In this study, we seek to build a probability of default model for a microfinance institution in Ghana and determine the validity of the linearity assumption for the fitted model for both logistic regression and survival analysis models.

CHAPTER THREE

RESEARCH METHODS

Introduction

In this chapter, the methods employed in the study are presented. They include logistic regression and survival analysis models. A brief description of the data used for the study is also presented.

Research Design

The research method used in this study is quantitative method. Quantitative methods of research include surveys and experiments. They make use of numeric data, identifies variables to be studied, tests or verifies hypothesis or explanations, among others. This study uses data that contains details of customers of a microfinance institution.

Data

The data used for this study was customers of a microfinance institution in Ghana. The dataset is made up of all individual loans disbursed from January 2012 to January 2019. It consisted of a total of 1060 clients with individual loans. The dataset was cleaned to remove all inputs with incomplete information.

Model Specification

In this section, we discuss the methods used to derive our mathematical model for the probability of default. The methods employed are logistic regression and survival analysis models. We first discuss the logistic regression model.

Logistic Regression

Logistic regression is generally used to describe the relationships that exist between a categorical dependent variable and one or more categorical or continuous independent variables. It applies logit transformation to the dependent variable. That is, the logistic model predicts the logit of the categorical dependent

variable (Y) from the independent variable (X) which can either be continuous or categorical.

Definition 35

The *odds of an event* is the ratio of the probability of an event occurring to the probability of an event not occurring. It is defined as:

$$\text{Odds} = \frac{p}{1-p}, \tag{3.1}$$

where $p \in (0,1)$ is the probability of an event occurring.

Definition 36

The *logit* is defined as the natural logarithm of odds.

Therefore the logit of Y is given as:

$$\text{logit}(Y) = \ln\left(\frac{p}{1-p}\right). \tag{3.2}$$

The logistic regression model can be of the following types: binary and multinomial logistic models.

1. Binary logistic model is a logistic regression model in which the dependent variable is dichotomous. The dependent variable is usually coded as 0 or 1.

2. Multinomial logistic model has a polytomous dependent variable which can be ordered or unordered.

In this study, we consider the binary logistic regression model.

Proposition 2

The binary logistic regression model with a single independent variable (also known as simple logistic regression model) is given as:

$$p = \frac{1}{1+\exp(-\beta_0 - \beta_1 X)}, \tag{3.3}$$

where p is the probability of an event occurring, the β's are the regression coefficients to be estimated and X is the independent variable.

Proof. In deriving Equation (3.3) for predicting the probability of the occurrence of an event of interest, p, we first take the antilog of the right side of Equation (3.2) to obtain:

$$\begin{aligned}
\frac{p}{1-p} &= \exp(\beta_0+\beta_1 X).\\
\text{That is, } p &= \exp(\beta_0+\beta_1 X)(1-p)\\
&= \exp(\beta_0+\beta_1 X) - p(\exp(\beta_0+\beta_1 X)).\\
\text{Thus, } p + p(\exp(\beta_0+\beta_1 X)) &= \exp(\beta_0+\beta_1 X)\\
p(1+\exp(\beta_0+\beta_1 X)) &= \exp(\beta_0+\beta_1 X)\\
\text{Then we get, } p &= \frac{\exp(\beta_0+\beta_1 X)}{1+\exp(\beta_0+\beta_1 X)}\\
p &= \frac{1}{1+\exp(-\beta_0-\beta_1 X)},
\end{aligned}$$

where $0 \leq p \leq 1$. □

Remark 13

Similarly, the logistic regression model with multiple predictors is given as:

$$\begin{aligned}
p &= \frac{\exp(\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\cdots+\beta_n X_n)}{1+\exp(\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\cdots+\beta_n X_n)},\\
&= \frac{1}{1+\exp(-(\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\cdots+\beta_n X_n))}.
\end{aligned} \tag{3.4}$$

Odds Ratio

In this part, we define some concepts that helps to measure effects of the independent variables on the dependent variable in a logistic regression model.

Definition 37

The *odds ratio* compares the odds of two events. In simple terms, an odds ratio is a measure of the extent to which two probabilities of an event occurring differ.

That is, for any two events A and B, the odds ratio of A to B is given as:

$$\begin{aligned}\text{Odds ratio } (A \; vs. B) &= \frac{\text{Odds of } A}{\text{Odds of } B} \\ &= \frac{p_A(1-p_B)}{p_B(1-p_A)},\end{aligned} \tag{3.5}$$

where p_A is the probability of event A occurring and p_B is the probability of event B occurring.

Remark 14 1. In the binary logistic regression model, the regression coefficient, β, is the estimate of an increase in the log odds of the dependent variable as a result of a unit increase in the independent variable.

2. The exponential function of the regression coefficient, $\exp\beta$ is the odds ratio associated with a unit increase in the independent variable.

3. The odds ratio is used to determine whether a particular independent variable has an effect on the dependent variable and compare the various independent variables for that dependent variable.

Proposition 3 1. Suppose X is a categorical variable with two levels, $k = 0, 1$, where zero stands for the reference group and one for the non-reference group the odds ratio for the non-reference group is given as:

$$\text{Odds ratio} = \exp(\beta_1). \tag{3.6}$$

2. Suppose X is a continuous variable, the odds ratio of an increase in an independent variable is given as:

$$\text{Odds ratio} = \exp(\beta_1 \phi). \tag{3.7}$$

Proof. (Collett, 2002)

1. The logistic regression model for a categorical variable with two levels,

$k = 0, 1$, is given as:

$$p_k = \exp(\beta_0 + \beta_1 x_k), \tag{3.8}$$

where x_k is an indicator variable which takes a value of zero for the reference group and one for the non-reference group. The odds ratio for the non-reference group is given as:

$$\begin{aligned} \text{Odds ratio} &= \frac{\text{Odds of } (x_k = 1)}{\text{Odds of } (x_k = 0)} \\ &= \frac{\exp(\beta_0 + \beta_1 x_1)}{\exp(\beta_0 + \beta_1 x_0)} \\ &= \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} \\ &= \exp(\beta_1). \end{aligned} \tag{3.9}$$

2. For a continuous variable, the simple logistic regression model is given as:

$$p_j = \exp(\beta_0 + \beta_1 x_j), \tag{3.10}$$

where x_j is the jth continuous variable. The odds ratio of an increase in the jth variable is given as:

$$\begin{aligned} \text{Odds ratio} &= \frac{\text{Odds of } (x_j + \phi)}{\text{Odds of } (x_j)} \\ &= \frac{\exp\{\beta_0 + \beta_1 (x_j + \phi)\}}{\exp\{\beta_0 + \beta_1 x_j\}} \\ &= \exp(\beta_1 \phi). \end{aligned} \tag{3.11}$$

□

We discuss the assumptions of the logistic regression model in the next section.

Assumptions of the Logistic Regression Model

The logistic regression model has several assumptions. They include:

1. The dependent variable is categorical.
2. Each observation is independent.
3. The independent variables have a linear relationship with the log odds of an event.
4. The binomial distribution is the underlying distribution of the binary logistic regression model.

In the next section, we discuss the underlying distribution of the binary logistic regression model.

Binomial Distribution

The binomial distribution is defined as follows.

Definition 38

Let X be the random variable counting the number of draws in a sample size of n observations. The probability that $X = x$ is given by:

$$\begin{aligned} P(X = x) &= \binom{n}{x} p^x \, (1-p)^{n-x} \\ &= \frac{n!}{x!(n-x)!} p^x \, (1-p)^{n-x}. \end{aligned} \tag{3.12}$$

The underlying distribution of the binary logistic regression model is necessary in determining the parameters of the model. The next section discusses the steps involved in estimating the parameters of the binary logistic regression model.

Estimating the Parameters of the Binary Logistic Regression Model

The maximum likelihood method can be used to estimate the regression coefficients, $\beta_i, i = 1, \cdots, n$ for $n \in \mathbb{N}$, in Equation (3.4).

Definition 39

A *likelihood function* estimates the probability of observing data, given unknown parameters.

Remark 15

The maximum likelihood maximizes the likelihood of reproducing the data given the parameter estimates.

Suppose that each y_i represents a binomial count in the ith population, then the likelihood function is given as:

$$f(\beta \mid y) = \prod_{i=1}^{K} \frac{n_i!}{y_i!(n_i - y_i)!} p_i^{y_i} \ (1 - p_i)^{n_i - y_i}. \tag{3.13}$$

Proposition 4

The maximum likelihood estimates of β that maximizes Equation (3.13) can be obtained by checking the necessary and sufficient conditions for optimality which is given by the following conditions:

1. $\sum_{i=1}^{K} y_i X_{ij} - n_i p_i X_{ij} = 0$.
2. $\sum_{i=1}^{K} n_i X_{ij} p_i (1 - p_i) X_{ij'} < 0$.

Proof. (Czepiel, 2002) Since y_i is a binomial count, its likelihood function is given by Equation (3.13). The factorial terms do not contain any β_i, hence they are treated as constants. Then Equation (3.13) is written as:

$$\prod_{i=1}^{K} \left(\frac{p_i}{1 - p_i} \right)^{y_i} (1 - p_i)^{n_i}. \tag{3.14}$$

Equation (3.4) can be rewritten as:

$$p_i = \frac{\exp(\sum_{j=0}^{n} X_{ij} \beta_j)}{1 + \exp(\sum_{j=0}^{n} X_{ij} \beta_j)}. \tag{3.15}$$

Substituting Equation (3.15) into Equation (3.14) and simplifying, gives:

$$\begin{aligned}&\prod_{i=1}^{K}\Big(\exp(\sum_{j=0}^{n}X_{ij}\beta_j)\Big)^{y_i}\Big(1-\frac{\exp(\sum_{j=0}^{n}X_{ij}\beta_j)}{1+\exp(\sum_{j=0}^{n}X_{ij}\beta_j)}\Big)^{n_i}\\ =&\prod_{i=1}^{K}\Big(\exp(\sum_{j=0}^{n}X_{ij}\beta_j)\Big)^{y_i}\Big(1+\exp(\sum_{j=0}^{n}X_{ij}\beta_j)\Big)^{-n_i}.\end{aligned}\tag{3.16}$$

Then taking the natural logarithm of Equation (3.16) gives the log likelihood function:

$$l(\beta)=\sum_{i=1}^{K}y_i\Big(\sum_{j=0}^{n}X_{ij}\beta_j\Big)-n_i\log\Big(1+\exp(\sum_{j=0}^{n}X_{ij}\beta_j)\Big).\tag{3.17}$$

The critical points of Equation (3.17) can be found by setting the first derivative with respect to each β to zero. Thus,

$$\begin{aligned}\frac{\partial l(\beta)}{\partial(\beta_j)}&=\sum_{i=1}^{K}y_iX_{ij}-n_i\frac{1}{1+\exp(X_{ij}\beta_j)}\exp\Big(\sum_{j=0}^{n}X_{ij}\beta_j\Big)X_{ij}\\ &=\sum_{i=1}^{K}y_iX_{ij}-n_ip_iX_{ij}.\end{aligned}\tag{3.18}$$

Setting $\frac{\partial l(\beta)}{\partial(\beta_j)}$ to zero gives the critical point. The critical point of Equation (3.18) can either be a maximum or a minimum. The critical point will be a maximum if every element on the diagonal of the matrix of the second partial derivatives is less than zero (negative definite). The general form of the matrix of the second partial derivatives is given as:

$$\begin{aligned}\frac{\partial^2 l\,(\beta)}{\partial\beta_j\partial\beta_{j'}}&=\frac{\partial}{\partial\beta_{j'}}\sum_{i=1}^{K}y_iX_{ij}-n_ip_iX_{ij},\\ &=-\sum_{i=1}^{K}n_iX_{ij}\frac{\partial}{\partial\beta_{j'}}\Big(\frac{\exp(\sum_{j=0}^{n}X_{ij}\beta_j)}{1+\exp(\sum_{j=0}^{n}X_{ij}\beta_j)}\Big),\\ &=-\sum_{i=1}^{K}n_iX_{ij}p_i(1-p_i)X_{ij'}.\end{aligned}\tag{3.19}$$

□

The estimates of β_j can be computed, after verifying that the matrix of the

second partial derivatives is negative, using an iterative process known as the Newton-Raphson method.

The Newton-Raphson method is given as follows:

Step 1. Start with an initial guess for the solution, that is $\beta_j^{(0)}$ which represents the vector of initial approximation for each β_j .

Step 2. Use the first two terms of the Taylor polynomial evaluated at the initial guess to get an estimate that is closer to the solution. Let $l'(\beta_j)$ denote $\frac{\partial l(\beta)}{\partial(\beta_j)}$. Then we have

$$\beta_j^{(1)} = \beta_j^{(0)} + [l''(\beta_j^{(0)})]^{-1} l'(\beta_j^{(0)}). \tag{3.20}$$

Step 3. Step 2 continues until there is no change in the elements of β_j from one iteration to the next.

The parameters in the logistic regression model help in predicting the dependent variable given some information about the independent variables. These predicted values are usually compared with the actual or observed values using residuals. The following section considers the concept of residuals in logistic regression analysis.

Residuals for Logistic Regression

Residuals measure the agreement between an observation on the independent variable and a corresponding fitted value. The residuals for the logistic regression model include Pearson, deviance, and partial residuals.

Definition 40

Pearson residuals measure the contribution of each observation to a statistic, which is a summary measure of the goodness of fit of the logistic regression model. It is defined as:

$$X_i = \frac{y_i - n_i\hat{p}_i}{\sqrt{\{n_i\hat{p}_i(1-\hat{p}_i)\}}}, \tag{3.21}$$

where n_i is the number of the different responses and $\hat{p}_i$ is the predicted probability of event.

Remark 16

The Pearson χ^2 statistic is given as:

$$X^2 = \sum X_i^2. \tag{3.22}$$

Definition 41

Deviance residuals also measure the contribution of each observation to the logistic regression model. Deviance residuals are defined as:

$$d_i = \text{sign}(y_i - \hat{y}_i)\left\{2y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + 2(n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)\right\}, \tag{3.23}$$

where $\text{sign}(y_i - \hat{y}_i)$ is the function that makes d_i positive when $y_i \geq \hat{y}_i$ and negative when $y_i < \hat{y}_i$. The total deviance is given as $D = \sum d_i^2$.

Definition 42

Partial residual plots are used to determine whether an independent variable should be transformed to give a non-linear term or not. The test statistic is defined as:

$$r_{ik} = \frac{y_i - n_i\hat{p}_i}{\{n_i\hat{p}_i(1-\hat{p}_i)\}} + \hat{\beta}_k x_{ki}, \tag{3.24}$$

where $\hat{\beta}_k$ is the coefficient of the kth independent variable in the model.

The main objective of this study is to develop a mathematical model for predicting the probability of default of clients of a microfinance institution. The logistic regression model is usually used in developing such models. However, the model does not take into account the time dynamics. That is, the model cannot compute the probability of default at a specific time. This makes it difficult to use the model when we want to include time characteristics. As a result of

this major drawback, we employ the survival analysis model which allows these time characteristics to be included.

Survival Analysis

In this subsection, we discuss the survival analysis model. Survival analysis is a collection of specialized tools that focus on time and duration until an event of interest occurs. The term “failure" is used to the occurrence of an event. Depending on where it is being applied, survival analysis can be called reliability analysis, lifetime data analysis, time to event analysis or event history analysis. The dependent variable is the time until an event of interest occurs. It can be continuous or discrete. Most of the concepts discussed in this section were borrowed from (Harrell Jr, 2015; Kleinbaum & Klein, 2010; Liu, 2012) and (Leung et al., 1997). There are several concepts used in survival analysis described as follows.

Definition 43

Survival time is the time taken for an event of interest to occur and usually denoted by a random variable T which takes values that are greater than or equal to zero, $(T \geq 0)$, since time is non-negative.

The actual survival time, denoted as t, is a value of T. The values of T have a probability distribution with a probability density function $f(t)$ and a cumulative distribution function $F(t)$. The cumulative distribution function for T is defined as:

$$F(t) = P(T \leq t) = \int_0^t f(u)\mathrm{d}u. \tag{3.25}$$

An important function related to the survival time is the survival function.

Definition 44

Survival function is the probability that a subject survives beyond a specified

time t and denoted as $S(t)$. It is given as:

$$S(t) = P(T > t) = 1 - F(t) = \int_t^{\infty} f(u)\mathrm{d}u. \tag{3.26}$$

According to Kleinbaum & Klein (2010), survival functions have these theoretical properties: they are non-increasing functions, that is at $t = 0, S(t) = 1$ and at $t = \infty, S(t) = 0$. This implies, $S(t) \in [0, 1]$.

The survival function is related to an important function in survival analysis known as the hazard function.

Definition 45

The *hazard function* measures the potential failure at time t given that the subject has survived up to time t. It is given as:

$$h(t) = \lim_{\Delta t \downarrow 0} \frac{P(t < T \leq t + \Delta t \mid T > t)}{\Delta t}, \tag{3.27}$$

where Δt is a small time interval.

Remark 17

The hazard function is not a probability. It can be expressed as a ratio between the density and survival functions.

Proposition 5

Let $f(t)$ and $S(t)$ be the density and the survival functions respectively of an arbitrary distribution. The hazard function can be expressed in terms of $f(t)$ and $S(t)$ as:

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ &= -\frac{\partial \log S(t)}{\partial t}. \end{aligned}$$

Proof. (Harrell Jr, 2015) Using the law of conditional probability, it follows

from Proposition 5 that:

$$\begin{aligned} h(t) &= \lim_{\Delta t \downarrow 0} \frac{1}{\Delta t} \frac{P(t < T \leq t + \Delta t)}{P(T > t)} \\ &= \frac{\lim\limits_{\Delta t \downarrow 0} \frac{P(t<T\leq t+\Delta t)}{\Delta t}}{P(T > t)}. \end{aligned}$$

From first principle approach, we obtain

$$\begin{aligned} h(t) &= \frac{\frac{P(T\leq t)-P(t\geq t+\Delta t)}{\Delta t}}{P(T > t)} \\ &= \frac{\frac{1-P(T<t)-(1-P(T<t+\Delta t))}{\Delta t}}{S(t)} \\ &= \frac{P(T < t + \Delta t) - P(T < t)}{S(t)} \\ &= \frac{f(t)}{S(t)}. \end{aligned}$$

In addition we have,

$$\begin{aligned} h(t) &= \frac{\frac{\partial F(t)}{\partial t}}{S(t)} \\ &= \frac{\frac{\partial (1-S(t))}{\partial t}}{S(t)} \\ &= -\frac{\partial \log S(t)}{\partial t}. \end{aligned} \tag{3.28}$$

□

The hazard function is preferred to the survival function because it gives more information about understanding the mechanism of failure and it helps to determine the underlying statistical model of a dataset.

The cumulative hazard function is the accumulation of the hazard function up to time t. It is given as:

$$H(t) = \int_0^t h(u)\mathrm{d}u. \tag{3.29}$$

The relationship between the cumulative hazard function and the survival function is given as:

$$H(t) = -\log(S(t)). \tag{3.30}$$

An important concept related to the hazard function is the hazard ratio.

Definition 46

Hazard ratio is the ratio of the risk of event occurring in one group at a given time compared with another group at that very time. That is, the ratio of the hazard rates between two groups. It measures the effect of the independent variables in the model. It is similar to the odds ratio in logistic regression. For independent variables X_1 and X_2, the hazard ratio is given as:

$$HR(t \mid X_1, X_2) = \frac{h(t \mid X_1)}{h(t \mid X_2)}. \tag{3.31}$$

Remark 18

If $HR = 1$, then there is no effect. If $HR > 1$, then there is an increase in the hazard and a reduction in the survival. If $HR < 1$, then there is a reduction in the hazard and an increase in the survival.

A feature of survival analysis that makes it unique from other statistical models is the ability to incorporate censored information.

Censoring

In this section we discuss the meaning, types and mechanisms of censoring. Censoring is a major feature that differentiates survival analysis from the other statistical techniques used in analysing data. It occurs when there is incomplete information about the survival time of the subjects in a study. Censoring usually occurs when a subject does not experience the event of interest before the study ends or is lost to follow-up during the study period or withdraws from the study. It is usually denoted by C. Censoring can be either fixed or random.

Fixed censoring occurs when either time or the number of failures is fixed. There are two types of fixed censoring.

1. Type I censoring occurs when every subject is observed or followed until a specified time t.

2. Type II censoring occurs when a study is terminated after a predetermined number of failures are observed, that is k failures out of n observations.

Random censoring occurs when subjects who are censored at time t are a representative of those remain at risk at time t considering their survival experience. That is, the rate of failure for the censored group should be the same as the rate of failure for the uncensored group.

A subject's survival time can be considered as right, left or interval censored. Right censoring occurs when the actual survival time of a subject exceeds the observed survival time of the study. Left censoring occurs when the actual survival time of a subject is less than the observed survival time of the study. Interval censoring occurs when the actual survival time of a subject occurs within a time interval.

The censoring mechanism can also be independent and non-informative. Independent censoring occurs when the censored subjects in a subgroup of interest at time t are a representative of all the subjects still at risk in that subgroup in terms of their survival experience. Mathematically, independent censoring occurs when the distribution of the survival times, denoted T, is independent of the distribution of the censoring times, denoted C. Non-informative censoring occurs when T provides no information about C.

Estimating the Specialised Functions

In this part, we discuss how to estimate the specialised functions used in survival analysis. These functions include the hazard and survival functions.

1. Hazard function. The Nelson-Aalen estimator is used to estimate the cu-

mulative hazard function. It is defined as:

$$\hat{H}(t) = \begin{cases} 0, & \text{if } t \leq t_i \\ \sum_{t_i \leq t} \frac{d_i}{n_i}, & \text{if } t_i \leq t \end{cases}, \tag{3.32}$$

where d_i is the number of events that occur at time t_i and n_i is the number of subjects at risk at time t_i and $i \in \mathbb{N}$.

2. Survival Function. The survival function can be estimated using nonparametric, semi-parametric and parametric models. Nonparametric models do not make any assumption about the shape of the hazard function and the effect of the independent variables on the shape of the hazard function. Semi-parametric models, like the nonparametric models do not make any assumption about the hazard function but makes assumptions about the effect of the independent variables on the shape of the hazard function. Parametric models make assumptions about the shape of the hazard function and the effect of the independent variables on the shape of the hazard function. In this study, we consider the Kaplan Meier estimator as (as a nonparametric model) and the Cox proportional hazards regression model (as a semi-parametric model).

   (i) The Kaplan Meier Estimator.

   Definition 47

   It estimates the survival function using both censored and uncensored data. It is also called the product limit estimator. For all values of t where there is data, it is defined as:

$$\hat{S}(t) = \begin{cases} 1, & t < t_1, \\ \prod_{t_i \leq t}[1 - \frac{d_i}{n_i}], & t_i \leq t, \end{cases} \tag{3.33}$$

   where d_i is the number of events that occur at time i and n_i is the

number of subjects at risk at time i.

The survival times are ordered in ascending order, before estimating the survival function. At any given time, the estimator is obtained by multiplying a sequence of estimated conditional survival probabilities. The graph of a Kaplan-Meier estimator, called the survival curve, is a step function with jumps at each observed event. The size of the jumps depend on the number of events at each time t_i and the pattern of censored observations before t_i The Kaplan-Meier estimator estimates the survival function without independent variables.

The Kaplan-Meier estimator can also be used to estimate the cumulative hazard function as follows:

$$\hat{H}(t) = -\ln[\hat{S}(t)]. \tag{3.34}$$

(ii) Cox Proportional Hazards Regression Model

Definition 48

It is used to model the relationship between the hazard function and independent variables. The Cox proportional hazards model is used to test the effect of independent variables on the survival times of different subjects. It is usually stated in terms of the hazard function as:

$$h(t \mid X) = h_0(t) \exp \sum_{i=1}^{n} (X_i \beta_i), \tag{3.35}$$

where $h_0(t)$ is the baseline hazard when there are no independent variables, X_is are the independent variables and β_is are the regression coefficients to be estimated.

The cumulative hazard function of the Cox model is given as:

$$H(t) = \int_0^t h_0(u) \exp \sum_{i=1}^{n} (\beta_i x_i) \mathrm{du} \tag{3.36}$$
$$= \exp \sum_{i=1}^{n} \beta_i x_i \left(\int_0^t h_0(u) \mathrm{du} \right),$$

where $h_0(u)$ is the baseline hazard function.

The survival function of the Cox regression model is given as:

$$S(t) = \exp \left(- \int_0^t h_0(u) \mathrm{du} \exp \sum_{i=1}^{n} (\beta_i x_i) \right)$$
$$= \exp \left(- H_0(t) \exp \sum_{i=1}^{n} (\beta_i x_i) \right).$$

$$\text{But } S_0(t) = \exp(-H_0(t)).$$

$$\text{Then } S(t) = S_0(t)^{\exp \sum_{i=1}^{n} (\beta_i x_i)}, \tag{3.37}$$

where $S(t)$ is the survival function at time t, $S_0(t)$ is the baseline survival function at time t and $H_0(t)$ is the baseline cumulative hazard function at time t .

A measure of effect of the independent variables in the Cox regression model on the dependent variable is the hazard ratio. In the next section, we discuss the hazard ratio.

Proposition 6

The hazard ratio in a Cox regression model estimates the effect of each independent variable on the dependent variable in the model.

1. Suppose X is a categorical variable with two levels, $p = 0, 1$, where zero stands for the reference group and one for the non-reference group the hazard ratio for the non-reference group is given as:

$$\text{Hazard ratio} = \exp(\hat{\beta}_p). \tag{3.38}$$

2. Suppose X is a continuous variable, the hazard ratio of an increase in an independent variable by ω is given as:

$$\text{Hazard ratio} = \exp(\omega\hat{\beta}_c). \tag{3.39}$$

Proof. (Collett, 2002)

1. Consider a dichotomous variable x_p with $x_{p1} = 1$ and $x_{p0} = 0$ and assume all the other covariates in the model are equal to zero. The hazard ratio is given as:

$$\begin{aligned} HR_p &= \frac{h_0(t)\exp(x_{p1}\hat{\beta}_p)}{h_0(t)\exp(x_{p0}\hat{\beta}_p)} \\ &= \exp[(x_{p1} - x_{p0})\hat{\beta}_p] \\ &= \exp(\hat{\beta}_p), \end{aligned} \tag{3.40}$$

where $\hat{\beta}_p$ is the regression coefficient estimate for the covariate x_p.

2. Consider two successive integer values, x_c and $x_c + \omega$ for a continuous independent variable, the hazard ratio is given as:

$$\begin{aligned} HR_{(x_c, x_c+\omega)} &= \frac{h_0(t)\exp[(x_c + \omega)\hat{\beta}_c]}{h_0(t)\exp(x_c\hat{\beta}_c)} \\ &= \exp[(x_c + \omega - x_c)\hat{\beta}_c] \\ &= \exp(\omega\hat{\beta}_c). \end{aligned} \tag{3.41}$$

□

Remark 19

For a continuous independent variable, the hazard ratio shows the extent to which the risk increases ($HR > 1$) or decreases ($HR < 1$).

The Cox proportional hazards regression model is based on certain assumptions. These assumptions are discussed below.

Assumptions of the Cox Proportional Hazards Regression Model

In this section, we discuss the assumptions of the Cox proportional hazards regression model. The Cox proportional hazards regression model is based on the following assumptions:

1. the covariates acts multiplicatively on the hazard at each point in time
2. the hazard of the event in a group is a constant multiple of the hazard in any other group (proportionality assumption)

Estimating Parameters in Cox Proportional Hazards Regression Model (No tied events)

The parameters in the Cox proportional hazards regression model are estimated using a partial likelihood but not a full likelihood. This is due to the assumption of proportional hazards. That is, the baseline hazard do not change and cancels out in the likelihood function since they are constants thereby resulting in a partial likelihood.

Definition 49

Tied events are events with the same survival times.

Consider a sample of n individuals ordered according to the rank of their survival times, $t_1 < t_2 < t_3 < \cdots < t_n$, and assuming there are no observation ties. Let $R(t_j)$ be the risk set which consists of all individuals who do not experience the event and are uncensored just before t_j. The conditional probability that individual j experiences the event given the risk set $R(t_j)$ in a discrete time interval $(t_j, t_j + \Delta t_j)$ is given as:

$$\frac{P[\text{individual } j \text{ with covariates } x_j \text{ experiences event in interval } (t_j, t_j + \Delta t_j)]}{\sum_{k \in R(t_j)} P[\text{individual } k \text{experiences event in } (t_j, t_j + \Delta t_j)]}. \tag{3.42}$$

The intervals between the successive survival times do not provide information on the multiplicative effects of the covariates (independent variables) on the hazard function since the baseline hazard has an arbitrary form in Cox regression.

Without specifying the distribution of the baseline hazard ($h_0(t)$), the incomplete hazard function is given as:

$$\begin{aligned}\frac{h(t_j;x_j)}{\sum_{k\in R(t_j)} h(t_j;x_k)} &= \frac{h_0(t_j)\exp(x_j^T\beta)}{\sum_{k\in R(t_j)} h_0(t_j)\exp(x_k^T\beta)},\\ &= \frac{\exp(x_j^T\beta)}{\sum_{k\in R(t_j)}\exp(x_k^T\beta)},\end{aligned} \tag{3.43}$$

where x is a vector of the independent variables, and β is a vector of regression coefficients.

Remark 20

Let t be a continuous function, then the conditional probability changes into a continuous hazard function given that $\Delta t_i \to 0$. It is given as:

$$\frac{\text{hazard rate at } t_j \text{ for individual } j \text{ with covariates } x_j}{\sum_{k\in R(t_j)} \text{hazard rate at } t_j \text{ for individual } k}. \tag{3.44}$$

The joint likelihood for β is given as the product of Equation (3.43) over all the t_j values as:

$$L(\beta) = \prod_{j=1}^{n}\left[\frac{\exp(x_j^T\beta)}{\sum_{k\in R(t_j)}\exp(x_k^T\beta)}\right]^{\delta_j}, \tag{3.45}$$

where δ_j is the censoring indicator (such that $\delta_j = 1$ if t_j is an event time and $\delta_j = 0$ if t_j is a censored time). When $\delta_j = 1$, Equation (3.45) is the conditional probability of an event occurring given the risk set $R(t_j)$. When $\delta_j = 0$, Equation (3.45) is equal to 1, implying that the product of all right censored observations is equal to 1 which does not make any contribution to the partial likelihood. Therefore, right censored observations cannot be accounted for in the partial likelihood. Equation (3.45) can be simplified by multiplying the conditional

probabilities over all events as:

$$L(\beta) = \prod_{j=1}^{q} \left[\frac{\exp(x_j^T \beta)}{\sum_{k \in R(t_j)} \exp(x_k^T \beta)} \right], \tag{3.46}$$

where q is the total number of events, ordered by rank. This simplification does not mean that the censored observations have been excluded since at each t_j the risk set, $R(t_j)$, contains all observations censored at times later than t_j.

Proposition 7

For right censored data the maximum likelihood estimates of β that maximizes Equation (3.46) can be obtained by computing the first and second order derivatives given by the following conditions:

1. $\sum_{i=1}^{d} \Big(x_{jp} - \mathbb{E}[x_p \mid R(t_j)] \Big) = 0.$
2. $\sum_{j=1}^{d} \{\mathbb{E}[x_{jp} x_{jp'} \mid R(t_j)]\} - \sum_{j=1}^{d} \{\mathbb{E}[x_{jp} \mid R(t_j)] \mathbb{E}[x_{jp'} \mid R(t_j)]\} < 0.$

Proof. (Liu, 2012) For right censored data, the partial likelihood function is given by Equation (3.46). Then the log partial likelihood is given as:

$$\log L(\beta) = \sum_{j=1}^{q} \left[(x_j^T \beta) - \sum_{j=1}^{q} \log \Big(\sum_{k \in R(t_j)} \exp(x_k^T \beta) \Big) \right]. \tag{3.47}$$

The first partial derivative of the log partial likelihood known as the score statistic, denoted by $U(\beta)$, is given as:

$$\begin{aligned} U(\beta) &= \frac{\partial \log L(\beta)}{\partial \beta} \\ &= \sum_{i=1}^{d} \left(x_j - \left[\frac{\sum_{k \in R(t_j)} x_k \exp(x_k^T \beta)}{\sum_{k \in R(t_j)} \exp(x_k^T \beta)} \right] \right) \\ &= \sum_{i=1}^{d} \Big(x_j - \mathbb{E}[x \mid R(t_j)] \Big). \end{aligned} \tag{3.48}$$

The coefficients of the vector β can be estimated by equating Equation (3.48) to zero.

The second partial derivative known as the information matrix, denoted by $I(\hat{\beta})$ is given as:

$$I(\hat{\beta}) = -\Big(\frac{\partial^2 \log L(\beta)}{\partial \beta \beta'}\Big) \mid_{\beta=\hat{\beta}}, \tag{3.49}$$

with $\hat{\beta}$ such that $I(\hat{\beta})$ given as:

$$I(\hat{\beta}) = \sum_{j=1}^{d} \{\mathbb{E}[x_j x_{j'} \mid R(t_j)]\} - \sum_{j=1}^{d} \{\mathbb{E}[x_j \mid R(t_j)]\mathbb{E}[x_{j'} \mid R(t_j)]\}, \tag{3.50}$$

where

$$\mathbb{E}[x_j x_{j'} \mid R(t_j)] = \frac{\sum_{k \in R(t_j)} x_k x_{k'} \exp(x_k^T \beta)}{\sum_{k \in R(t_j)} \exp(x_k^T \beta)}. \tag{3.51}$$

□

There are instances where the survival times of the events in the Cox Proportional Hazards Regression model will be tied. This is discussed in the section below.

Estimating Parameters in Cox Proportional Hazards Regression Model (Tied events)

In this section, we discuss how parameters in the Cox proportional hazards regression model will be estimated in the case of tied events. Specifically, there are four approaches that have been adopted to estimate parameters in the Cox proportional regression hazards model with tied events. These approaches are discussed below.

1. Discrete-time Cox proportional hazards regression model. The partial likelihood function to a discrete-time regression model is a proposed method for handling tied events. The model is given by:

$$\frac{h(t;x)dt}{1-h(t;x)dt} = \exp(x^T \beta) \frac{h_0(t)dt}{1-h_0(t)dt}. \tag{3.52}$$

The likelihood function is given as:

$$L^{\text{discrete}}(\beta) = \prod_{i=1}^{n} \frac{\exp(s_i^T \beta)}{\sum_{k \in R_{d_i}(t_i)} \exp(s_k^T \beta)}. \tag{3.53}$$

The maximum likelihood method can be used to estimate the regression coefficients in Equation (3.48) in a similar manner as the Cox proportional hazards model without ties. The drawback of this model is that the partial likelihood is difficult to compute when the number of ties is large.

2. Exact or average partial likelihood method. It considers all the possible orderings and takes the average. The likelihood function for average partial likelihood for all survival times is given as:

$$L^{\text{exact}}(\beta) \propto \prod_{i=1}^{n} \left(\exp\left[\left(\sum_{c \in D_i} x_c \right)^T \beta \right] \left\{ \sum_{v \in A_i} \prod_{r=1}^{d_i} \left[\sum_{c \in R(t_i, v, r)} \exp x_c^T \beta \right] \right\}^{-1} \right). \tag{3.54}$$

This partial likelihood is difficult to compute if there is a large number of ties at given observed survival times.

3. Breslow method. The likelihood function is given as:

$$L^{\text{Breslow}}(\beta) = \prod_{i=1}^{n} \left[\frac{\exp\left[\left(\sum_{k \in D_i} x_k \right)^T \beta \right]}{\sum_{k \in R_{(t_i)}} \exp(x_k^T \beta)^{d_i}} \right]. \tag{3.55}$$

4. Efron method. This method improves upon the Breslow method. The likelihood function is given as:

$$L^{\text{Efron}}(\beta) = \prod_{i=1}^{n} \frac{\exp\left[\left(\sum_{k \in D_i} x_k \right)^T \beta \right]}{\prod_{r=0}^{d_i - 1} \left[\sum_{k \in R_{(t_i)}} \exp(x_k^T \beta) - \frac{r-1}{d_i} \sum_{k \in R_{(t_i)}} \exp(x_k^T \beta) \right]}. \tag{3.56}$$

This method is usually preferred to the other three methods.

Remark 21

The Efron, Breslow and exact methods of estimating parameters approximate methods for handling ties in the Cox proportional hazards model.

Generally, events often occur at random. This notion of randomness is usually associated with stochastic processes. The counting process approach makes it possible to describe some stochastic processes associated with survival analysis.

Counting Processes in Survival Analysis

In this section, the counting process models for survival analysis are discussed. Counting processes for survival analysis are discussed extensively in Liu (2012) and Fleming & Harrington (2011). In survival analysis, the counting process approach is used to explain some of the concepts of regression residuals of the Cox regression model and the occurrence of repeated events.

Suppose T is an event time variable and C is a censored time variable. Let T and C be nonnegative, independent random variables and assume that the distribution of T has a density. The variable $\tilde{T} = \min(T,C)$, is a censored observation of the event time variable T and $\delta = \mathbb{1}_{\{T \leq C\}}$ is the indicator variable for the event of an uncensored observation of T. The counting process $\{N(t); t \geq 0\}$ at time t is given by:

$$N(t) = \mathbb{1}_{\{\tilde{T} \leq t, \delta = 1\}} = \delta \mathbb{1}_{\{T \leq t\}}. \tag{3.57}$$

The counting process is basic to the martingale approach to censored data.

The at risk process, $Y(t)$ is given by:

$$Y(t) = \mathbb{1}_{\{\tilde{T} \geq t\}}. \tag{3.58}$$

Proposition 8

Let T and C be nonnegative, independent random variables. Suppose T is a

continuous function whose distribution has a density. The intensity function at t is given as:

$$\lambda(t) \approx \left[\frac{\mathbb{E}[N[(t+\Delta t)-] - N(t-) \mid T \geq t, C \geq t]}{\Delta t}\right]. \tag{3.59}$$

Proof. (Fleming & Harrington, 2011) Let $F(t) = P(T \leq t)$, $S(t) = 1 - F(t)$, and $\lambda(t) = \frac{-\mathrm{d}\log(S(t))}{\mathrm{d}t}$.

Since T and C are independent,

$$\begin{aligned}
\lambda(t) &= \frac{-\mathrm{d}\log(S(t))}{\mathrm{d}t} \\
&= \left[-S(t)^{-1}\left\{-\frac{\mathrm{d}}{\mathrm{d}t}S(t)\right\}\right] \\
&= \left[[P(T>t)]^{-1}\left\{\frac{\mathrm{d}}{\mathrm{d}t}P(T>t)\right\}\right] \\
&= \left[[P(T>t)]^{-1}P(t \leq T < t+\Delta t)\right] \\
&= \lim_{\Delta t \downarrow 0}\frac{1}{\Delta t}\{P(t \leq T < t+\Delta t)\}[P(T>t)]^{-1} \\
\lambda(t) &= \lim_{\Delta t \downarrow 0}\frac{P(t \leq T < t+\Delta t \mid T \geq t, C \geq t)}{\Delta t}.
\end{aligned} \tag{3.60}$$

It follows that,

$$\lambda(t)\Delta t + o(\Delta t) = \lim_{\Delta t \downarrow 0} P(t \leq T < t+\Delta t \mid T \geq t, C \geq t).$$

$$\text{Since } N(t-) = \lim_{s \uparrow t} N(s)$$

$$\lambda(t)\Delta t \approx P(N((t+\Delta t)-) - N(t-) = 1 \mid T \geq t, C \geq t).$$

Since $N((t+\Delta t)-) - N(t-)$ is a $0, 1-$ valued random variable,

$$\lambda(t)\Delta t \approx \mathbb{E}[N((t+\Delta t)-) - N(t-) \mid T \geq t, C \geq t].$$

$$\text{Then, } \lambda(t) \approx \left[\frac{\mathbb{E}[N[(t+\Delta t)-] - N(t-) \mid T \geq t, C \geq t]}{\Delta t}\right].$$

□

For a single event, the intensity function for counting processes is equivalent to the hazard function.

Proof. Given that T is independent of C, From Proposition 8, it follows that

$$\lambda(t)\Delta t + o(\Delta t) = \lim_{\Delta t \downarrow 0} P(t \leq T < t + \Delta t \mid T \geq t, C \geq t). \tag{3.61}$$

Since $N(t-) = \lim_{s \uparrow t} N(s)$ □

Definition 50

Let T represent an event time random variable with arbitrary distribution function $F(t) = P(T \leq t)$. The cumulative hazard function Λ for T is given by:

$$\Lambda(t) = \int_0^t \frac{\mathrm{d}F(u)}{1 - F(u-)}. \tag{3.62}$$

When $F(u) - F(u-) > 0$, $\Delta\Lambda(u) = P(T = u \mid T \geq u)$, which is the natural definition for a hazard function at points of positive probability for T.

Lemma 2

Suppose $S(t)$ is the survival function and $F(t)$ is the distribution function, the hazard function can be expressed as:

$$\lambda(t) = \frac{-\mathrm{d}\log S(t)}{\mathrm{d}t} = \frac{\mathrm{d}F(t)}{\mathrm{d}S(t)}. \tag{3.63}$$

Proof. (Fleming & Harrington, 2011). Let $\lambda(t) = \frac{-\mathrm{d}\log S(t)}{\mathrm{d}t}$. It follows that

$$\begin{aligned}
\lambda(t)\mathrm{d}t &= -\,\mathrm{d}\log S(t) \\
&= -\left[(S(t))^{-1}[-\mathrm{d}S(t)]\right] \\
&= -\left[(S(t))^{-1}[\mathrm{d}(1 - F(t))]\right] \\
\lambda(t)\mathrm{d}t &= \frac{\mathrm{d}F(t)}{\mathrm{d}S(t)}.
\end{aligned}$$

□

Theorem 8

The random variable which approximates the number of events at each fixed

time t by N over $(0,t]$ is the process A given by

$$A(t) = \int_0^t \mathbb{1}_{\{\tilde{T}>t\}} \lambda(u)\mathrm{d}u.$$

$$\text{Then,} \quad \mathbb{E}[N(t)] = \mathbb{E}[A(t)]. \tag{3.64}$$

Proof. (Fleming & Harrington, 2011). Let $C(c) = P(C > c)$, $F(t) = P(T \leq t)$, and $S(t) = 1 - F(t)$. Then, from Equation 3.57,

$$\begin{aligned}
\mathbb{E}[N(t)] =& P(\tilde{T} \leq t, \delta = 1) \\
=& P(T \leq t, T \leq C) \\
=& \int_0^t C(c-)\mathrm{d}F(c) \\
=& \int_0^t C(c-)S(c)\frac{\mathrm{d}F(c)}{S(c)}.
\end{aligned}$$

From Lemma 2, it follows that

$$\begin{aligned}
\mathbb{E}[N(t)] =& P(\tilde{T} \geq c)\lambda(c)\mathrm{d}c \\
=& \mathbb{E}\int_0^t \mathbb{1}_{\{\tilde{T}\geq c\}}\lambda(c)\mathrm{d}c \\
=& \mathbb{E}[A(t)].
\end{aligned}$$

□

In the next section the martingale theory approach in counting processes is discussed.

The Martingale Theory

The counting process approach to survival analysis is closely related to martingale theory. This is as a result of the fact that methods in martingale theory are used to derive the properties of statistical estimates and test procedures in counting processes. In this section, some concepts in martingale theory are explained.

Definition 51

The process $M(t)$, is a *martingale* for a censored event time variable T, given by the process:

$$\begin{aligned} M(t) =& \mathbb{1}_{\{\tilde{T}\leq t,\delta=1\}} - \int_0^t \mathbb{1}_{\{\tilde{T}\geq u\}}\lambda(u)\mathrm{d}u \\ =& N(t) - A(t). \end{aligned} \tag{3.65}$$

Let the information in $N(u)$ and $\mathbb{1}_{\{\tilde{T}\leq u,\delta=0\}}$ up to time t be

$$\mathscr{F}_t = \sigma\{N(u), \mathbb{1}_{\{\tilde{T}\leq u,\delta=0\}} : 0 \leq u \leq t\}, \tag{3.66}$$

and the information in $N(u)$ and $\mathbb{1}_{\{\tilde{T}\leq u,\delta=0\}}$ up to, but not including time t be

$$\mathscr{F}_{t-} = \sigma\{N(u), \mathbb{1}_{\{\tilde{T}\leq u,\delta=0\}} : 0 \leq u < t\}. \tag{3.67}$$

Suppose T and C are independent, then

$$\mathbb{E}[\mathrm{d}N(t) \mid \mathscr{F}_{t-}] = \mathbb{1}_{\{\tilde{T}\geq t\}}\lambda(t)\mathrm{d}(t) = \mathrm{d}A(t), \tag{3.68}$$

$$\text{and } \mathbb{E}[\mathrm{d}A(t) \mid \mathscr{F}_{t-}] = \mathbb{1}_{\{\tilde{T}\geq t\}}\lambda(t)\mathrm{d}(t) = \mathrm{d}A(t). \tag{3.69}$$

The change in $M = N - A$ over an infinitesimal interval $(t - \mathrm{d}t, t]$ is given as

$$\mathrm{d}M(t) = \mathrm{d}N(t) - \mathrm{d}A(t). \tag{3.70}$$

$$\begin{aligned} \mathbb{E}[\mathrm{d}M(t) \mid \mathscr{F}_{s-}] =& \mathbb{E}[\mathrm{d}N(t) - \mathrm{d}A(t) \mid \mathscr{F}_{s-}] \\ =& \mathbb{E}[\mathrm{d}N(t) \mid \mathscr{F}_{s-}] - \mathbb{E}[\mathrm{d}A(t) \mid \mathscr{F}_{s-}]. \end{aligned}$$

From Equations 3.68 and 3.69,

$$\mathbb{E}[\mathrm{d}M(t) \mid \mathscr{F}_{s-}] = 0. \tag{3.71}$$

In the next section, the nonparametric estimation of the survival function

using counting processes based on martingale theory is discussed.

Nonparametric Estimation of the Survival Function in Counting Processes

In this section, the nonparametric estimation of the survival function with respect to the martingale theory is discussed. The Nelson-Aalen and Kaplan-Meier estimators are the nonparametric estimators of the survival function. These estimators are discussed below.

1. Nelson-Aalen Estimator. This is a nonparametric estimator of the cumulative hazard rate for right censored data.The martingale theory can be used to formulate the Nelson-Aalen estimator as follows.

   Proposition 9

   Let $\tilde{N}(t) = \sum_{i=1}^{n} N_i(t)$ and $\tilde{Y}(t) = \sum_{i=1}^{n} \mathbb{1}_{\{\tilde{T}_i \geq t\}}$. The Nelson-Aalen estimator $\hat{\Lambda}(t)$ for counting processes is given by:

   $$\hat{\Lambda}(t) = \int_0^t \frac{\mathrm{d}\tilde{N}(u)}{\tilde{Y}(u)}. \tag{3.72}$$

   Proof. (Fleming & Harrington, 2011). Let $M_i(t) = N_i(t) - \int_0^t Y_i(s)\mathrm{d}\Lambda(s)$ be a martingale for each i with respect to $\{\mathscr{F}_t : t \geq 0\}$. Then, $\mathscr{M}(t) = \tilde{N}(t) - \int_0^t \tilde{Y}(s)\mathrm{d}\Lambda(s)$. Since the process given at time t by:

   $$\frac{\mathbb{1}_{\{\tilde{Y}(t)>0\}}}{\tilde{Y}(t)} = \begin{cases} \frac{1}{\tilde{Y}(t)} & \text{if } \tilde{Y}(t) > 0, \\ 0 & \text{if } \tilde{Y}(t) = 0 \end{cases}$$

   is a càglàd adapted process, and $\{\mathscr{M}(t) : t \geq 0\}$ given by:

   $$\begin{aligned} \mathscr{M}(t) &= \int_0^t \frac{\mathbb{1}_{\{\tilde{Y}(s)>0\}}}{\tilde{Y}(s)} \mathrm{d}\mathscr{M}(s), \\ &= \int_0^t \frac{\mathbb{1}_{\{\tilde{Y}(s)>0\}}}{\tilde{Y}(s)} [\mathrm{d}\tilde{N}(s) - \mathrm{d}\Lambda(s)] \\ &= \int_0^t \frac{\mathrm{d}\tilde{N}(s)}{\tilde{Y}(s)} - \int_0^t \mathbb{1}_{\{\tilde{Y}(s)>0\}} \mathrm{d}\Lambda(s), \end{aligned}$$

is a martingale. Since $\mathscr{M}(0) = 0$,

$$\mathbb{E}\int_0^t \frac{\mathrm{d}\tilde{N}(s)}{\tilde{Y}(s)} = \mathbb{E}\int_0^t \mathbb{1}_{\{\tilde{Y}(s)>0\}}\mathrm{d}\Lambda(s).$$

□

2. Kaplan-Meier Estimator. This estimator is a nonparametric estimator for estimating the survival function.

Proposition 10

Let $\hat{\Lambda}(t) = \int_0^t \frac{\mathrm{d}\tilde{N}(u)}{\tilde{Y}(u)}$ be the Nelson-Aalen estimator. The Kaplan-Meier estimator of the survivor function for counting processes is given as:

$$\hat{S}(t) = \prod_{s\leq t}\Big[1-\hat{\Lambda}(t)\Big]. \tag{3.73}$$

Proof. (Fleming & Harrington, 2011). Recall that $S(t) = 1 - F(t)$ and $\Lambda(t) = \int_0^t \{1-F(u-)\}^{-1}\mathrm{d}F(u)$. Therefore,

$$\mathrm{d}\Lambda(u) = \frac{\mathrm{d}F(u)}{1-F(u-)}.$$

Since $F(0) = 0$,

$$F(t) = \int_0^t \mathrm{d}F(u) = \int_0^t \{1-F(u-)\}\mathrm{d}\Lambda(u). \tag{3.74}$$

Using the definition $\hat{S}(u) = 1 - \hat{F}(u)$ and inserting the Nelson-Aalen estimator given in Equation 3.72 into Equation 3.74 gives:

$$\hat{S}(t) = \int_0^t \hat{S}(u-)\mathrm{d}\hat{\Lambda}(u)$$

$$\mathrm{d}\hat{S}(t) = \hat{S}(u-)\mathrm{d}\hat{\Lambda}(u).$$

$$\text{Hence } \Delta\hat{S}(t) = \hat{S}(u-)\Delta\hat{\Lambda}(s).$$

But $\hat{S}(t-) - \hat{S}(t) = -\Delta\hat{S}(t)$. This implies that

$$\begin{aligned}\Delta\hat{S}(t) =& \hat{S}(t-)\frac{\Delta\tilde{N}(t)}{\tilde{Y}(t)}\\ \text{Then } \hat{S}(t) =& \hat{S}(t-)\left[1 - \frac{\Delta\tilde{N}(t)}{\tilde{Y}(t)}\right],\\ \text{and } \hat{S}(t) =& \prod_{s\leq t}\left[1 - \frac{\Delta\tilde{N}(s)}{\tilde{Y}(s)}\right].\end{aligned}$$

□

The counting process approach can also be applied to the Cox proportional hazards regression model. The details are discussed in the next section.

Application of Counting Processes to Cox Regression

In this subsection, the counting process approach of the Cox proportional hazards model is discussed. In the traditional approach, survival processes in the Cox model are represented by a function of three lifetime indicators: t_i (the survival time), δ_i (the censoring indicator) and x_i (independent variables). However, a survival process in the counting process approach is denoted by the triple $\{N_i(t), Y_i(t), X_i\}$ of counting paths. From Equations 3.57 and 3.58, let

$$N_i(t) = \mathbb{1}_{\{\tilde{T}_i \leq t, \delta_i = 1\}}, \text{ and } Y_i(t) = \mathbb{1}_{\{\tilde{T}_i \geq t\}}. \tag{3.75}$$

The vector of independent variables X_i is given by $X_i = X_1, \cdots, X_n$. It is usually specified as time dependent, $X_i(t)$, in the counting process model.

Let $\mathscr{F}_t$ be a sub-σ-algebra of the σ-algebra $\mathscr{F}$ which is continuous when $\mathscr{F}_{t+} = \mathscr{F}_t$. The σ-algebra generated by the triple $\{N_i(t), Y_i(t), X_i(t)\}$ is given by the filtration:

$$\mathscr{F}_t = \sigma\{N_i(s), Y_i(s+), X_i(s),\ i = 1, \cdots, n;\ 0 \leq s \leq t\}. \tag{3.76}$$

The values of $\tilde{T}_i$ and δ_i such that $\tilde{T}_i < t$ generates a filtration on $[0,t)$ given as:

$$\mathscr{F}_{t-} = \sigma\{N(t), Y(t), X(t)\}. \tag{3.77}$$

To obtain the discrete realization of the σ-algebra $\mathscr{F}_{t-}$ for individual i given that the value $Y_i(t)$ is 0 or 1, Equation 3.59 can be written as:

$$\begin{aligned}\mathbb{E}[\Delta N_i(t) \mid \mathscr{F}_{t-}] =& \mathbb{E}[\Delta N_i(t) \mid Y_i(t)] \\ =& P\{t \leq T_i < t + \Delta t, C_i \geq t \mid Y_i(t)\} \\ =& Y_i(t)\lambda(t)\mathrm{d}t. \end{aligned} \tag{3.78}$$

The last line in Equation 3.78 is an expected value and is called the compensator of $N_i(t)$ with respect to the filtration $\mathscr{F}_t$.

The integrated intensity process for the i-th individual is given as:

$$\Lambda_i(t) = \int_0^t Y_i(u)\lambda_i(u)\mathrm{d}u. \tag{3.79}$$

The intensity function is usually written in terms of the differential of the integrated intensity process as:

$$\lambda_i(t)\mathrm{d}t = \mathrm{d}\Lambda_i(t). \tag{3.80}$$

The intensity function for the ith individual at time t is given by the counting process as:

$$\lambda_i(t) = \lambda_0(t)\exp[X_i\beta], \tag{3.81}$$

where β is an $(L \times 1)$ vector of regression coefficients. Therefore, it follows from Equations 3.81 and 3.79 that the intensity function for $N_i(t)$ with the vector of

independent variables X_i is written as:

$$Y_i(t)\mathrm{d}\Lambda(t,X_i) = Y_i(t)\exp[X_i(\beta)]\mathrm{d}\Lambda_0(t). \tag{3.82}$$

The partial likelihood for n independent triples $[N_i, Y_i, X_i]$, where $i = 1, \cdots, n$ for the Cox proportional hazards regression model on the time interval $[0, s]$ in counting processes is given as:

$$L(\beta) = \prod_{i=1}^{n}\prod_{t\geq 0}\left\{\frac{Y_i(t)\exp[X_i^T\beta]}{\sum_k Y_k(t)\exp[X_k^T\beta]}\right\}^{\mathrm{d}N_i(t)}, \tag{3.83}$$

and the log partial likelihood function is given as:

$$\log L(\beta) = \sum_{i=1}^{n}\int_0^{\infty} Y_i(t)[X_i^T\beta] - \log\left\{\sum_k Y_k(t)\exp[X_k^T\beta]\right\}\mathrm{d}N_i(t). \tag{3.84}$$

The estimator $\hat{\beta}$ in the log partial likelihood function is the solution to the equation:

$$U(\beta) = \frac{\partial}{\partial\beta}\log L(\beta)\mid_{\beta=\hat{\beta}} = 0. \tag{3.85}$$

Proposition 11

The first derivative of the log partial likelihood function:

$$U(\beta) = \frac{\partial}{\partial\beta}\log L(\beta), \tag{3.86}$$

is a martingale.

Proof. Gill (1984). Let $L(\beta)$ be the likelihood function for β and

$$\overline{X}(\beta) = \sum_{k=1}^{n}\frac{Y_k(t)X_k\exp[X_k^T\beta_0]}{\sum_{k=1}^{n}Y_k(t)\exp[X_k^T\beta_0]}.$$

Then from Equation 3.83 we have that:

$$\frac{\partial}{\partial\beta}\log(\beta_0) = \sum_{k=1}^{n}\sum_{t\leq s}(X_i - \sum_{k=1}^{n}\frac{Y_k(t)X_k\exp[X_k^T\beta_0]}{\sum_{k=1}^{n}Y_k(t)\exp[X_k^T\beta_0]})$$
$$= \sum_{k=1}^{n}\int_{t=0}^{s}(X_i - \overline{X})\mathrm{d}N_i(t).$$

Since $\mathrm{d}M_i(t) = \mathrm{d}N_i(t) - \Lambda_i(t)\mathrm{d}(t)$, and

$$\sum_{k=1}^{n}\int_{t=0}^{s}(X_i - \overline{X})\mathrm{d}N_i(t) = \sum_{k=1}^{n}\int_{t=0}^{s}(X_i - \overline{X})\Lambda_i(t).$$

We have that

$$\sum_{k=1}^{n}\int_{t=0}^{s}(X_i - \overline{X})\Lambda_i(t) = \sum_{k=1}^{n}X_iY_i(t)\lambda_0(t)\exp(X_k^T\beta_0) - \overline{X}_i(\sum_{k=1}^{n}Y_i(t)\lambda_0(t)\exp(X_k^T\beta_0))$$
$$=0.$$

Hence,

$$\frac{\partial}{\partial\beta}\log(\beta_0) = \sum_{k=1}^{n}\int_{t=0}^{s}(X_i - \overline{X})\mathrm{d}M_i(t).$$

□

In the next section, the residuals for the Cox regression model are discussed.

Residuals for the Cox Regression Model

The residuals for the Cox regression model include Cox-Snell, Schoenfeld, martingale, deviance and partial or score residuals. Liu (2012) and Harrell Jr (2015) discuss residuals for the Cox regression model. Some of these residuals are discussed below.

Schoenfeld Residual

Definition 52

The Schoenfeld residual is the difference between the covariate vector for the

individual at event time t_j and the expectation of the covariate over the risk set $R(t_j)$. It is defined as:

$$r^{Schoenfeld} = x_j - \mathbb{E}[x_j \mid R(t_j)]. \tag{3.87}$$

Schoenfeld residuals are used to test the proportional hazards assumption.

Cox-Snell Residual

Definition 53

Suppose a Cox proportional hazards model, $h(t,X) = h_0(t)\exp(X_k^T\beta)$, has been fitted to the data (t_k, δ_k, X_k), for $k = 1, \cdots, p$. The Cox-Snell residuals are given as:

$$r_k^{Cox} = \hat{H}_0(t_k)\exp\Big(\sum_{k=1}^{p} X_k\beta_k\Big). \tag{3.88}$$

The Cox-Snell residual can be used to assess the fit of the Cox proportional hazards model.

Remark 22

If the model is correct and the estimated $\hat{\beta}$s are close to the true values of β, then the Cox-Snell residuals should look like a censored sample from a unit exponential distribution. Determining whether the Cox-Snell residuals behave like a sample from an exponential distribution involves computing the Nelson-Aalen estimator of the hazard function of the Cox-Snell residuals. If the unit exponential distribution fits the data, then this estimator should be approximately equal to the cumulative hazard function of the unit exponential.

Martingale Residual

Definition 54

It is an estimate of the excess events in the data that are not predicted by the model. The martingale residuals for the Cox model without time dependent

covariates is given as:

$$\hat{M_l}(t) = \delta_l - \hat{\Lambda}_0(t_l)\exp(X_l^T\hat{\beta}), \tag{3.89}$$

where $\hat{\Lambda}_0(t_l)$ is the estimated baseline cumulative hazard function at the observed survival time t_l.

Remark 23

From Definition 54:

1. Martingale residuals are based on the martingale theory.
2. Martingale residuals are used to assess the best functional form for a given covariate using an assumed Cox model for the remaining covariates.

Regression models are evaluated after they have been fitted to determine their efficiency. In the section that follows, the steps involved in evaluating a fitted regression model are discussed.

Evaluating Regression Models

Parameters estimated in regression models are evaluated to determine how good the models fit the data. There are several steps involved in evaluating regression models. They include assessing the overall model performance, the significance of each independent variable, goodness of fit statistics, and validating the predicted probabilities of the model.

Step 1: Assessing the significance of independent variables. The significance of each independent variable in a regression model can be tested using the Wald statistic and likelihood ratio test.

Definition 55

The *Wald Statistic* is the ratio of the square of the regression coefficient to the square of the standard error of the coefficient. It is given as:

$$W = \left(\frac{\hat{\beta}}{S.E_{\beta}}\right)^2. \tag{3.90}$$

Remark 24

From Definition 55:

(a) The Wald statistic is asymptotically distributed as a χ^2 distribution.

(b) It is also compared with a χ^2 distribution with one degree of freedom.

(c) The reliability of the Wald statistic is sometimes questionable even though it is easy to compute.

Definition 56

The *likelihood ratio test* for an independent variable compares the likelihood of obtaining the data when the value of the variable is zero (L_0) to the likelihood of obtaining the data when the value of the variable is evaluated at the maximum likelihood estimate (L_1). The test statistic is given as:

$$G = -2(\ln L_0 - \ln L_1). \tag{3.91}$$

Remark 25 (a) The statistic is computed with a χ^2 distribution with one degree of freedom.

(b) The contribution of each independent variable in a model is assessed by entering the variables hierarchically and comparing each new model with the previous model to determine the significance of the variable.

Step 2. Assessing the overall model performance. The overall fit of a model describes the strong relationship between all the independent variables in a model, when taken together, and the dependent variable. This can be done using the likelihood ratio test.

Definition 57

The *likelihood ratio test for a model* compares the fit of the model with

and without the independent variables. The test statistic is given as:

$$G = -2(\ln L_0 - \ln L_1), \tag{3.92}$$

where L_0 is the likelihood of the model with no independent variables (null model) and L_1 is the likelihood of the model with independent variables.

Remark 26

If the p-value is less than 0.05, it implies that at least one of the independent variables contributes to the prediction of the outcome.

Step 4. Goodness of fit statistics. These statistics assess the fit of the logistic regression model against actual outcomes and determines whether the variation in the residuals of the model is small. The goodness of fit statistic for logistic regression is Le Cessie van Houwelingen Copas-Hosmer sum of squares test. The Nagelkerke R^2, C-statistic, and the Chi-square goodness of fit tests can be used for both logistic regression and survival analysis models.

Definition 58

The *Nagelkerke* R^2 *index* is used to quantify the predictive strength of the model. It is defined as:

$$R_N^2 = \frac{1 - \exp(\frac{-LR}{n})}{1 - \exp(\frac{-L_0}{n})}, \tag{3.93}$$

where LR is the global log likelihood ratio statistic for testing the importance of all the predictors in the model and L_0 is the log likelihood for the null model and n is the sample size.

Definition 59

Chi-square tests involve residuals, $(y_i - \hat{y}_i)$, where y_i is the observed dependent variable for the ith subject and $\hat{y}_i$ is the corresponding prediction from the model.

Remark 27 (a) A standardized residual is defined as:

$$r_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i(1-\hat{y}_i)}}, \tag{3.94}$$

where the standard deviation of the residuals is $\hat{y}_i(1-\hat{y}_i)$.

(b) A χ^2 statistic can be formed from the standardized residual as: $\chi^2 = \sum_{i=1}^{n} {r_i}^2$.

(c) This statistic follows a χ^2 distribution with $n-(k+1)$ degrees of freedom, and p-values can be calculated.

Definition 60

The *le Cessie van Houwelingen Copas- Hosmer sum of squares* test statistic is given as:

$$\hat{T}_r = \sum_{i=1}^{n} \frac{\hat{r}_{si}^2}{\hat{var}(\hat{r}_{si}^2)}, \tag{3.95}$$

where $\hat{r}_{si}^2$ is the smoothed standardized residuals and $\hat{var}(\hat{r}_{si}^2)$ is the variance of the smoothed standardized residuals.

Definition 61

C-statistic is the proportion of pairs with different observed probabilities for which the model predicts correctly a higher probability for observed events than for unobserved events. The C statistic is given as:

$$C = \sum_{\substack{l=1 \\ Y_l=0}}^{r} \sum_{\substack{k=1 \\ Y_k=1}}^{r} \frac{I(P_k > P_l) + \frac{1}{2} I(P_k = P_l)}{N_o N_1}, \tag{3.96}$$

where P_d denotes an estimate of $P(Y_d = 1 \mid X_d)$ and N_0, N_1 the sample sizes of probabilities $Y = 0$ and $Y = 1$ respectively.

Step 4. Validating predicted probabilities. A fitted model can be validated by using the model with its estimated coefficient to predict the dependent vari-

able of another dataset and checking its residuals. Model validation can either be external or internal.

Definition 62

External validation involves testing a final model developed in one setting on subjects in another setting at another time.

Definition 63

Internal validation involves fitting and validating the model carefully by using one series of subjects.

Remark 28 (a) Internal validation techniques involve data splitting, cross validation and bootstrapping.

(b) Data splitting involves a random split of the dataset into two, using one to fit the model and the other to validate the model.

(c) Cross validation is repeated data splitting with alternating development and validation samples.

(d) Bootstrapping involves sampling with replacement repeatedly from the observed dataset and forming a large number of bootstrap datasets (samples), each of the same size as the original data.

In model validation, a measure of association is used to express the degree to which predicted probabilities agree with actual outcomes. There are several measures of association which include Somer's D statistic, the *C*-statistic, Nagelkerke R^2, among others.

Definition 64

Somer's D is a transformed version of the c- statistic. It is given as:

$$D_{xy} = 2(C - 0.5). \tag{3.97}$$

The steps discussed above provide efficient means to evaluate regression models.

In regression models, the presence of a nonlinear relationship between a dependent variable Y and an independent variable X means that including X as a single independent variable will not be sufficient to model the relationship. Therefore, regression splines are used to model such relationships. The concept of regression splines are discussed in the next section.

Regression Splines

In this section, we discuss the concept of regression splines. We give more attention to cubic spline functions. Regression splines are discussed in Korn & Graubard (2011); Smith (1979); Wegman & Wright (1983), and Wold (1974).

Definition 65

Splines that have been computed with respect to a regression model are known as *regression splines.*

Remark 29 1. Regression splines are used in regression models that specify the expected dependent variable as a function of the linear combination of the independent variables. For example logistic regression, Cox proportional hazards regression, among others.

2. Regression splines have great flexibility for analysing data and can be used as graphical tools for communicating complex findings naturally.

Recall the definition of a spline function of a variable x given in Definition 28. A family of splines is defined by the degree m of the spline function, the number of knots P, and the position of the knots ($k_1 < \cdots < k_P$). In order to use a spline function of a variable x in a multivariable regression model, we have to introduce $P+m$ new variables. That is, x^a for $a = 1, \cdots, m$ and $(x-k_i)_+^m$ for $i = 1, \cdots, P$. Then the regression coefficients β_{ia} are determined in addition with the regression coefficients for the other variables.

Definition 66

Cubic splines is defined as:

i. a cubic function between adjacent elements of a set of fixed knots $k_1 < \cdots < k_P$ in the range of x and

ii. continuous and has continuous first and second order derivatives.

Mathematically, it is given as:

$$s(x) = \sum_{a=0}^{3} \beta_{0a} x^a + \sum_{i=1}^{P} \beta_{i3} (x - k_i)_+^3. \tag{3.98}$$

Remark 30

Cubic splines are usually used because they are visually smooth and have greater flexibility in fitting data. However, they have abnormal behaviour beyond the outermost knots. This drawback results in a restriction of these splines to be linear beyond the outermost knots, hence the notion of restricted or natural cubic splines.

Definition 67

A *restricted cubic spline* is defined as:

i. a cubic function between adjacent elements of a set of fixed knots $k_1 < \cdots < k_P$ in the range of x.

ii. a linear function for $x < k_1$ and $x > k_P$, and

iii. continuous and has continuous first and second order derivatives.

Mathematically, it is given as:

$$\begin{aligned} s(x) =& \beta_{00} + \beta_{01} x + \sum_{i=1}^{P-2} \beta_{i3} \Big[(x - k_i)_+^3 - (x - k_{P-1})_+^3 \Big(\frac{k_P - k_i}{k_P - k_{P-1}} \Big) \\ &+ (x - k_P)_+^3 \Big(\frac{k_{P-1} - k_i}{k_P - k_{P-1}} \Big) \Big]. \end{aligned} \tag{3.99}$$

Remark 31

In the restricted cubic spline,

a. for $x < k_1$, it implies that $\beta_{02} = \beta_{03} = 0$.

b. for $x > k_P$, it implies that $\sum_{i=0}^{P} \beta_{i3} = 0$ and $\sum_{i=1}^{P} \beta_{i3} k_i = 0$.

After specifying the method to be used to transform nonlinear relationships in the regression model, we discuss how the model should be built.

Model Building Process

In this section, we discuss the processes involved in building the models. The main aim of model building is to select as less variables as possible but, the model, still reflects the true outcomes of the data . There are a lot of model building strategies which include stepwise regression, best subsets and purposeful selection of variables. (Hosmer Jr et al., 2013, Chapter 4) and Zhang (2016) discussed the method of purposeful selection of covariates for building statistical models. In this study, we employed the method of purposeful selection of covariates in building the models. The steps involved in the purposeful selection of variables are discussed below.

Step 1. Univariable Analysis.

The first step in purposeful selection of variables involves univariable analysis to examine the unadjusted association between the independent and dependent variables. Each of the independent variables is used to fit the model one at a time.

Step 2. Multivariable Analysis.

First of all, a multivariable model is fitted with the significant variables identified in the univariable stage. The model is assessed for significant variables. Then, a new multivariable model is fitted with the variables that were not significant in the univariable analysis and the significant variables determined in the initial multivariable analysis. The significant variables for this new model are determined. The model at the end of this step is called the preliminary main effects model.

Step 3. Analysis of the functional form of the variables in the model.

In this step, we examine whether the variables in the preliminary main effects model obey the linearity assumption or not. This is usually done for the continuous variables. This can be done using a significance test or residual plots. The significance test is done using polynomials, transformations, fractional polynomials or restricted cubic splines. Partial residual plots are used in the logistic regression model while martingale residual plots are used in the survival analysis model. The model at the end of this step is called the main effects model.

Step 4. Test for Interactions.

In this step, we check for interactions among the variables in the model. The interaction variables are created as the product of the main effect variables. The existence of interaction between two variables implies that the effect of one variable on the dependent variable is not constant over the levels of the other variable. The significance of the interaction variables are then assessed.

Step 5. Model diagnostics.

In this step, we assess the adequacy and fit of the model using residuals and goodness of fit tests described earlier for the models.

Chapter Summary

The models used in predicting the probability of default were discussed. First of all, the logistic regression model, its assumptions, how to estimate its parameters and its residuals were discussed. Also, survival analysis, its assumptions, its parameter estimation methods and residuals were discussed. The counting process approach to survival analysis was also considered. After that, the steps involved in evaluating regression models were presented. Finally, the method involved in transforming continuous covariates in the presence of non-linear relationships were presented.

CHAPTER FOUR

RESULTS AND DISCUSSIONS

Introduction

In this chapter the results obtained from fitting logistic regression and survival analysis models applied to a dataset from a microfinance institution for predicting the probability of default are discussed. A brief description of the dataset used for the analysis is discussed. The model building process for both logistic regression and survival analysis models are discussed in detail. Also, the results for evaluating model assumptions are discussed. A mathematical expression for both logistic regression and survival analysis models for predicting the probability of default are given.

Data Description

The data used for the study was obtained from a microfinance institution in Ghana. The data in its raw form could not be used for the analysis, therefore it required some cleaning. It contained details of customers who took individual and group loans. All entries of the data which contained missing information were not considered.

The data consisted of loans disbursed from January 2012 to January 2019. However, the main focus of the study was to estimate the probability of default of individual loans so all the group loans were ignored. Also, it contained information on different types of repayment times such as monthly, weekly, and fortnightly but the study concentrated on only customers who paid monthly. The total sample used for the study was 1060 customers. These included customers who had finished paying their loans and those who had not finished paying their loans. Those who had not finished paying their loans were considered as the censored observations.

A loan was considered to be in default when the customer does not pay his or her loan to the microfinance in full.The events were coded as 0 for non-default and 1 for default. The variables in the data that were considered as independent

variables and their description are discussed below. First of all, the variables from the application data, often called application characteristics, which basically contain information about the customers and the loan characteristics are discussed. The variables are described in Table 1 below.

Table 1: Description of Variables used in the Model

Variable	**Description**
Branch Name	Name of the branch where the loan was taken.
Mobile Number	The mobile number of a customer.
Principal	The total amount of money taken as a loan.
Age	The age of the customer.
Gender	The gender of the customer.
Rate	The interest rate on the loan.
Product	The type of loan taken.
Marital Status	The marital status of the client.
Number of Repayment	The length of time taken to repay the loan.
Date Disbursed	The date the loan was given out.
Repayment Start Date	The date to start repayment.
Repayment End Date	The date to end repayment.

Source: Working Data (2019)

Table 1 shows the application characteristics of customers who took loans from the microfinance institution. The variables Date Disbursed, Repayment Start Date and Repayment End Date were used to determine the default status. It must be noted that some of the characteristics had different levels while others were used to derive a new variable. For instance, mobile network was derived from mobile number. The variables were considered as continuous variables while others were considered as categorical variables based on the nature of the inputs they contained. The continuous variables were the variables that contained numeric inputs, such as Principal, Age, Number of Repayment and Rate.

The categorical variables were the variables that consisted of different levels, such as Branch Name, Gender, Marital Status, Product, and Mobile Network. The different levels of the categorical variables are given in Table 2 below.

Table 2: Description of Categorical Variables

Variable	**Level**
Branch Name	Adabraka Branch
	Ashaiman Branch
	Bolgatanga Branch
	Pokuase Branch
	Tamale Branch
Gender	Female
	Male
Marital Status	Divorced
	Married
	Not Specified
	Single
Product	Edwumapa Loan
	Personal Loan
	Salary Loan
	Staff Loan
	Others
Mobile Network	Airtel Tigo
	MTN
	Voda
	Others

Source: Working Data (2019)

Table 2 shows the description of the various levels of each of the categorical variables from the application characteristics.

Apart from these application characteristics, some other derived variables were used in building the models. They include macroeconomic variables and variables denoting some specific periods in Ghana. The macroeconomic variables used include the average inflation rate, average unemployment rate, average annual GDP growth rate, Bank of Ghana lending rate and the average foreign exchange rate. The variables denoting some specific periods in Ghana include some festive occasions such as New Year, Easter, Ramadan and Christmas, and beginning of an academic year. The macroeconomic variables were considered as continuous variables while the variables denoting some specific periods in Ghana were considered as categorical variables, coded as 1 for if the repayment period lies within such periods and 0 otherwise.

In the next section, the descriptive statistics of the application characteristics are discussed.

Descriptive Statistics

In this section, the results obtained from the descriptive statistics of the continuous application variables are discussed. The descriptive statistics of Age, Rate, Principal and Number of Repayment are shown in Table 3 below.

Table 3: Descriptive Statistics for Continuous Application Variables

Variable	Minimum	Maximum	Median	Mean
Age	0.00	70.00	40.00	40.95
Principal	300	250000	3000	4425
Rate	0.00	1.80	0.60	0.58
Number of Repayment	1.00	24.00	6.00	6.22

Source: Working Data (2019)

Table 3 shows the descriptive statistics of the continuous variables of the application characteristics of the customers of the microfinance institution. The mean age of the customers of the microfinance institution who take individual loans is approximately 41 years and the average rate is 0.58. Also, the average

time taken for a loan to be repaid is approximately 6 months and the average amount of money given out as loans is 4425.

In the next section, the logistic model for predicting probability of default is discussed.

Model Building for Logistic Regression

In this study, the method of purposeful selection of covariates is used to build the logistic regression model for predicting the probability of default. This method has been discussed under the model building process on page 71. The results obtained from building the logistic regression model are discussed below.

A total sample of 863 customers were used for the logistic regression model. This is due to the fact that the model considers customers whose loan term has matured, that is those whose time for repayment is complete. Those who finish paying their loan at the end of their repayment time are classified as non-defaulters while those who do not finish paying at the end of their repayment time are classified as defaulters. The application and macroeconomic variables were used in building the logistic regression model. The variables that are linked to the social environment of Ghana were not included in this model because they were not significant.

Let $X_i,\ i = 1, \cdots, n$ denote the variables used in the logistic regression model. The meaning of the X_i's are shown in Table 4 below.

Table 4: Variables for Logistic Regression

Variable	Meaning
X_1	Age
X_2	Rate
X_3	Gender
X_4	Branch Name
X_5	Marital Status
X_6	Number of Repayment
X_7	Mobile Network
X_8	Principal
X_9	Product
X_{10}	Average Unemployment Rate
X_{11}	Average Annual GDP Growth Rate
X_{12}	Average Inflation Rate
X_{13}	Bank of Ghana Lending Rate
X_{14}	Average Foreign Exchange Rate

Source: Working Data (2019)

Table 4 shows the independent variables that were used to fit the logistic regression model. The results obtained from following the steps in the method of purposeful selection of covariates are discussed below.

1. The first step in the method of purposeful selection of covariates is univariable analysis. The following results were obtained from the univariable analysis.

   Proposition 12

   The covariates that were significant in the univariable analysis with their corresponding p-values are shown in Table 5 below.

Table 5: Significant Covariates from Univariable Logistic Regression Analysis.

Variable	***p*-value**
Age	0.0143
Rate	< 0.0001
Branch Name	< 0.0001
Marital Status	< 0.0001
Number of Repayment	0.0013
Network	< 0.0001
Average Annual GDP Growth Rate	0.0217
Average Inflation Rate	< 0.0001
Bank of Ghana Lending Rate	< 0.0001
Average Foreign Exchange Rate	< 0.0001

Source: Working Data (2019)

Table 5 shows the results for the significant covariates in the logistic regression model. The p-value was set at an alpha level of 0.05.

Proof. These results follow from using the likelihood ratio test to determine the significance of the independent variables which was discussed on page 64. □

2. The next step is mulitvariable analysis. The following results were given by the multivariable model.

Proposition 13

The significant variables from the fitted multivariable model with their respective p-values are shown in Table 6 as follows.

Table 6: Significant Covariates from Multivariable Logistic Regression Analysis.

Variable	***p*-value**
Rate	0.0004
Branch Name	0.0004
Number of Repayment	0.0088
Average Inflation Rate	0.0002
Average Foreign Exchange Rate	0.0006

Source: Working Data (2019)

Table 6 shows the significant variables in the fitted multivariable model at an alpha level of 0.05.

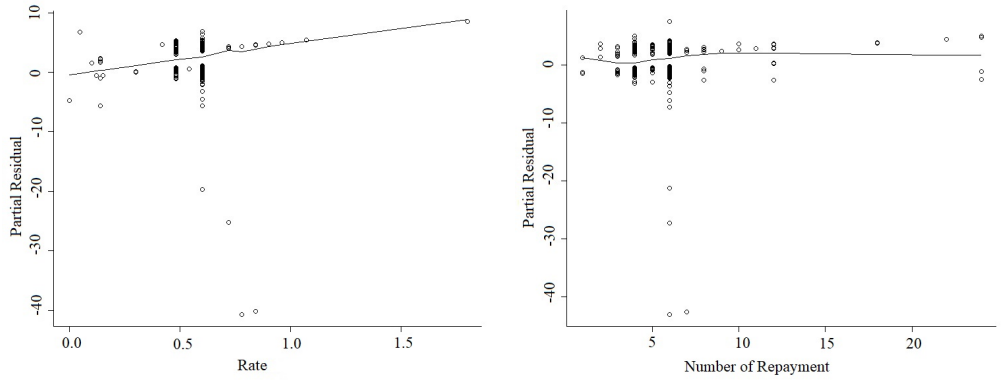
Proof. This result is based on the steps involved in fitting multivariable models discussed under model building process on page 71. □

3. The next step is to assess the functional form of the continuous variables in the fitted multivariable model. This helps to determine whether the linearity assumption holds or not. The continuous variables from the significant variables in the multivariable model shown in Table 6 are Rate, Number of Repayment, Average Inflation Rate and Average Foreign Exchange Rate. The functional form of the continuous variable can be assessed using partial residuals and linearity test.

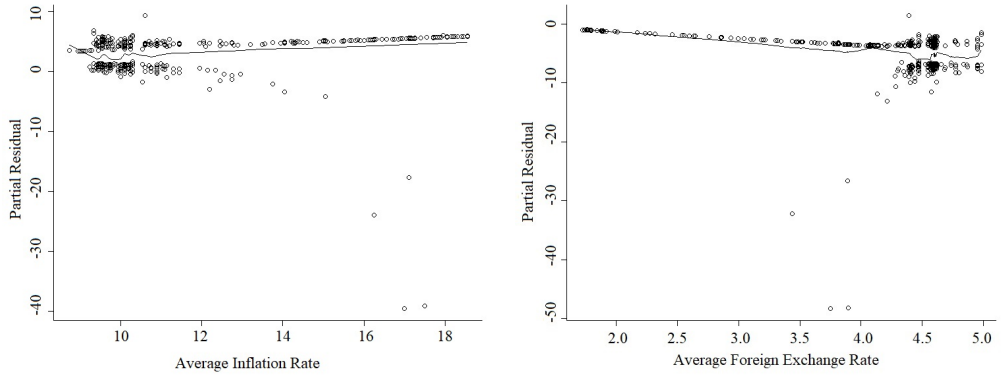
   Proposition 14

   The results from the assessment of the functional form of the continuous covariates are shown below.

   (a) The partial residual plots for the continuous covariates are shown in Figure 1 below.



(a) Rate (b) Number of Repayment



(c) Average Inflation Rate (d) Average Foreign Exchange Rate

Figure 1: Partial Residual Plots for Continuous Covariates

Figure 1 shows the partial residual plots for the continuous covariates in the logistic regression model. Figure 1(a) shows the partial residual plot for Rate, Figure 1(b) for Number of Repayment, Figure 1(c) for Average Inflation Rate and Figure 1(d) for Average Foreign Exchange Rate. Figure 1(a) seems to be increasing, while Figure 1(b) decreases slightly at the initial stages between 1 and 8 and remains constant. Figures 1(c) and 1(d) tend to decrease slowly. These plots are not sufficient to conclude that the continuous covariates affects the log odds in a nonlinear relationship. Hence a test of significance is conducted to confirm the results from the partial residual plots.

(b) The results from the linearity test, that is the test to determine which continuous variables obey the linearity assumption in the multivari-

able model are shown in Table 7 below.

Table 7: Results from Linearity Test for Logistic Regression

Variable	***p*-value**
Rate	0.1251
Nonlinear	0.9931
BranchName	0.0014
Number of Repayment	0.0007
Nonlinear	0.0013
Average Inflation Rate	0.6515
Nonlinear	0.8803
Average Foreign Exchange Rate	0.1914
Nonlinear	0.2279

Source: Working Data (2019)

The results in Table 7 show that the variable Number of Repayment does not obey the linearity assumption since the p-value is significant at alpha level of 0.05. Therefore, a restricted cubic spline is used to transform it. From Table 4, X_6 = Number of Repayment. The restricted cubic spline transformation for X_6, with k_i, $i = 1, \cdots, P$ knots is given as:

$$s(X_6) = \beta_{00} + \beta_{01}X_6 + \sum_{i=1}^{P-2} \beta_{i3}\Big[(X_6 - k_i)_+^3 - (X_6 - t_{P-1})_+^3\Big(\frac{t_P - k_i}{k_P - k_{P-1}}\Big) + (X_6 - k_P)_+^3\Big(\frac{k_{P-1} - k_i}{k_P - k_{P-1}}\Big)\Big] \tag{4.1}$$

Proof. The results above follow from partial residuals discussed under regression residuals for logistic regression on page 36 and restricted cubic splines under regression splines on page 69. □

4. After checking the linearity assumption and making the necessary transformations for the continuous covariates, a test for interactions is con-

ducted. In the fitted model, none of the interaction terms were significant.

Based on the results obtained above, a theorem is deduced to provide the mathematical model for predicting the probability of default of clients who take individual loans from the microfinance institution.

Theorem 9

The logistic regression model for predicting the probability of default of a customer who takes individual loans from the microfinance institution is given by:

$$P(\text{Default}) = \frac{1}{1+\exp(-f(\text{X}))}, \tag{4.2}$$

where $f(X) = \beta_0 + \beta_2 \text{X}_2 + \beta_4 \text{X}_4 + s(\text{X}_6) + \beta_{12}\text{X}_{12} + \beta_{14}\text{X}_{14}$, with $s(\text{X}_6)$ given in Equation 4.1 and the X_i's defined in Table 4.

5. After obtaining a mathematical model for predicting the probability of default, the model needs to be assessed to determine its efficiency. The results are shown below.

Proposition 15

The measures for assessing the performance of the overall model in this study are the Nagelkerke R^2 and le Cessie van Houwelingen Copas-Hosmer sum of squares test. The Nagelkerke R^2 value for the fitted model is 0.336 which implies that the fitted model performs better than the null model by approximately 34%.

The results for le Cessie van Houwelingen Copas-Hosmer sum of squares test is given in Table 8 as follows.

Table 8: Results for le Cessie van Houwelingen Copas-Hosmer sum of Squares Test

Residuals	
Sum of squared errors	157.5399
Expected value | H_0	157.2232
Standard Deviation	0.9364
Z-score	0.3382
p-value	0.7352

Source: Working Data (2019)

Table 8 shows the results of le Cessie van Houwelingen Copas-Hosmer sun of squares test with a p-value of 0.7352 which is not significant at an alpha level of 0.05. This means that the model is a good model.

The model statistics for the logistic regression model are shown in Table 9 below.

Table 9: Model Statistics for Logistic Regression

Model Likelihood Ratio Test		**Discrimination Indexes**	
LR χ^2	245.87	R^2	0.336
d.f.	9	C	0.777
Pr($>\chi^2$)	<0.0001	D_{xy}	0.554

Source: Working Data (2019)

Table 9 shows the model statistics for the logistic regression model. The model likelihood ratio test (LR χ^2) has a p-value (Pr($>\chi^2$)) of < 0.0001 which indicates a significant overall model at an alpha level of 0.05. The R^2 index is 0.336 showing that the model performs better than the model without covariates by approximately 34%. The C-statistic for the logistic regression model is 0.777 indicating that the probability that the prediction

is higher for a default than a non-default by the fitted model is approximately 78%.

Proof. The results follow from the statistics used for assessing the overall model performance discussed on page 71. □

The fitted model is then validated and calibrated to determine its performance on future subjects. The results from the validated logistic regression model are shown in Table 10 below.

Table 10: Results from Validated Logistic Regression Model.

Index	**Original Sample**	**Training Sample**	**Test Sample**	**Optimism**	**Corrected Index**	*n*
D_{xy}	0.5537	0.5666	0.5450	0.0217	0.5320	200
R^2	0.3357	0.3468	0.3270	0.0198	0.3160	200

Source: Working Data (2019)

In Table 10, the first column, Index, shows the different discrimination indexes that are being validated. The second column, Original Sample, gives the value of the measure calculated using the model fitted to the original data and evaluated on the original data. The third column, Training Sample, gives the mean of the index on the bootstrap samples. The fourth column, Test Sample, gives the mean of the index by applying the models fitted to the bootstrap datasets evaluated on the original data. The fifth column, Optimism, is the difference between Training and Test Samples, which is the decrease in the performance between the Training and Test Samples. The sixth column, Corrected Index, is the difference between Original Sample and Optimism. The last column, n, shows the total number of bootstrap samples. The value of the Optimism helps the fitted model to adjust for over-fitting. The Optimism Corrected index of the D_{xy} of the fitted model is 0.5320, which indicates that the fitted model has a good discrimination. The second row also shows the validated results for the

R^2. The R^2 index for the Original Sample is 0.3357, while that of the Corrected Index is 0.3160.

The results obtained from calibrating the logistic regression model are shown in Figure 2 below.

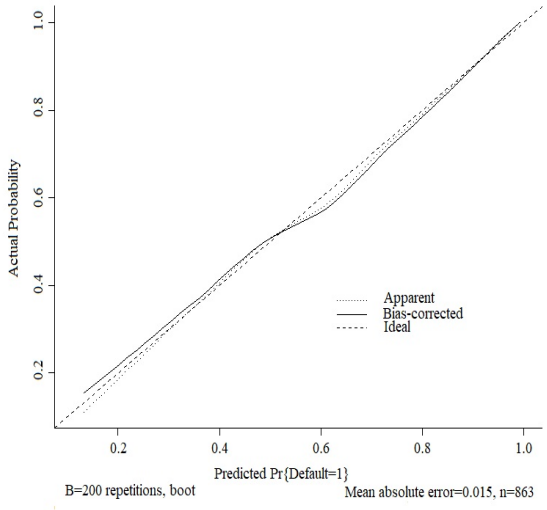


Figure 2: Plot of Calibration of Fitted Model.

Figure 2 shows a plot of the results obtained from calibrating the fitted model. The Ideal line represents perfect prediction as the predicted probabilities equal the observed probabilities. The apparent line is the in-sample calibration. The bias-corrected line is derived through the resampling procedure which helps to add uncertainty to the calibration plot to get an idea of how the model might perform out of sample and adjust for optimism calibration.

The apparent and bias-corrected lines are compared with the ideal line. When either of the two lines is above the ideal line, the model under-predicts in that range of predicted probabilities. And when either of the lines is below the ideal line, the model over-predicts in that range of predicted probabilities. Hence, we can conclude from Figure 2 that the fitted model will under-predict between the range of 0.6 to 0.75 on both new and the current dataset.

The mean absolute error is the average absolute difference between predicted probability and actual probability. This implies that the average error of

the fitted model is 0.014, which is relatively smaller.

After fitting the model and assessing its significance some predictions were made. The next section provides a summary of these predictions.

Predictions Using the Fitted Logistic Regression

Predictions are used as forecast to determine what happens at a future time given current information. Since the main objective of this study was to formulate a mathematical model for predicting the probability of default of clients of a microfinance institution, it became necessary to use the fitted logistic regression model to make predictions to determine how effective the model will be when it is being used. The results of the predictions based on a sample dataset are shown in Table 11 below.

Table 11: Predictions from Fitted Logistic Regression Model

Customer Number	Branch Name	Rate	No. of Repay.	Avg. Inf. Rate	Avg. For. Ex. Rate	Predictions	Actual Observation
1	Ada.	0.6	6	12.8	2.17	0.9810	1
2	Ada.	0.78	6	13.2	2.32	0.9840	1
3	Pok.	0.84	6	17.25	3.89	0.9787	1
4	Ada.	0.78	6	17	3.75	0.9744	0
5	Ada.	0.72	6	16	3.10	0.9821	1
6	Pok.	0.6	6	15.05	4.21	0.9127	1
7	Ash.	0.6	4	10.00	4.55	0.3114	1
8	Ash.	0.6	4	9.80	4.60	0.2866	0
9	Tam.	0.6	6	9.10	1.77	0.9902	1
10	Tam.	0.72	4	14.90	2.86	0.9952	1
11	Tam.	0.48	4	10.00	4.55	0.4752	0
12	Bol.	0.72	6	16.25	3.44	0.9611	0
13	Bol.	0.72	6	15.25	2.86	0.9735	1
14	Ash.	0.6	5	9.70	4.59	0.3637	0
15	Ash.	0.6	4	10.15	4.56	0.3170	0
16	Pok.	0.72	4	16.15	4.07	0.9598	1
17	Pok.	0.84	8	17.4	3.90	0.9808	1
18	Bol.	0.6	6	11.45	4.41	0.6060	1
19	Bol.	0.48	6	9.55	4.60	0.3956	0
20	Tam.	0.6	4	9.35	4.98	0.5505	0

Source: Working Data (2019)

Table 11 shows the results from predictions from the fitted logistic regression model. The columns represent the Customer number, Branch name, Rate, Number of Repayment, Average Inflation Rate, Average Foreign Exchange Rate, Predictions and Actual observed default (1) or non-default (0) respectively. It can be seen that the predicted probability of default for Customer 1 is 0.9810 and the actual observation is a default. However, the predicted probability of default for Customer 4 is 0.9744 and the actual observation is a non-default. This may be attributed to the high rate and average inflation rate when compared to

Customer 1. Also, Customer 7 had a lower predicted probability of default but the actual observation was a default and this can be associated with a smaller number of repayment and a lower average inflation rate.

In the next section, the results obtained from the logistic regression are compared with previous related literature.

Discussion of Results in the Logistic Regression Model

The variables that were used in building the logistic regression model are shown in Table 4. The results obtained from the logistic regression model implied that in determining the probability of default for clients of the microfinance institution under study, the variables that should be considered are Rate, Branch Name, Number of Repayment, Average Inflation Rate, and Average Foreign Exchange Rate.

Schreiner (2004b) found that the variables gender and branch are relevant in determining the risk of a client. Agbemava et al. (2016) found out that the variable loan duration is significant in predicting the probability of default. Bensic et al. (2005) also found tat the length of time spent in repaying loans and interest rate are significant in determining the default rate of clients.

In the following section, the results obtained from using survival analysis models on the dataset are presented.

Model Building for Survival Analysis

In this section, the results obtained from the application of survival analysis model for the probability of default model are presented. Since survival analysis provides time specific estimates of probability of default, the variable Number of Repayment is used in addition the Default variable (0 for default and 1 for non default) as the dependent variable.

The sample used for building the survival analysis model for probability of default consists of 1060 customers. The application characteristics of these clients, macroeconomic variables, and variables that are linked to the social environment of Ghana were used.

The results obtained from the nonparametric estimation of the survival function, that is the model with no independent variables, are as follows.

Proposition 16

The result from Kaplan-Meier estimate of the survival function is shown in Figure 3 below.

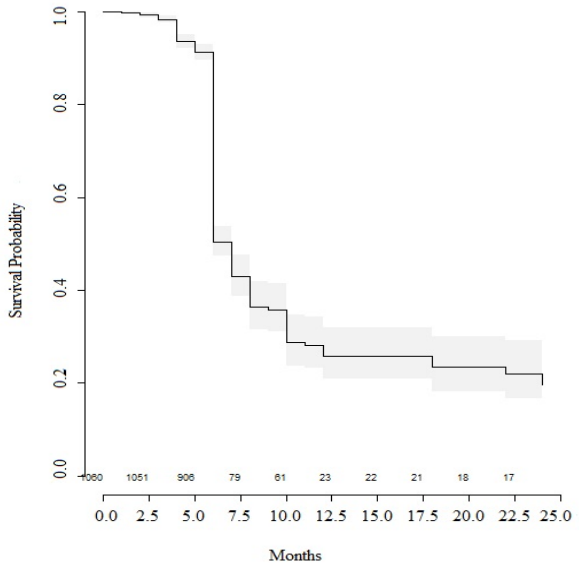


Figure 3: Plot of Kaplan-Meier Estimates of the Survival Function.

Figure 3 shows the survival function estimated by the Kaplan-Meier estimator. The y-axis represent the survival probability, the x-axis represent the time in months, and the values on top of the x-axis represent the number of individuals at risk at a specific time.

Proof. The results above can be derived from the methods for estimating the survival function discussed on page 42. □

The Cox proportional hazards model involves the use of independent variables in estimating the survival function. Let $X_i, i = 1, \cdots, n$ denote the independent variables. The X_i's are explained in Table 12 below.

Table 12: Variables for Cox Proportional Hazards Regression.

Variable	Meaning
X_1	Age
X_2	Rate
X_3	Gender
X_4	Branch Name
X_5	Marital Status
X_6	Mobile Network
X_7	Principal
X_8	Product
X_9	Average Unemployment Rate
X_{10}	Average Annual GDP Growth Rate
X_{11}	Average Inflation Rate
X_{12}	Bank of Ghana Lending Rate
X_{13}	Average Foreign Exchange Rate
X_{14}	Christmas
X_{15}	Easter
X_{16}	New Year
X_{17}	Ramadan
X_{18}	Academic Year

Source: Working Data (2019)

The results obtained from building the Cox proportional hazards model for probability of default following the model building process described on page 71.

1. The results obtained from the univariable analysis are as follows.

   Proposition 17

   The following covariates, with their respective p-values, shown in Table 13 were significant in the univariable Cox regression analysis.

Table 13: Significant Covariates from Univariable Cox Regression Analysis.

Variable	***p*-value**
Rate	< 0.0001
Branch Name	< 0.0001
Marital Status	< 0.0001
Product	< 0.0001
Network	< 0.0001
Average Annual GDP Growth Rate	< 0.0001
Average Inflation Rate	< 0.0001
Bank of Ghana Lending Rate	< 0.0001
Average Foreign Exchange Rate	< 0.0001
Christmas	< 0.0001
New Year	< 0.0001
Easter	0.0002
Ramadan	0.0017

Source: Working Data (2019)

The significant variables from the univariable Cox proportional hazards regression analysis are shown in Table 13.

Proof. The results in Table 13 above follow from the likelihood ratio test determine the significance of each covariate in a model discussed on page 64. □

2. In the multivariable analysis, the following results were obtained.

Proposition 18

The results for the significant variables from multivariable analysis are shown in Table 14 below.

Table 14: Significant Covariates from Multivariable Cox Regression Analysis.

Variable	p**-value**
Rate	< 0.0001
Branch Name	0.0001
Product	0.0030
Average Inflation Rate	0.0007
Bank of Ghana Lending Rate	< 0.0001
Average Foreign Exchange Rate	< 0.0001
Average Unemployment Rate	0.0002
Christmas	0.0278
New Year	< 0.0001
Easter	0.0009
Ramadan	< 0.0001

Source: Working Data (2019)

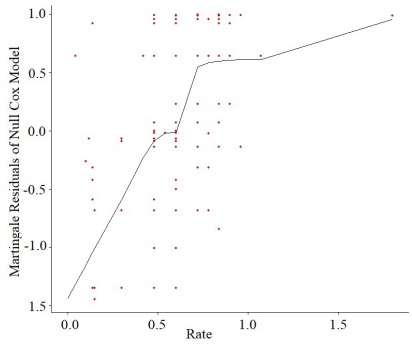
Table 14 shows the results from the multivariable analysis of the Cox proportional hazards regression.

Proof. The above results in Table 14 follow from the processes involved in building multivariable models discussed on page 71. □

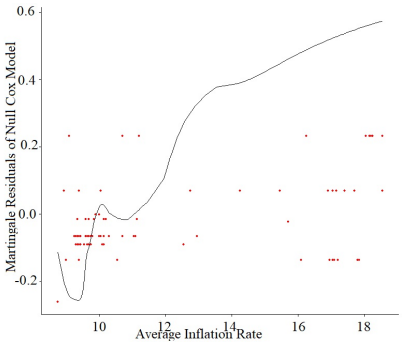
3. The next step involves assessing the functional form of the continuous covariates in the multivariable Cox regression model. From Table 14, the continuous covariates are Rate, Average Inflation Rate, Bank of Ghana Lending Rate, Average Foreign Exchange Rate, and Average Unemployment Rate. The following results were obtained.

   Proposition 19

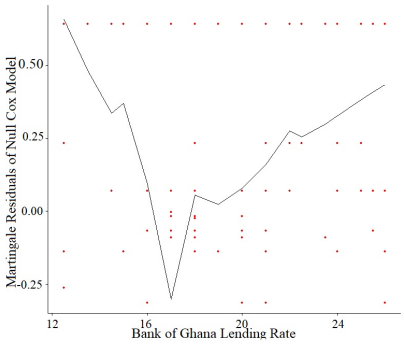
   The functional form of the continuous covariates can be assessed using martingale residuals and linearity test. The martingale residual plots for the continuous covariates are shown in Figure 4 below.



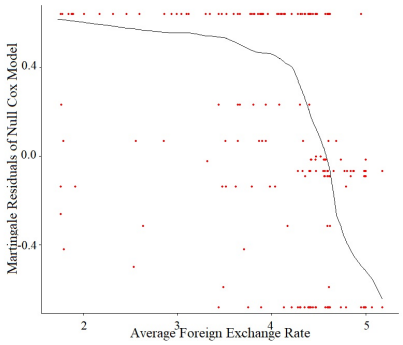
(a) Rate



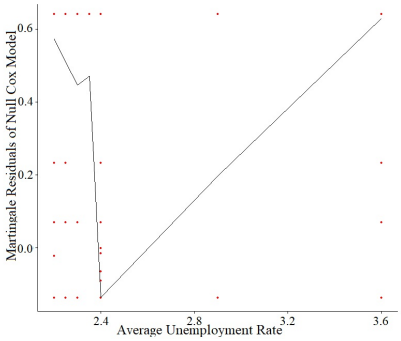
(b) Average Inflation Rate



(c) Bank of Ghana Lending Rate



(d) Average Foreign Exchange Rate



(e) Average Unemployment Rate

Figure 4: Martingale Residual Plots for Continuous Covariates in Multivariable Cox Regression Analysis

Figure 4 above shows the functional form of the continuous covariates using martingale residuals. Figure 4(a) shows the martingale residual plot for Rate, Figure 4(b) for Average Inflation Rate, Figure 4(c) for Bank of Ghana Lending Rate, Figure 4(d) for Average Foreign Exchange Rate and Figure 4(e) for Average Unemployment Rate. It can be seen that all the continuous variables in the Cox regression model are not linear in the log hazards. This graphical representation is usually not sufficient to conclude that all the continuous covariates affect the log hazard nonlinearly in the

Cox regression model, therefore a linearity test is conducted.

The results obtained from the linearity test in assessing the functional form of the continuous covariates are shown in Table 15 below.

Table 15: Results from Linearity Test for Cox Regression Model

Variable	***p*-value**
Rate	< 0.0001
Nonlinear	0.8276
Branch Name	0.0033
Product	0.0355
Christmas	0.7649
New Year	< 0.0001
Easter	0.0001
Ramadan	< 0.0001
Average Inflation Rate	< 0.0001
Nonlinear	< 0.0001
Average Unemployment Rate	< 0.0001
Nonlinear	< 0.0001
Bank of Ghana Lending Rate	0.6496
Nonlinear	0.5245
Average Foreign Exchange Rate	0.0040
Nonlinear	0.0719

Source: Working Data (2019)

Table 15 shows the results from the linearity test of the Cox Regression model. The results show that the continuous variables Average Inflation Rate and Average Unemployment Rate are nonlinear at an alpha level of 0.05. From Table 12, X_9= Average Unemployment Rate and X_{11}= Average Inflation Rate. The restricted cubic spline transformation of X_9 and X_{11}

with $k_i, i = 1, \cdots, n$ knots respectively are:

$$s(X_9) = \beta_{00} + \beta_{01}X_9 + \sum_{i=1}^{P-2} \beta_{i3}\Big[(X_9 - k_i)_+^3 - (X_9 - t_{P-1})_+^3\Big(\frac{t_P - k_i}{k_P - k_{P-1}}\Big) + (X_9 - k_P)_+^3\Big(\frac{k_{P-1} - k_i}{k_P - k_{P-1}}\Big)\Big] \tag{4.3}$$

and

$$s(X_{11}) = \beta_{00} + \beta_{01}X_{11} + \sum_{i=1}^{P-2} \beta_{i3}\Big[(X_{11} - k_i)_+^3 - (X_{11} - t_{P-1})_+^3\Big(\frac{t_P - k_i}{k_P - k_{P-1}}\Big) + (X_{11} - k_P)_+^3\Big(\frac{k_{P-1} - k_i}{k_P - k_{P-1}}\Big)\Big] \tag{4.4}$$

It can also be seen from Table 12 that some of the variables lose their significance due to the transformation of the continuous covariates. The significant variables in the multivariable are Rate, Branch Name, Product, New Year, Easter, Ramadan, Average Inflation Rate, Average Unemployment Rate, and Average Foreign Exchange.

Proof. The results above follow from the third step in model building process discussed on page 71, martingale residuals on page 62 and regression splines on page 69. □

4. The next result gives the interaction between the covariates.

Proposition 20

The following results in Table 16 show the significant interaction terms obtained from the interaction test between the covariates in the multivariable Cox regresssion model.

Table 16: Results from Interaction Test for Cox Regression Model

Variable	p**-value**
Rate and Product	< 0.0001
Ramadan and New Year	0.0051
Easter and Average Inflation Rate	0.0255

Source: Working Data (2019)

Proof. Using the test for interactions discussed under model building process on page 71, the above results were obtained. □

Based on the above results, the following theorem, which provides a mathematical model for probability of default, can be deduced.

Theorem 10

The hazard function of a client is given as:

$$h(t) = h_0(t)\exp(f(X;\beta,s,\alpha)), \tag{4.5}$$

$$\begin{aligned}\text{where } f(X;\beta,s,\alpha) =& \beta_2X_2 + \beta_4X_4 + \beta_8X_8 + s(X_9) + s(X_{11}) + \beta_{13}X_{13} \\ &+ \beta_{15}X_{15} + \beta_{16}X_{16} + \beta_{17}X_{17} + \alpha_1[X_2X_8] \\ &+ \alpha_2[s(X_{11})X_{15}] + \alpha_3[X_{16}X_{17}],\end{aligned}$$

with the X_i's given in Table 12, $s(X_9)$ in Equation 4.3 and $s(X_{11})$ given in Equation 4.4.

The survival function is given as:

$$S(t) = S_0(t)^{\exp(f(X;\beta,s,\alpha))}. \tag{4.6}$$

Then the probability of default is

$$P(\text{Default}) = 1 - S(t). \tag{4.7}$$

5. The next step involves model diagnostics. Under the Cox proportional hazards regression model diagnostics involve assessing the proportional hazards assumption, the overall goodness of fit of the model, among others. The results are as follows.

Proposition 21

The residuals used in this study for the Cox regression model include the Cox-Snell and Schoenfeld residuals. The results obtained from the Schoenfeld residuals are shown in Table 17 as follows.

Table 17: Results from Schoenfeld Residuals for Cox Regression Model

Variable	***p*-value**
Rate	0.9019
Branch Name	0.6869
Product	0.3793
Ramadan	0.6322
New Year	0.7249
Easter	0.3405
Average Inflation Rate	0.2071
Average Unemployment Rate	0.0950
Average Foreign Exchange Rate	0.1166
Rate and Product	0.6203
Ramadan and New Year	0.1836
Easter and Average Inflation Rate	0.3439
GLOBAL	0.0873

Source: Working Data (2019)

Table 17 shows the results from the Schoenfeld residuals or the proportional hazards assumption test. It can be seen that all the variables and the overall model (GLOBAL) are not significant at an alpha level of 0.05. This shows that the proportional hazards is satisfied by all the variables

and the overall model as well. The results obtained from the Cox-Snell residuals are shown in Figure 5 as follows.

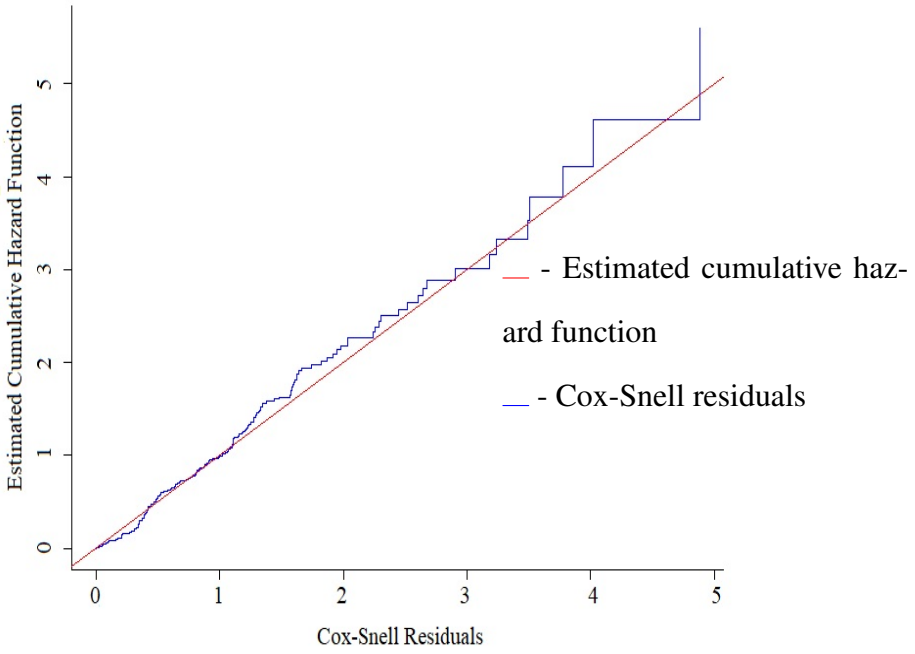


Figure 5: Plot of Cox-Snell Residuals

Figure 5 shows the plot of the Cox-Snell Residuals. The red line shows the estimated cumulative hazard function and the blue line shows the Cox-Snell residuals. The Cox-Snell residuals are narrowly scattered around the estimated cumulative hazard function. This indicates that the model is a good fit.

The model statistics for the fitted Cox Regression model are shown in Table 18 as follows.

Table 18: Model Statistics for Cox Regression Model

Model Tests		**Discrimination Indexes**	
LR χ^2	579.43	R^2	0.422
d.f.	30	C	0.817
Pr($>\chi^2$)	0.0000	D_{xy}	0.633
Score χ^2	697.37		
Pr($>\chi^2$)	0.0000		

Source: Working Data (2019)

Table 18 shows the model statistics for the Cox regression model. The model likelihood test (LR χ^2) is significant at an alpha level of 0.05 with a p-value (Pr($>\chi^2$)) of 0.0000. The score statistic is also significant at with a p-value of 0.0000. The R^2 index is approximately 42%, indicating how the fitted performs in comparison to the null model (model without independent variables). The C-statistic is approximately 82%, showing how the fitted model associates a higher prediction for a default than a non default.

Proof. Using the methods discussed under step 5 of model building processes on page 71 and residuals for the Cox regression model on page 62, the above results can be obtained. □

The next thing is to validate the fitted model. The results obtained from validating the fitted Cox regression model are shown in Table 19 as follows.

Table 19: Results from Validated Cox Regression Model

Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n
D_{xy}	0.6334	0.6232	0.5930	0.0302	0.6032	102*
R^2	0.4219	0.3879	0.3504	0.0375	0.3844	102*

* Divergence or singularity in 98 samples

Source: Working Data (2019)

Table 19 shows the results from the validated Cox regression model. The columns represent those of Table 10 on page 85. The $*$ (Divergence or singularity in 98 samples) indicate that 98 samples of the bootstrap did not yield results for the measures that are being validated. The D_{xy} for the original sample is 0.6334 and the optimism corrected is 0.6032 which shows the model will perform better on future observations. The R^2 index is 0.4219 and that of the optimism corrected is 0.3844, which also indicates a good fit.

After fitting and evaluating the fitted survival analysis model, predictions were made using a sample dataset. The results are presented in the next section.

Predictions Using the Survival Analysis Model

The survival analysis model was used to make predictions after it was fitted and evaluated. The probability of default for this model is given at a specified time since the model presents time bound estimates. Therefore we considered only predictions at time $t = 6$ months. It is worthy to note that a person is said to default if the time of the estimated probability of default is the same as the number of repayment. That is, if a person is paying for 6 months, the predicted probability of default is given at time $t = 6$. The results for the predicted probability of default for a sample of clients at time $t = 6$ with 6 months repayment time are shown in Table 20 below.

Table 20: Predictions from Fitted Cox Regression Model

Customer No.	Branch Name	Rate	Product (Type of Loan)	Ramadan	New Year	Easter	Avg. Inf. Rate	Avg. Unemp. Rate	Avg. For. Rate	Predictions	Actual Observation
1	Ada.	0.6	Personal	0	1	0	12.8	2.2	2.17	0.8692	1
2	Ada.	0.72	Personal	1	0	0	16	2.2	3.10	0.9507	1
3	Ada.	0.9	Others	0	1	1	16.95	2.2	3.60	0.7472	1
4	Ada.	0.15	Staff	0	1	0	9.5	2.4	4.74.	0.0247	0
5	Pok.	0.84	Edwumapa	1	0	0	17.25	2.2	3.90	0.9642	1
6	Pok.	0.6	Personal	0	1	1	9.35	2.4	5.00	0.0845	0
7	Pok.	0.78	Others	1	0	1	17.1	2.2	3.82	0.6726	1
8	Pok.	0.48	Edwumapa	0	1	0	9.55	2.4	4.95	0.2834	1
9	Ash.	0.6	Personal	1	0	0	11.95	2.4	4.39	0.5792	0
10	Ash.	0.6	Edwuamapa	0	1	1	10.6	2.4	4.39	0.6378	1
11	Ash.	0.84	Personal	1	0	0	17.2	2.4	3.91	0.8143	1
12	Ash.	0.48	Others	0	1	0	9.55	2.4	4.95	0.1854	0
13	Bol.	0.72	Personal	1	0	1	15.5	2.2	2.94	0.7547	1
14	Bol.	0.48	Edwumapa	1	1	1	10.9	2.4	4.47	0.0859	1
15	Bol.	0.6	Salary	1	0	1	10.1	2.4	4.59	0.4930	0
16	Bol.	0.6	Others	1	0	1	10.1	2.4	4.59	0.5169	0
17	Tam.	0.6	Personal	0	1	1	9.1	3.6	1.77	0.8540	1
18	Tam.	0.84	Edwumapa	1	0	0	17.4	2.2	4.06	0.9950	1
19	Tam.	0.48	Edwumapa	1	0	1	10.25	2.4	4.57	0.8018	0
20	Tam.	0.6	Personal	1	0	1	10.1	2.4	4.59	0.8087	1

Source: Working Data (2019)

Table 20 shows the predicted probability of default of a sample of 20 clients at time $t = 6$ with 6 months repayment time. The columns represent Customer number, Branch Name, Rate, Product, Ramadan, New Year, Easter, Average Inflation Rate, Average Unemployment Rate, Predictions and Actual Observations from the dataset respectively. It can be seen that Customer 1's predicted probability of default is 0.8692 and the actual observation is a default (1). Also, Customer 4's predicted probability is 0.0247 and the actual observation is a non-default (1). However, Customer 19's actual observation is a non-default but the predicted probability of default is 0.8018 which can be attributed to a higher average foreign exchange rate as compared to Customer 1.

Chapter Summary

The results obtained from analysing the data with the logistic regression and survival analysis models were presented. From the logistic regression model, it was found that the variables Rate, Branch Name, Number of Repayment, Av-

erage Inflation Rate and Average Foreign Exchange Rate were significant in predicting the probability of default. There was also a transformation of Number of Repayment to include nonlinear terms. The residuals also showed that the model has a good fit.

The results obtained from the survival analysis model showed that the variables Rate, Branch Name, Product, New Year, Easter, Ramadan, Average Inflation Rate, Average Unemployment Rate, Bank of Ghana Lending Rate and Average Foreign Exchange Rate were significant in determining the probability of default. There were nonlinear terms included in the model for Average Inflation Rate and Average Unemployment Rate. Also, there were interactions between the variables Rate and Product, Ramadan and New Year, and Easter and Average Inflation Rate. The residuals also indicated that the model has a good fit.

CHAPTER FIVE

SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

Overview

In this thesis, a mathematical model for predicting the probability of default of clients of a microfinance institution has been presented. The variables used include characteristics of the borrower, loan characteristics, macroeconomic indicators, and some specific seasons in Ghana. The methods used in building the mathematical model were discussed with their results. A summary, some conclusions and recommendations of the results obtained are presented in this chapter.

Summary

Most financial institutions, including microfinance, face the problem of loan default from their loan clients. This affects the profit margin of these institutions. It therefore becomes very necessary for these financial institutions to control the level of clients' loan default. The credit scoring technique has since been used as one of the measures in controlling loan default of clients.

In this thesis, a mathematical model for predicting the probability of default of clients of a microfinance institution was formulated using logistic regression and survival analysis models. The main focus was on clients who take individual loans.

First of all, a logistic regression model was used as the mathematical model. Application variables including loan characteristics and borrower information, and macroeconomic variables were used in the model. This model used a dataset of clients whose repayment time is past due which includes those who finished and those who did not finish paying their loans on or before the repayment time. The variables Rate, Branch Name, Number of repayment, Average Inflation Rate and Average Foreign Exchange Rate were significant in the model as shown in Table 6. The functional form of each continuous variable in the model was assessed and Number of Repayment as shown in Table 7 was

found to be nonlinear in relation to the log odds, so it was transformed with restricted cubic splines to have a linear relationship with the log odds in Equation 4.1. A test for interactions was done for the variables in the model but none was significant. The final mathematical model for predicting the probability of default for clients with individual loans was presented in Theorem 9. The model was evaluated based on the steps outlined on page 64. The model was also validated with the results showing that it had a good fit as shown in Table 10. Predictions were then made based on the fitted model and the results are presented in Table 11.

Then, a survival analysis model was also used as the mathematical model. Here, a new variable indicating some seasons or periods of the year were added to the application and macroeconomic variables. The dataset used included all loans whose repayment time is past due and those who are still paying. The variables Rate, Branch Name, Product, New Year, Easter, Ramadan, Average Inflation rate, Average Unemployment rate and Average Foreign Exchange rate were significant in the model. The functional form of the continuous variables in the model was examined and the results in Table 15 showed that the variables Average Inflation rate and Average Unemployment rate were nonlinear in relation to the log hazards. Hence, they were transformed with restricted cubic splines to meet the linearity assumption as shown in Equations 4.3 and 4.4 for Average Inflation rate and Average Unemployment rate respectively. An interaction test was conducted and it showed interactions between the variables Ramadan and New Year, Rate and Product, and Easter and Inflation Rate as shown in Table 16. The model for predicting probability of default was presented in Theorem 10. The fitted model was evaluated with the measures outlined on page 64 and validated with results shown in Table 19 which indicates a good fit. After that, predictions were made based on the fitted model at time $t = 6$ with loans with 6 months repayment time. The results are shown in Table 20.

The next section presents conclusions drawn from the methods used in

building a mathematical model for predicting probability of default of clients with individual loans from a microfinance institution.

Conclusion

The main objective of this study was to build a mathematical model for predicting the probability of default of clients taking individual loans from a microfinance institution in Ghana. Logistic regression and survival analysis models were used in building the mathematical model.

The logistic regression model was the first model applied on a dataset from a microfinance institution to build a mathematical model for the institution. The logistic regression model showed that the variables Rate, Number of Repayment, Average Inflation rate and Average Foreign Exchange Rate were significant in predicting the probability of default of a client who takes individual loans. The Number of Repayment variable was transformed using restricted cubic splines. The evaluation and validation results showed that the model was a good model. Most of the predictions of the model were similar to those of the observed outcomes however, there were some few which did not match the actual outcomes and it could be attributed to an extreme value in one of the variables. This model provided predictions usually at the end of a repayment time.

The survival analysis model, specifically the Cox proportional hazards regression model was used to introduce time dynamics in the model for predicting the probability of default. New variables linked to the Ghanaian environment were introduced in the model. The results showed that some of these variables such as Easter, New Year and Ramadan should be included in the model for predicting the probability of default. The fitted model was also evaluated and validated with good results.

Recommendations

Based on the results found in this study, the following recommendations are made.

First of all, the microfinance institution should adopt better data manage-

ment systems or practices. They include keeping records of clients' repayment history and arrears, loans that have been closed, among others.

The microfinance institution should also adopt the model proposed in the study to be able to make informed decisions about, and objective assessment of credit risk.

The microfinance institution should provide some incentives for those who pay their loans on time which will serve as a positive reinforcement for them and those who do not pay also.

The interest rate on loans of microfinance institutions are usually high. This can be attributed to the cost involved in providing small loans. It therefore becomes necessary to compute the optimal interest rate that would favour both the lender and borrower. This has been left for further research.

REFERENCES

Abid, L., Masmoudi, A., & Zouari-Ghorbel, S. (2016). The consumer loan's payment default predictive model: An application in a tunisian commercial bank. *Asian Economic and Financial Review*, *16*(1), 27.

Agbemava, E., Nyarko, I. K., Adade, T. C., & Bediako, A. K. (2016). Logistic regression analysis of predictors of loan defaults by customers of non-traditional banks in ghana. *European Scientific Journal*, *12*(1).

Asiama, J., & Osei, V. (2007). A note on microfinance in ghana. *Bank of Ghana, Accra*.

Badugu, D., & Tripathi, V. K. (2016). Role of microfinance institutions in the development of entrepreneurs. *International Journal of Accounting Research*, *42*(3419), 1–15.

Banasik, J., Crook, J. N., & Thomas, L. C. (1999). Not if but when will borrowers default. *Journal of the Operational Research Society*, *50*(12), 1185–1190.

Bellotti, T., & Crook, J. (2009). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, *60*(12), 1699–1707.

Bensic, M., Sarlija, N., & Zekic-Susac, M. (2005). Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. *Intelligent Systems in Accounting, Finance & Management: International Journal*, *13*(3), 133–150.

Blavy, M. R., Basu, M. A., & Yülek, M. Â. (2004). *Microfinance in africa: Experience and lessons from selected african countries (epub)* (No. 4-174). International Monetary Fund.

Boateng, A. A. (2015). An examination of challenges and prospects of micro-finance institutions in ghana. *Journal of Economics and Sustainable Development*, *6*(4), 52–60.

Bojanov, B. D., Hakopian, H., & Sahakian, B. (2013). *Spline functions and multivariate interpolations* (Vol. 248). Springer Science & Business Media.

Castañeda, L. B., Arunachalam, V., & Dharmaraja, S. (2012). *Introduction to probability and stochastic processes with applications*. Wiley Online Library.

Collett, D. (2002). *Modelling binary data*. CRC Press.

Coravos, A. R. (2010). *Measuring the likelihood of small business loan default* (Unpublished doctoral dissertation). Duke University.

Czepiel, S. A. (2002). Maximum likelihood estimation of logistic regression models: theory and implementation. *Available at czep. net/stat/mlelr. pdf*.

Dirick, L., Claeskens, G., & Baesens, B. (2017). Time to default in credit scoring using survival analysis: a benchmark study. *Journal of the Operational Research Society*, *68*(6), 652–665.

Fatemi, A., & Fooladi, I. (2006). Credit risk management: a survey of practices. *Managerial Finance*, *32*(3), 227–233.

Fleming, T. R., & Harrington, D. P. (2011). *Counting processes and survival analysis* (Vol. 169). John Wiley & Sons.

Fund, U. N. C. D. (2006). *Building inclusive financial sectors for development*. United Nations Publications.

Gill, R. D. (1984). Understanding cox's regression model: a martingale approach. *Journal of the American Statistical Association*, *79*(386), 441–447.

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *160*(3), 523–541.

Harrell Jr, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

Hassan, M. K., Brodmann, J., Rayfield, B., & Huda, M. (2018). Modeling credit risk in credit unions using survival analysis. *International Journal of Bank Marketing*, *36*(3), 482–495.

Heumann, C., Schomaker, M., et al. (2016). *Introduction to statistics and data analysis*. Springer.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Kleinbaum, D. G., & Klein, M. (2010). *Survival analysis* (Vol. 3). Springer.

Kocenda, E., & Vojtek, M. (2009). Default predictors and credit scoring models for retail banking. *CESifo Working Paper Series*.

Korn, E. L., & Graubard, B. I. (2011). *Analysis of health surveys* (Vol. 323). John Wiley & Sons.

Kwofie, C., Owusu-Ansah, C., & Boadi, C. (2015). Predicting the probability of loan-default: An application of binary logistic regression. *Research Journal of Mathematics and Statistics*, *7*(4), 46–52.

Lafourcade, A.-L., Isern, J., Mwangi, P., & Brown, M. (2005). Overview of the outreach and financial performance of microfinance institutions in africa. *Microfinance Information eXchange, Washington, DC. Retrieved from http://www. mixmarket. org/medialibrary/mixmarket/ Africa_Data_Study. pdf*.

Leung, K.-M., Elashoff, R. M., & Afifi, A. A. (1997). Censoring issues in survival analysis. *Annual review of public health*, *18*(1), 83–104.

Liu, X. (2012). *Survival analysis: models and applications*. John Wiley & Sons.

Malik, M., & Thomas, L. C. (2010). Modelling credit risk of portfolio of consumer loans. *Journal of the Operational Research Society*, *61*(3), 411–420.

Morduch, J. (1999). The microfinance promise. *Journal of economic literature*, *37*(4), 1569–1614.

Njeru Warue, B. (2012). Factors affecting loan delinquency in microfinance institutions in kenya. *International Journal of Management Sciences and Business Research*, *1*(12).

Noh, H. J., Roh, T. H., & Han, I. (2005). Prognostic personal credit risk model considering censored information. *Expert Systems with Applications*, *28*(4), 753–762.

Ofori, K. S., Fianu, E., Omoregie, K., Odai, N. A., & Oduro-Gyimah, F. (2014). Predicting credit default among micro borrowers in ghana. *Research Journal of Finance and Accounting*, *5*(12), 96–104.

Ojiako, U., Nguyen, T.-D., & Shen, S.-W. (2013). Modelling the predictive performance of credit scoring. *Professional Accountant*, *13*(1), 1–12.

on Banking Supervision, B. C. (2004). International convergence of capital measurement and capital standards: a revised framework. *Basel Committee on Banking Supervision, Bank for International Settlements*.

Otero, M. (1999). Bringing development back, into microfinance. *Journal of Microfinance/ESR Review*, *1*(1), 2.

Rychnovskỳ, M. (2018). Survival analysis as a tool for better probability of default prediction. *Acta Oeconomica Pragensia*, *2018*(1), 34–46.

Schreiner, M. (2004a). Benefits and pitfalls of statistical credit scoring for microfinance/ventajas y desventajas del scoring estadístico para las microfinanzas/vertus et faiblesses de l'évaluation statistique (credit scoring) en microfinance. *Savings and Development*, 63–86.

Schreiner, M. (2004b). Scoring arrears at a microlender in bolivia. *Journal of Microfinance/ESR Review*, *6*(2), 5.

Smith, P. L. (1979). Splines as a useful and convenient statistical tool. *The American Statistician*, *33*(2), 57–62.

Stepanova, M., & Thomas, L. (2002). Survival analysis methods for personal loan data. *Operations Research*, *50*(2), 277–289.

Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, *16*(2), 149–172.

Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). *Credit scoring and its applications*. SIAM.

Viswanathan, P. K., & Shanthi, S. K. (2017). Modelling credit default in microfinance — an indian case study. *Journal of Emerging Market Finance*, *16*(3), 246–258.

Wegman, E. J., & Wright, I. W. (1983). Splines in statistics. *Journal of the American Statistical Association*, *78*(382), 351–365.

Westgaard, S., & Van der Wijst, N. (2001). Default probabilities in a corporate bank portfolio: A logistic model approach. *European Journal of Operational Research*, *135*(2), 338–349.

Wold, S. (1974). Spline functions in data analysis. *Technometrics*, *16*(1), 1–11.

Yeboah, B. E. (2012). *Predicting microfinance credit default (a study of nsoatreman rural bank, sunyani)* (Unpublished doctoral dissertation). University of Health and Allied Sciences.

Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, *36*(2), 2473–2480.

Zhang, Z. (2016). Model building strategy for logistic regression: purposeful selection. *Annals of translational medicine*, *4*(6), 111–114.

APPENDIX

Logistic Regression Code

```
library(car)
library(caret)
library(lattice)
library(ROCR)
library(pROC)
library(MASS)
library(splines)
library(ResourceSelection)
library(MKmisc)
library(pscl)
library(rms)
library(gmodels)
library(ggplot2)
library(readxl)
library(rpart)
library(Hmisc)
library(ggfortify)
library(broom)
library(sure)
library(dplyr)
library(mgcv)
```

Univariable Analysis

```
Mn1 <- lrm(Default=Age,data = MN)
Mn1
Mn1A <- lrm(Default=Rate,data = MN)
```

```
Mn1A
Mn1B <- lrm(Default=Principal,data = MN)
Mn1B
Mn1C <- lrm(Default=Gender,data =MN)
Mn1C
Mn1D <- lrm(Default=BranchName,data =MN)
Mn1D
Mn1E <- lrm(Default=MaritalStatus,data =MN)
Mn1E
Mn1F <- lrm(Default=Product,data =MN)
Mn1F
Mn1G <- lrm(Default=No.ofRepay.,data=MN)
Mn1G
Mn1H <- lrm(Default=Network,data=MN)
Mn1H
Mn1I <- lrm(Default=Christmas,data=MN)
Mn1I
Mn1J <- lrm(Default=NewYear,data=MN)
Mn1J
Mn1K <- lrm(Default=Easter,data=MN)
Mn1K
Mn1L <- lrm(Default=AcademicYear,data=MN)
Mn1L
Mn1M <- lrm(Default=Ramadan,data=MN)
Mn1M
Mn1N <- lrm(Default=Festival,data=MN)
Mn1N
Mn1O <- lrm(Default=AverageUR,data=MN)
Mn1O
```

```
Mn1P <- lrm(Default=AAGDP,data=MN)
Mn1P
Mn1Q <- lrm(Default=AverageIR,data=MN)
Mn1Q
Mn1R <- lrm(Default=BoGLR,data=MN)
Mn1R
Mn1S <- lrm(Default=AEFXRate,data=MN)
Mn1S
```

Initial Multivariable model with significant variables from univariable analysis

```
Mn4A <- lrm(Default = Age+Rate+BranchName+MaritalStatus +No.ofRepay.
+ Network+AAGDP+AverageIR+BoGLR+AEFXRate ,data=MN,x=TRUE,
y=TRUE)
Mn4A
anova(Mn4A)
```

Significant variables from Initial Multivariable Analysis

```
Mn4A. <- lrm(Default = Rate+ BranchName+No.ofRepay.+AverageIR
+AEFXRate, data=MN, x=TRUE, y=TRUE) Mn4A.
anova(Mn4A.)
```

Second Multivariable Analysis

```
Mn4B <- lrm(Default = Rate+BranchName +No.ofRepay.+AverageIR
+AEFXRate+Principal +Product,data=MN,x=TRUE,y=TRUE)
Mn4B
anova(Mn4B)
```

Significant parameters from Second Multivariable Analysis

```
Mn4B. <- lrm(Default=Rate+BranchName +No.ofRepay.+AverageIR
+AEFXRate,data=MN,x=TRUE,y=TRUE)
Mn4B.
anova(Mn4B.)
```

Linearity Test

```
resid(Mn4B.,'partial',pl=TRUE)
Mn4C <- lrm(Default=rcs(Rate)+BranchName +rcs(No.ofRepay.)
+rcs(AverageIR)+rcs(AEFXRate),data=MN,x=TRUE,y=TRUE)
Mn4C
anova(Mn4C)
```

Interaction Test

```
Mn4D <- lrm(Default=BranchName+AEFXRate
+ AvereIR+Rate*rcs(No.ofRepay.),data=MN,x=TRUE,y=TRUE)
Mn4D
anova(Mn4D)
resid(Mn4D,'gof')
```

Validation of fitted Model

```
set.seed(10)
validate(Mn4D,B=200)
plot(calibrate(Mn4D,B=200))
```

Predictions

```
Mn4E <- predict(Mn4D,type = 'fitted')
```