



**University of Venda**

**VARIABLE SELECTION IN DISCRETE SURVIVAL  
MODELS**

By

**Mabvuu Coster**

**14012421**

submitted in fulfilment of the requirements for the degree of Master of  
Science in Statistics

in the

Department of Statistics  
School of Mathematical and Natural Sciences  
University of Venda  
Thohoyandou, Limpopo  
South Africa

Supervisor : Dr. A. Bere

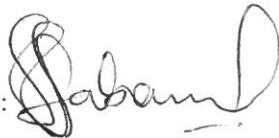
Co-Supervisor : Dr. C. Sigauke

Date submitted: 27 February 2020

## Declaration

I, **Coster Mabvuu** (student number 14012421), declare that this dissertation titled: “**Variable selection in discrete survival models**” hereby submitted for the **Master of Science degree in Statistics** at the **University of Venda** is my original work and has not been submitted for any degree at any other university or institution. The dissertation does not contain any other persons’ writings unless specifically acknowledged and referenced accordingly.

Signature:



Date: 27/02/2020

# Table of Contents

<b>Table of Contents</b>	<b>iii</b>
<b>Abstract</b>	<b>vi</b>
<b>Acknowledgements</b>	<b>vii</b>
<b>1</b>	<b>1</b>
1.1 Introduction . . . . .	1
1.1.1 Censoring . . . . .	2
1.1.2 Hazard function ( $\lambda(t)$ ) . . . . .	2
1.1.3 Survival function . . . . .	3
1.1.4 Cox Model . . . . .	3
1.1.5 Discrete versus continuous survival models . . . . .	4
1.1.6 Discrete Survival Model . . . . .	4
1.2 Variable Selection . . . . .	5
1.2.1 Stepwise procedures . . . . .	5
1.2.2 Penalised likelihood approaches . . . . .	6
1.2.3 Boosting . . . . .	6
1.3 Statement of the Problem . . . . .	7
1.4 Aims and Objectives . . . . .	8
1.4.1 Aim of the study . . . . .	8
1.4.2 Objectives . . . . .	8
1.5 Proposed methodology . . . . .	9
1.6 Layout of the project . . . . .	9
<b>2 Literature Review</b>	<b>10</b>
2.1 Stepwise procedures . . . . .	10
2.2 Penalised likelihood variable selection . . . . .	12
2.3 Boosting . . . . .	13

2.4	Unobserved heterogeneity . . . . .	14
2.5	Age at first marriage . . . . .	15
<b>3</b>	<b>Methodology</b>	<b>18</b>
3.1	Basic discrete survival model . . . . .	18
3.2	Commonly used link functions . . . . .	19
3.2.1	Logistic Discrete Hazard Model . . . . .	19
3.2.2	Grouped Proportional Hazard Model . . . . .	20
3.2.3	Probit Model . . . . .	20
3.3	The baseline hazard . . . . .	21
3.4	Penalised estimation and variable selection . . . . .	22
3.4.1	Models without unobserved heterogeneity . . . . .	22
3.4.2	Models with unobserved heterogeneity . . . . .	25
3.4.3	Software packages for penalised regression . . . . .	28
3.5	Boosting estimation . . . . .	28
3.6	Inferences on predictors . . . . .	31
3.6.1	Wald's test . . . . .	32
3.6.2	Likelihood ratio test . . . . .	32
3.6.3	Score test . . . . .	32
3.7	Goodness-of fit of the model . . . . .	32
3.8	Model Prediction Accuracy . . . . .	33
3.8.1	Predictive Deviance . . . . .	34
3.8.2	Concordance Index . . . . .	34
3.9	Application: Age at first marriage . . . . .	36
3.9.1	Data description . . . . .	36
3.9.2	Model building . . . . .	37
<b>4</b>	<b>Simulation study</b>	<b>39</b>
4.1	Simulation approach . . . . .	39
4.2	Simulation Results . . . . .	42
4.3	Conclusion Remarks . . . . .	45
<b>5</b>	<b>Empirical data analysis results</b>	<b>46</b>
5.1	Descriptive Results . . . . .	46
5.1.1	Univariate analysis . . . . .	46
5.1.2	Bivariate analysis . . . . .	48
5.2	Estimated linear effects . . . . .	53
5.2.1	Estimation from <i>lasso</i> and <i>gradient boosting</i> without random effect . . . . .	53

5.2.2	Estimation from <i>lasso</i> and <i>gradient boosting</i> with random effect	54
5.2.3	Boosting iterations . . . . .	55
5.3	Model diagnostics . . . . .	57
5.3.1	Residual analysis . . . . .	57
5.3.2	Concordance Index . . . . .	58
5.4	Model comparison . . . . .	58
5.5	Interpretation of the results based on the model with <i>lasso</i> variable selection . . . . .	59
<b>6</b>	<b>Conclusion</b>	<b>64</b>
6.1	Introduction . . . . .	64
6.2	Conclusion . . . . .	64
6.3	Limitations of the study . . . . .	65
6.4	References . . . . .	66
6.5	Appendix . . . . .	73

# Abstract

Selection of variables is vital in high dimensional statistical modelling as it aims to identify the right subset model. However, variable selection for discrete survival analysis poses many challenges due to a complicated data structure. Survival data might have unobserved heterogeneity leading to biased estimates when not taken into account. Conventional variable selection methods have stability problems. A simulation approach was used to assess and compare the performance of Least Absolute Shrinkage and Selection Operator (Lasso) and gradient boosting on discrete survival data. Parameter related mean squared errors (MSEs) and false positive rates suggest Lasso performs better than gradient boosting. Frailty models outperform discrete survival models that do not account for unobserved heterogeneity. The two methods were also applied on Zimbabwe Demographic Health Survey (ZDHS) 2016 data on age at first marriage and did not select exactly the same variables. Gradient boosting retained more variables into the model. Place of residence, highest educational level attained and age cohort are the major influential factors of age at first marriage in Zimbabwe based on Lasso.

**Keywords:** *Boosting, discrete-time hazard model, Lasso, penalised variable selection methods, unobserved heterogeneity,*

# Acknowledgements

My deepest gratitude to God, The Almighty for empowering and guiding me throughout this research.

Secondly, my deep acknowledgement to my supervisors Dr Bere A and Dr Sigauke C for exemplary guidance, helpful suggestions, valuable feedback and continual motivation throughout the duration of this research.

Lastly, I want to thank my family and friends for their kind co-operation, spiritual and financial support and encouragement which helped me in completing this research.

## List of Abbreviations

AIC	Akaike's Information Criterion.
BIC	Bayesian Information Criterion.
CI	Confidence Interval.
ENET	Elastic Net.
FNR	False Negative Rates.
FPR	False Positive Rates.
GAM	Generalised Additive Model.
GLMM	Generalised Linear Mixed Model.
LASSO	Least Absolute Shrinkage and Selection Operator.
MLE	Maximum Likelihood Estimation.
MSE	Mean Squared Error.
SCAD	Smooth Clipped Absolute Deviation.
se	Standard Error.
UNICEF	United Nations Children's Fund.
ZDHS	Zimbabwe Demographic Health Survey.
ZIMSTATS	Zimbabwe National Statistics Agency.



# Chapter 1

This chapter briefly discusses basic concepts in survival analysis and variable selection, problem statement, objectives, proposed methodology and layout of the study.

## 1.1 Introduction

Survival analysis is a compilation of statistical methods used to describe, explain, or predict the occurrence and timing of events. The event can be unemployment, occurrence of disease, death, marriage or divorce. The duration prior to the occurrence of a particular event can be measured in days, weeks, months or years. Survival analysis is a highly active area of research with applications in various fields of study which include engineering, physical, biological and social sciences (Gould et. al., 2008; Klein and Goel, 1992). The other names for survival analysis are event history analysis, failure time analysis, duration analysis or transition analysis. The happening of the event of interest is termed ‘failure’. Survival time means the duration for the failure to occur usually denoted by  $T$ . The random variable  $T$  is presumably to be positive . The methods of survival analysis are able to handle right censoring (as explained in subsection 1.1.1 below), a common phenomenon in longitudinal data.

### 1.1.1 Censoring

Survival data is special because of censoring. Censoring represents a particular type of missing data or data with partial information. It is a phenomenon that occurs when there is no fully observation of an individual's survival time. The precise duration within one condition is not known. Censoring for instance, occurs if there is an early drop out from a study by an observation (observations are lost prior to the occurrence of a phenomenon), or if the happening of an observation occurs later than the time an investigation has been concluded (Groll and Tutz, 2016). Censoring has three forms namely right, left as well as interval censoring. With right censoring, the starting of a spell is well known even though there is no observation on the time a transition takes place. It is a phenomenon that occurs when the observed time is shorter than the survival time. Right censoring can be an instance if a subject experience different event separately from the cause of interest, the study has been finished though the subject endures. Left censoring is when the event of interest has happened prior to the observation date though it is unknown exactly when. The entry to state is unknown though the end of the spell is observed. For example, a situation whereby the day of medical examination disclosing a sickness is set on yet when a patient has been infected is not observed. Lastly, interval censoring is when the event arise in a time span but it is unknown precisely when in the interval. Censoring is independent when it is unrelated to event occurrence.

### 1.1.2 Hazard function ( $\lambda(t)$ )

The hazard function relates the instant rate of occurrence of a phenomenon in subjects who are currently at risk of the event (Austin and Leeds, 2016). It is defined as

the conditional probability for the failure within the interval  $[a_{t-1}, a_t)$  provided the interval is arrived at. The hazard function is defined as:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \quad (1.1.1)$$

The hazard function takes on any shape of a non-negative function.

### 1.1.3 Survival function

The survival function provides the likelihood of surviving later than a specified time period  $t$  and can be defined as :

$$S(t) = P(T > t) \quad (1.1.2)$$

$$= 1 - F(t), \quad (1.1.3)$$

where  $F(t)$  is the cumulative density function of a given distribution (Kalbfleisch and Prentice, 2002). The survival function is a downward sloping curve, non-increasing and is estimated by using the Kaplan- Meier method. Note that  $S(0)=1$ ,  $S(t) \rightarrow 0$  as  $t \rightarrow \infty$

### 1.1.4 Cox Model

The Cox model is a statistical technique applied to associate various risk factors or, exposures, considered simultaneously, with survival time (Cox, 1972). The Cox model is a continuous-time specification. In a Cox proportional hazard model, the measure of the effect is the hazard rate, which is the possibility of failure given that the participator has lived up to time  $t$ . The proportional hazards model specifies that:

$$\lambda(t | \mathbf{x}) = \gamma_0(t) \exp(\mathbf{X}^T \boldsymbol{\gamma}),$$

where  $\gamma_0(t)$  is the baseline hazard function (the expected hazard without the effect of the considered factors). The  $\exp(\mathbf{X}^T\boldsymbol{\gamma})$  gives the influence of predictors on the hazard. The baseline hazard function can assume any form while the covariates enter linearly hence the model is semi-parametric.

### 1.1.5 Discrete versus continuous survival models

Survival time is either continuous or discrete. The assumptions of continuous models are that the event to be modelled can arise at anytime and survival time is continuous. When the exact time at which the events occur is known it is appropriate to use these models. However, discrete survival-time models assume duration time is banded or grouped into discrete intervals (Hess and Persson, 2010). Discrete models are better models in the presence of ties (at least two individuals experiencing a phenomenon at the same recorded time). Discrete time models do not have a strict proportional hazard assumption. They can be formulated as binary responses models. This enables use of software and methods for binary response models for estimation. In this study the discrete survival model is adopted because of its advantages and also because the response in the application data is measured in whole years.

### 1.1.6 Discrete Survival Model

In the case of discrete time, the discrete hazard function is given by:

$$\lambda(t|\mathbf{x}) = P(T = t|T \geq t, \mathbf{x}), t = 1, 2, \dots \quad (1.1.4)$$

Equation (1.1.5) is the discrete hazard as a function of covariates. It is the conditional likelihood for failure in the interval  $[a_{t-1}, a_t)$  provided the interval is arrived at.

A discrete survival model has the form:

$$\lambda(t|x) = h(\gamma_{0t} + \mathbf{X}^T \gamma), \quad (1.1.5)$$

where  $h(\cdot)$  is a set response function,  $\gamma_{0t}$  is a time varying intercept taken as a baseline effect ignoring any set of predictors,  $\gamma$  and  $\mathbf{X}$  are the weights and covariates, respectively.

This is a statistical technique applied to associate risk factors with ‘survival’ time when the events arise within a given time period. The risk is evaluated as a conditional probability that a phenomenon will happen. Discrete time hazard model makes an assumption of simultaneous occurrence of two or more observations. Hence, discrete-time hazard models are proposed in the case of heavy ties (Singer and Willet, 2003).

## 1.2 Variable Selection

The major aim of variable selection is to choose an ideal subset of predictors. It helps in the identification of null predictors thereby improving the prediction performance of the fitted model. Classical techniques such as stepwise, penalised likelihood and bootstrap variables selection methods are used.

### 1.2.1 Stepwise procedures

#### Backward elimination

The backward elimination procedure commences with a model with all covariates and removes predictors with the highest  $p$  – values greater than a pre-specified threshold. The model is re-estimated and the procedure ends when all  $p$  – values in the model do not exceed the significance level. There is no re-addition of previously dropped variables.

### **Forward selection**

The forward selection procedure starts a model without covariates except for the constant. A predictor with the lowest  $p$  – *value* below a pre-specified threshold is added. The procedure stops when none of the variables is significant at the threshold value.

### **Stepwise selection**

Stepwise selection integrates backward and forward selection. The addition or removal of variables at each step is on the basis of  $p$  – *values*. The procedure continues up until all the remaining variables are significant at the threshold value and every dropped variable is not significant or until variable to be added is same as last removed variable.

## **1.2.2 Penalised likelihood approaches**

These are also known as shrinkage methods. They use a technique that shrinks other regression coefficients towards zero by minimising the penalised likelihood. They include the ridge regression, Least Absolute Shrinkage and Selection Operation (LASSO) (Tibshirani, 1996), Smoothly Clipped Absolute Deviation (SCAD) (Fani and Li, 2001), Elastic Net (EN) (Zou and Hastie, 2005) and adaptive LASSO (Zou, 2006) selection methods.

## **1.2.3 Boosting**

Boosting is a model fitting and variable selection technique in settings of high dimension. It is a machine learning and optimisation procedure for fitting extensive regression models with various feasible forms of results and predictor effects (Mayr et. al., 2014a, b). It is an ensemble learning method wherein many models (weak

learners) are integrated to build a more robust model (strong learner). Models are fitted in series with the aim of reducing errors sequentially.

### 1.3 Statement of the Problem

Continuous time-to-event data has many variable selection procedures available (Groll and Tutz, 2014). However, the presence of ties and proportional hazards assumption pose a challenge to models for continuous time, for example, Cox model. Kalbfleisch and Prentice (2002) indicated that, the prevalence of ties gives rise to asymptotic bias in both the estimation of the regression coefficients and in the estimation of the corresponding covariance matrix. When the proportional hazards assumption is violated (might fail to hold since the effects of explanatory variables on the hazard may be intrinsically non-proportional), the estimated covariates effects are likely to be biased (Hess and Persson, 2010). In numerous applications time is quantified in days, weeks, months or years i.e. a discrete scale. Discrete models assume duration time is branded or grouped into discrete intervals.

In this study we aim to fit models on age at first marriage for women in Zimbabwe using Demographic Health Survey (DHS 2015-16) data. The DHS data set is based on a multistage stratified sampling. It consists of clustered observations. The data is clustered at individual, household and community level. Clustered observations usually imply responses not being independent. Thus for instance women belonging to the same family might tend to give approximately the same age at first marriage because of cultural or religious beliefs. The same effect can also be observed at community level due to exposure to the same job opportunities and or educational facilities. Therefore, ignoring the clustering in the data can result in biased estimates

of parameters.

Traditional variable selection methods (backward elimination, forward elimination, stepwise) are argued to have stability problems. They disregard the stochastic errors projected by the selection process and are mathematically impractical when the number of covariates gets large. Penalised variable selection shrinks some of the regression coefficients towards zero by minimising the penalised likelihood estimator. They also take into account the stochastic errors projected by selection process. Penalised and boosting methods are relatively new and not widely explored. This study explores the use of penalised and boosting variable selection techniques in a discrete survival model with unobserved heterogeneity.

## **1.4 Aims and Objectives**

### **1.4.1 Aim of the study**

To explore the use of lasso and gradient boosting variable selection methods.

### **1.4.2 Objectives**

The objectives of the study are to:

1. conduct a simulation study to evaluate and compare the effectiveness of lasso and gradient boosting variable selection methods.
2. apply lasso and gradient boosting variable selection techniques to build a model on the age at first marriage in Zimbabwe.
3. identify the determinants of age at first marriage for women in Zimbabwe.



## 1.5 Proposed methodology

A simulation study will be conducted to assess and compare the effectiveness of the methods. The methods are also going to be applied to data on age at first marriage for women in Zimbabwe.

## 1.6 Layout of the project

The entire dissertation is organised as follows. Chapter two presents the literature review. Chapter three discusses the methodology. Chapter four presents a simulation study and chapter five discusses application to data on age at first marriage in Zimbabwe. Lastly, chapter six presents conclusion, recommendations and limitations of the study.

# Chapter 2

## Literature Review

The chapter discusses related literature on lasso and boosting variable selection methods, unobserved heterogeneity and age at first marriage in Zimbabwe and elsewhere in the world.

### 2.1 Stepwise procedures

The stepwise approaches fit regression models. An automated procedure carries out the choice of a predictive variable. At every step, a variable is examined for inclusion or elimination from the set of explanatory variables on the basis of some predefined criterion. These criteria include adjusted  $R^2$ , Mallows'  $C_p$ , Bayesian Information Criterion (BIC), Akaike's Information Criterion (AIC) and p-values. Stepwise regression is a procedure in selecting a model developed by statisticians when model uncertainty was started to be taken into consideration. It is available in almost any statistical software. Stepwise regression with a range of criteria is found in the R package *leaps* (Lumley, 2017). There are three approaches to stepwise regression namely stepwise selection, backward elimination and forward selection.

With forward selection, the model commences without covariates. A univariate model for each variable is fitted. The variable having the largest partial F-statistic is entered into the base model. The previous strategy is redone up until no new variable may enter the model anymore. Forward selection results to minimal error on prediction and bias when compared to other regressions.

Backward elimination starts with every covariate included in the model. Sequentially, the least useful predictor (predictor having the highest  $p$  – value greater than a predefined threshold) is withdrawn from the model. The model is re-evaluated and the procedure stops when all  $p$  – values in the model are below the significance level. Variables eliminated cannot be re-added to the model.

Stepwise selection is an integration of forward and backward selection techniques. Variables are included or withdrawn at each step on the basis of the  $p$  – values. The procedure continues until every variable not included is not significant or up until a variable to be entered is same as previous deleted variable (Ekman, 2017).

An effective variable selection results in a better model prediction and interpretation. However, traditional stepwise procedures are argued to ignore the stochastic errors inherited in variable selection stages and lack stability (Breiman, 1996). The computations in stepwise procedures are infeasible when there is a large-sized number of covariates. Hurvich and Tsai (1990) as well as Steyerberg and Marius (1999) demonstrated biasness in estimation and inconsistent selection of stepwise regression. Some alternatives to traditional stepwise variable selection procedures are regularisation techniques which are penalty methods and boosting.

## 2.2 Penalised likelihood variable selection

Penalised regression methods are also known as shrinkage methods. The variables are selected by optimising a penalised likelihood. Lasso is the most well known penalised likelihood variable selection method. Lasso conducts variable selection and shrinkage simultaneously. With lasso, a weighted penalty is added to the log-likelihood function. The absolute values of all the variables are contained in the penalty. The lasso penalty enforces variable selection. All or limited coefficients of the variables are shrunk to zero. This depends on the size of the penalty. Small coefficients are shrunk to be precisely zeros. This results in a sparse solution (Friedman et. al., 2007).

In discrete survival modelling, estimation of parameters is complicated by the presence of the baseline hazard parameters as the estimates tend to be unstable even for small sample sizes. Stability of the baseline hazard estimates is crucial. To ensure that stability, a second penalty is added to the log-likelihood. The ridge type penalty is the most convenient approach when the baseline hazard is expressed as a sum of dummy variables representing time periods at which observations are made. However, if the baseline hazard is expressed as the summation of basis functions as in Efron, 1988; Fahrmeir, 1994; Möst et. al., 2016, a difference penalty can be used. The penalties ensure that parameters are estimated smoothly and give sufficient estimates of the baseline.

In R, there are packages that can be used for penalised variable selection in discrete survival models. Well known packages include *glmnet* (Friedman et. al., 2010), *penalised* (Goeman, 2010), *glmmlasso* (Groll and Tutz, 2017), *grplasso* (Meier et. al., 2008) and *grpreg* (Breheny and Jian, 2015). However, *glmnet*, *grplasso* and *grpreg*

cannot penalise the baseline parameters separately as the functionality is not available. For effective penalised likelihood variable selection, *glmmlasso* is chosen among the packages while *glmnet* and *penalised* by Goeman have shown weak performance as no random effects can be incorporated. For discrete survival models both with and without unobserved heterogeneity, the *glmmlasso* package has displayed high performance (Groll and Tutz, 2017).

## 2.3 Boosting

Boosting is a statistical variable selection as well as model fitting technique which originated from machine learning community designed as a tool to predict binary outcomes (Mayr et. al., 2014a, b). The *AdaBoost* algorithm was the first boosting method (Freund and Schapire, 1996). However, boosting was afterwards adapted into the area of statistical modelling. Boosting is now employed in the selection and estimation of the impact of covariates in a wider range of regression models. Like *lasso*, boosting is applicable and stable in large-sized settings where the number of covariates surpasses the number of observations ( $p > n$ ). In large-sized setting, most conventional estimation algorithms for regression setting fail. In addition, statistical boosting algorithms are very flexible in terms of the type of explanatory variables that enter into final model. Computerised variable selection together with model choice are included in the fitting procedure (Mayr et. al., 2014a, b). Boosting produces a prediction model by iteratively applying simple models (learners). The solutions of the models are combined to produce a much better prediction outcome. Prominent boosting algorithms are gradient algorithm (Friedman, 2001; Bühlmann and Hothorn, 2007) and likelihood-based boosting (Tutz and Binder, 2006). In gradient boosting,

errors are reduced successively as models are fitted in series. Like lasso's tuning parameter, the stopping iteration ( $m_{\text{stop}}$ ) is vital for variable selection in the boosting algorithm. It controls the shrinkage effects of estimates. Optimal  $m_{\text{stop}}$  values ensure early stopping that prevents overfitting and improves prediction accuracy. Cross validation is used to decide on the  $m_{\text{stop}}$  optimal value.

In R software, the gradient boosting algorithm is executed in the package *mboost* (Hothorn et. al., 2015). The *mboost* package provides a sizeable number of families and base learners which may be combined by the user resulting in a wide-range possibilities for nearly every statistical setting where regression models can be applied. We are not aware of any work where *mboost* and *glmmlasso* functions are compared.

## 2.4 Unobserved heterogeneity

In trying to capture disparity in the hazard rate across individuals in the population, predictors are included in a discrete survival model. In most instances, not all key variables are usually included. Some variables are not included since they are not in a data set. Unobserved heterogeneity is described as variation between individuals that is because of unmeasured characteristics. The excluded variables give rise to model misspecification. If unobserved heterogeneity is not accounted for, biased parameters are the result (Guo and Rodriguez, 1992). The discrete time logit model can be expanded to account for heterogeneity (Lewis and Raftery, 1999; Biggeri et. al., 2001; Manda and Meyer, 2005). As a remedy for unobserved heterogeneity, a random effect is included within the model. The random effect is well known as frailty. It is presumed to have a normal distribution. Frailty constitutes unobserved risk

factors that are particular to an individual and fixed overtime. The significance of the frailty term is tested by using the estimated variance of the frailty effect. A zero variance shows the independence of observations from the same household or community. However, a large variance shows large heterogeneity across a given household or community. This implies greater correlation among individuals from the same household or community.

## 2.5 Age at first marriage

In many communities the most vital life event for all men and women is marriage. It signifies the emergence to adulthood (Jensen and Thornton, 2003). It marks the transition to adulthood and beginning of even social allowable time for sensual activity and bearing of children (Palamuleni, 2011). The age at first marriage is defined as the age at which the respondent in this case, a woman began inhabiting with first husband or partner as a wife irrespective of the legality or otherwise of their union (ZIMSTATS, 2012).

The Zimbabwe Demographic Health Survey reported the median age at first marriage amongst women of 19.7 in 2011. It was amongst the highest in African states (ZIMSTATS, 2012 and Harwood-Lejuence, 2001). Most countries declared 18 years as the minimum legal age at which a woman can enter into marriage or union. In Sub-Saharan Africa, Zimbabwe is one of the countries with marriage age rising along with Kenya and Senegal (UNICEF, 2001). However, the marriage of children and adolescents under 18 years is yet a common phenomenon. Globally, about sixty millions of women aged between 20 and 24 years entered marriage before 18 years of age

(UNICEF, 2007). Conversely, women were married in their mid to late twenties prior to the Industrial Revolution (Gaskin, 1978 and Hajnal, 1953).

The important aspect of women's reproductive conduct with extensive consequences especially for the sexual health and social standing is the timing of the first marriage or union (Singh and Samara, 1996). It is a measure of vulnerability to the risk of pregnancy and childbearing (Blesoe and Cohen, 1993) and affects fertility level (McDevitt et. al., 1996) and population growth especially in countries with low contraceptive rate (Lesthaeghe et. al., 1989). Entry into motherhood lengthens the reproductive period thereby increasing fertility (Islam, 1999).

Women with low age at first marriage or union are possibly to perceive motherhood the only focal point of the woman's life at the expense of some areas of development for instance, education, job experience and self-development (Singh and Samara, 1996). It is also associated with early pregnancy and childbirth resulting to heightened chances of premature, labour complications during delivery and morbidity. Researchers have shown that children born to young mothers face low scores in achievement tests and high scores on the problem behaviour index (Hoffertn and Reid, 2002; Mirrowsky, 2005; Kamal, 2012).

Various studies show that the demographic, social and economic variables influence age at which a woman get married for the first time (Pamuleni, 2011; Garrence, 2004; Bracher and Santow, 1998; Axinn and Thornton, 1992). Amongst these variables are education, wealth status, region, place of dwelling, religion, ethnicity and work status. There is considerably minimal studies, if not none, looking at age at first marriage in



Zimbabwe. Many previous studies were focusing on fertility and contraceptive use. The research sought to ascertain determinants of age at first marriage at national level through penalised and boosting variable selections.

# Chapter 3

## Methodology

This chapter covers a more detailed theoretical background of the proposed methodology. Discrete time survival models, penalised and boosting estimation and model accuracy are discussed. In addition, given is a brief description of the real data the variable selection methods are to be applied on.

### 3.1 Basic discrete survival model

Let  $T$  be the a non-negative random variable ‘discrete time’ with feasible values  $T \in \{1, \dots, k\}$ . This implies  $T = t$  is observed if an event happens within the interval  $[a_{t-1}, a_t)$ . The discrete hazard function is defined by

$$\lambda(t|\mathbf{x}) = P(T = t|T \geq t, \mathbf{x}), t = 1, \dots, k, \quad (3.1.1)$$

where  $\mathbf{x}$  is a set of predictors. The hazard function is the conditional probability for failure in the interval  $[a_{t-1}, a_t)$  provided the interval is arrived (Tutz and Schmid, 2016). The corresponding survival function is defined by

$$S(t|\mathbf{x}) = P(T > t|\mathbf{x}) = \prod_{s=1}^t (1 - \lambda(s|\mathbf{x})). \quad (3.1.2)$$

It is the likelihood that an incident happens after a specified discrete time  $t$ .

The discrete survival model shows the hazard as a function of the covariates through an equation written below:

$$\lambda(t|x) = g(\gamma_{0t} + \mathbf{X}^T \gamma), \quad (3.1.3)$$

where  $g(\cdot)$  is a fixed response function. It connects the response probability with the linear predictor  $\gamma_{0t} + \mathbf{X}^T \gamma$ . It is presumed to be stringently increasing unconditionally. The parameter  $\gamma_{0t}$  represents the baseline hazard which captures the influence of time on the hazard. The vector of predictors is denoted by  $x^T$  and  $\gamma$  are the parameter estimates.

There are many discrete hazard models. A discrete hazard model depends on the chosen response function linking the response probability to the linear predictor which contains the effects of covariates. Below are some widely used discrete survival models.

## 3.2 Commonly used link functions

### 3.2.1 Logistic Discrete Hazard Model

It is a binate regression model. It uses the logistic distribution function obtained from the log-Burr distribution of the following form:

$$g(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)} \quad (3.2.1)$$

The logistic model is written in the form:

$$\lambda(t|\mathbf{x}) = \frac{\exp(\gamma_{0t} + \mathbf{X}^T \gamma)}{1 + \exp(\gamma_{0t} + \mathbf{X}^T \gamma)}. \quad (3.2.2)$$

Another expression for the above model is:

$$\log \left( \frac{P(T = t|\mathbf{x})}{P(T > t|\mathbf{x})} \right) = \gamma_{0t} + \mathbf{X}^T \gamma. \quad (3.2.3)$$

It makes a comparison between the likelihood of an occurrence at  $t$  and the likelihood of an occurrence after  $t$  (Tutz and Schmid, 2016).

### 3.2.2 Grouped Proportional Hazard Model

It is a grouped form of the Cox's proportional hazard model. It applies the Gompertz distribution of the form:

$$g(\eta) = 1 - \exp(-\exp(\eta)) \quad (3.2.4)$$

as the response function which is asymmetrical.

The Gompertz model has the form:

$$\lambda(t|\mathbf{x}) = 1 - \exp(-\exp(\gamma_{0t} + \mathbf{X}^T \gamma)) \quad (3.2.5)$$

or alternatively

$$\log(-\log(P(T > t|T \geq t))) = \gamma_{0t} + \mathbf{X}^T \gamma. \quad (3.2.6)$$

It is also called the *complementary log-log* ("clog-log") model. The hazards in various groups are assumed to be proportional.

### 3.2.3 Probit Model

It make use of the cumulative standard normal distribution  $\phi(\cdot)$ .

The Probit Model has the form:

$$\lambda(t|\mathbf{x}) = \phi(\gamma_{0t} + \mathbf{X}^T \gamma). \quad (3.2.7)$$

The probit model is most preferred in economic applications over logistic model in modelling binate responses. This study make use of the logistic link function. The logistic discrete hazard model is used on data application.

### 3.3 The baseline hazard

In the Model (3.1.3),  $\gamma_{0t}$  denotes the baseline hazard function which is the risk for a discrete when the entire covariates are set to zero. For  $q$  number of time points, the parameters of the baseline hazard function are given by  $\gamma_0(1), \gamma_0(2), \dots, \gamma_0(q)$ . Therefore, the baseline hazard is expressed as:

$$\gamma_{0t} = \gamma_0(1)D_1 + \gamma_0(2)D_2 + \dots, \gamma_0(q-1)D_{q-1}, \quad (3.3.1)$$

where  $\gamma_0(s)$  for  $s = 1, \dots, q-1$  being parameters to be estimated and  $D_i$  for  $s = 1, 2, \dots, q$  are fixed basis functions. If the number of time points is large, the number of parameters becomes larger as well (Tutz and Schmid, 2016). This might result in erroneous maximum likelihood estimates. In this study models which consider time as a smooth function are used.

The usual approach to fit a smooth function is to make an assumption that the function can be approximated by a finite summation of basis functions. Let the parameters  $\gamma_0(1), \gamma_0(2), \dots, \gamma_0(q)$  be approximated by

$$\gamma_{0t} = \sum_{s=1}^m \gamma_{0s} \phi_s(t), \quad (3.3.2)$$

where  $\phi_s(\cdot)$  are basis functions being fixed. Usual options are polynomial splines in the form of the truncated power series and B-splines (Tutz and Schmid, 2016).

## 3.4 Penalised estimation and variable selection

### 3.4.1 Models without unobserved heterogeneity

Considering the model of the form in equation (3.1.3), the unknown parameter estimates can be attained by the maximum likelihood method. However, observations in survival analysis are subjected to censoring. Right censoring is considered in this study. Suppose  $C_i$  represents the censoring time and  $T_i$  the exact failure time of the observation  $i$ .  $T_i$  and  $C_i$  are assumed independent. The observed time is given by  $t_i = \min(T_i; C_i)$  as the minimal of survival time  $T_i$  and censoring time  $C_i$  (Groll and Tutz, 2016). The introduction of an indicator variable for censoring is vital. It is given by :

$$\delta_i = \begin{cases} 1 & \text{if } T_i \leq C_i \\ 0 & \text{if } T_i > C_i \end{cases}$$

where the assumption is that the censoring happens at the end of the interval (Tutz and Schmid, 2016). The chances of observing the exact survival time ( $t_i, \delta_i = 1$ ) is given by

$$P(T_i = t_i, \delta_i = 1) = P(T_i = t_i)P(C_i \geq t_i). \quad (3.4.1)$$

The possibilities of observing censoring at time  $t_i$  is defined as

$$P(C_i = t_i, \delta_i = 0) = P(T_i > t_i)P(C_i = t_i). \quad (3.4.2)$$

Under *random censoring* (the assumption that the random variables  $T_i$  and  $C_i$  are independent and that the finite sequence  $(T_i, C_i), i = 1, \dots, n$  are measured independently), combining probability equation (3.4.1) and equation (3.4.2) leads to the likelihood contribution or the probability of observing  $(t_i, \delta_i)$  given by

$$P(t_i, \delta_i | x_{it}) = P(T_i = t_i)^{\delta_i} P(T_i > t_i)^{1-\delta_i} P(C_i \geq t_i)^{\delta_i} P(C_i = t_i)^{1-\delta_i}. \quad (3.4.3)$$

Let the contribution of the censoring distribution be  $c_i$  defined by

$$c_i = P(C_i \geq t_i)^{\delta_i} P(C_i = t_i)^{1-\delta_i}. \quad (3.4.4)$$

If the assumption is that  $c_i$  is independent of the variables determining the survival time, meaning uninformative of the censoring mechanism (Kalbfleisch and Prentice, 2002) then the minimised likelihood function is of the form

$$P(t_i, \delta_i | x_{it}) = c_i P(T_i = t_i)^{\delta_i} P(T_i > t_i)^{1-\delta_i}. \quad (3.4.5)$$

Sequences of binary data are used in re-writing the probability and the corresponding likelihood in discrete survival modelling (Groll and Tutz, 2016). If the sequence for a non-censored observation ( $\delta_i = 1$ ) is defined by  $(y_{i1}, \dots, y_{ti}) = (0, \dots, 0, 1)$  and  $(y_{i1}, \dots, y_{ti}) = (0, \dots, 0)$  for a censored ( $\delta_i = 0$ ), the likelihood (including  $c_i$ ) can be presented as

$$P(t_i, \delta_i | x_{it}) = c_i \prod_{s=1}^{t_i} \lambda(s | x_{it})^{y_{is}} (1 - \lambda(s | x_{it}))^{1-y_{is}}. \quad (3.4.6)$$

For a discrete survival-time model, the log-likelihood is of the form:

$$l = \sum_{i=1}^n \sum_{s=1}^{t_i} (y_{is} \log \lambda(s | x_i)) + (1 - y_{is}) \log(1 - \lambda(s | x_i)). \quad (3.4.7)$$

Selection of variables can be achieved by penalising the log-likelihood. The most famous implementation of this is the lasso where a penalty of the form  $\nu \sum_{i=1}^p |\gamma_i|$  is added to equation (3.4.7). Another second penalty is added to stabilize baseline hazard estimates resulting in a penalised likelihood of the form:

$$l = \sum_{i=1}^n \sum_{s=1}^{t_i} (y_{is} \log \lambda(s | x_i)) + (1 - y_{is}) \log(1 - \lambda(s | x_i)) - \nu \sum_{i=1}^p |\gamma_i| - \gamma_s \text{Pen}(\gamma_0). \quad (3.4.8)$$

The penalty term puts restriction on  $\nu$  that enforces variable selection and relies on a scalar *tuning parameter*,  $\nu \geq 0$ . The strength of penalisation is controlled by the size of  $\nu$ . For  $\nu=0$ , no penalty, the ordinary maximum likelihood is obtained implying no variable selection. All variables are permitted into the model. However, as  $\nu \rightarrow \infty$ , all predictors are removed from the model.

One possibility for a penalty for the baseline hazard function is a ridge penalty of the form:

$$\text{Pen}(\gamma_0) = \sum_{t=1}^q \gamma_{0t}^2. \quad (3.4.9)$$

It is used if the baseline is expressed as a sum of dummy variables representing time periods at which observations are made.

Alternatively, if the baseline is expressed as the summation of basis functions as in Efron, 1988; Fahrmeir, 1994; Möst et. al., 2016, a difference penalty can be used. In this study use was made of B-splines of order three and difference penalty of the form:

$$\text{Pen}(\gamma_0) = \sum_{j=r+1}^m (\Delta^r \gamma_{0j})^2, \quad (3.4.10)$$

where  $\Delta$  is the difference operator, working on adjacent B-splines coefficients and  $r$  is the order of  $\Delta$ . This penalty ensures smooth parameter estimation with the level of smoothness set on the tuning parameter  $\gamma_s$ .

AIC, BIC and cross-validation are used to ascertain the tuning parameters  $\nu$  and  $\gamma_s$ . BIC uses all data to fit a model and places a heavier penalty than the AIC statistic on models with many covariates. This results in models with few covariates. However, through simulations, Groll and Tutz (2017) established that it is not necessary to select both parameters using either information criteria or cross-validation. They recommended carefully selection of  $\nu$  and an optimal value of  $\gamma_s$  to be used. In this



study, BIC was used as criterion for  $\nu$  and the value of  $\gamma_s$  was set at 10. The process involved creating a vector of possible values of  $\nu$ , fitting a model using each value and noting the value that gives the minimum value of the BIC. The parameter estimates of the prior fit were used as starting values for the next fit to enhance fast convergence at each value of  $\nu$ . It is essential to make sure that the  $\nu$  sequence starts at the value big enough such that all covariates are shrunk to zero.

### 3.4.2 Models with unobserved heterogeneity

Model (3.1.3) ignores heterogeneity among individuals that can lead to biased parameter estimates. In a discrete-time model, frailty is integrated by adding a random effect with a normal distribution to the linear predictor. The basic frailty model for the  $i^{th}$  observation at the time  $t$  has the form:

$$\lambda(t|\mathbf{x}_{it}, z_{it}, \mathbf{b}_i) = g(\gamma_{0t} + x_{it}^T \gamma + z_{it}^T b_i) \quad (3.4.11)$$

with explanatory variables  $x_{it}, z_{it}$ , that may change after a while, and random effect vector  $b_i$ . The frailties are usually presumed to follow a normal distribution with mean zero and variance  $\sigma^2$ . The variance for the random effect is the variance between individuals that is attributable to time-invariant characteristics not observed i.e. residual variance.

In a random effect model, the marginal probability is given by

$$P(t_i, \delta_i | x_{it}, z_{it}) = \int P(t_i, \delta_i | x_{it}, z_{it}, b_i) p(b_i) db_i \quad (3.4.12)$$

equivalent to

$$P(t_i, \delta_i | x_{it}, z_{it}) = \int P(T_i = t_i)^{\delta_i} \cdot P(T_i > t_i)^{1-\delta_i} p(b_i) db_i. \quad (3.4.13)$$

expressed as

$$P(t_i, \delta_i | x_{it}, z_{it}) = \int \prod_{s=1}^{t_i} \lambda(s | x_{it}, z_{it}, b_i)^{y_{is}} (1 - \lambda(s | x_{it}, z_{it}, b_i))^{1-y_{is}} p(b_i) db_i \quad (3.4.14)$$

the unconditional probability of a random effects model (frailty model) for well structured binary data.

For the distribution  $\rho(\cdot)$  of the random effect, it is presumed that  $b_i \sim N(0, Q(\rho))$  where  $\rho$  is a vector of parameters not known that determines the baseline hazard collected in  $\gamma_0^T = (\gamma_{01}, \dots, \gamma_{0k})$ . By maximising the marginal log-likelihood, the parameters are estimated. The marginal log-likelihood is as follows:

$$l(\gamma_0, \gamma, \rho) = \sum_{i=1}^n \log \left( \int \prod_{s=1}^{t_i} \lambda(s | x_{it}, z_{it}, b_i)^{y_{is}} (1 - \lambda(s | x_{it}, z_{it}, b_i))^{1-y_{is}} p(b_i) db_i \right). \quad (3.4.15)$$

Representing a discrete survival model by a binary regression model whose log-likelihood is in a more simplified form, the marginal log-likelihood will be of the form:

$$l(\gamma_0, \gamma, \rho) = \sum_{i=1}^n \log \left( \int f(y_i | \gamma_0, \gamma, \rho) p(b_i) db_i \right). \quad (3.4.16)$$

The random effects model is estimated using Gauss-Hermite integration (Hinde, 1982; Anderson and Aitkin, 1985). Gaussian quadrature can be used to approximate the marginal likelihood by numerical integration. Gaussian-Hermite quadrature is a form of Gaussian quadrature for approximating the value of integrals of the nature  $\int_{-\infty}^{+\infty} e^{-x^2} dx$ .

In this case

$$\int_{-\infty}^{+\infty} e^{-x^2} f(x) dx \simeq \sum_{i=1}^p a_i f(x_i), \quad (3.4.17)$$

where  $p$  is the number of sample points used. The  $x_i$  are the roots of the Hermite polynomial  $H_p(x)(i = 1, 2, \dots, p)$ , and the associated weights  $a_i$  are given by

$$a_i = \frac{2^{p-1} p! \sqrt{\pi}}{p^2 [H_{p-1}(x_i)]^2}. \quad (3.4.18)$$

Following ideas in Groll and Tutz (2017) it can be shown that  $l(\gamma_0, \gamma, \rho)$  can be presented as

$$l(\gamma_0, \gamma, \rho, b) = \sum_{i=1}^n \log(f(y_i | \gamma_0, \gamma, \rho) - \frac{1}{2} \mathbf{b}^T \mathbf{Q}_b(\rho)^{-1} \mathbf{b}), \quad (3.4.19)$$

where  $\mathbf{b}^T = (\mathbf{b}_1^T, \dots, \mathbf{b}_n^T)$  collects the random effects and  $\mathbf{Q}_b$  is the diagonal covariance matrix.

To incorporate penalisation in discrete survival as in equation (3.4.8), two penalty terms are added to equation (3.4.19) resulting in a penalised likelihood of the form:

$$l(\gamma_0, \gamma, \rho, b) = \sum_{i=1}^n \log(f(y_i | \gamma_0, \gamma, \rho)) - \frac{1}{2} b^T Q_b(p)^{-1} b - \gamma \sum_{i=1}^p |\gamma_i| - \gamma_s \text{Pen}(\gamma_0). \quad (3.4.20)$$

Before estimating the likelihood functions in equation (3.4.8) and equation (3.4.20) appropriate design matrices must be constructed. To transform discrete survival data into binary representation, the *dataLong* function is used. The *dataLong* function is found on the R add-on package *discSurv*.

For a subject who experiences the outcome of an event (non-censored),  $\delta = 1$  at time  $t_i$ , the augmented data matrix is given by:

$$\begin{bmatrix} \text{Binary observation} & \text{Dummies} & \text{Covariates} & & \\ & & & & \\ & 0 & 1 & x_{i1} & \cdots & x_{ip} \\ & 0 & 2 & x_{i1} & \cdots & x_{ip} \\ & \vdots & \vdots & \vdots & \vdots & \vdots \\ & 1 & t_i & x_{i1} & \cdots & x_{ip} \end{bmatrix}$$

Similarly, when the individual has been censored ( $\delta = 0$ ) the augmented matrix is of the form:

$$\begin{bmatrix} \textit{Binary observation} & \textit{Dummies} & \textit{Covariates} & & \\ 0 & 1 & x_{i1} & \cdots & x_{ip} \\ 0 & 2 & x_{i1} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & t_i & x_{i1} & \cdots & x_{ip} \end{bmatrix}$$

The binary observation column represents the binate responses  $y_{i1}, \dots, y_{it}$  while the dummies column is the time interval ranging from 1 to  $t_i$ . For fixed parameters for  $\gamma_{0t}$ , the column is a nominal variable such as ‘dummy variables’. The remaining columns are the effects of the explanatory variables.

### 3.4.3 Software packages for penalised regression

The *glmmlasso* package in **R** was selected to estimate models (3.4.8) and (3.4.20) based on high performance as demonstrated in Groll and Tutz (2017).

## 3.5 Boosting estimation

Gradient boosting starts with the response variable  $y$ , a set of covariates  $\mathbf{x}$  and a loss function  $\rho(y, g(x))$ . The loss function measures the deviation of the prediction function from the outcome. In Gaussian regression, the loss function is defined as:

$$\rho(y, g(x)) = (y - g(x))^2 \quad (3.5.1)$$

i.e. the squared error loss.

The loss function is supposed to be differentiable with respect to the prediction function  $g(x)$  i.e.

$$\frac{\partial \rho}{\partial g} = -2(y - g(x)). \quad (3.5.2)$$

Gradient boosting aims at modelling the link between  $y$  and  $\mathbf{x} := (x_1, \dots, x_p)^T$  and to get the optimal prediction of  $y$  given  $\mathbf{x}$ . This is achieved by minimising a loss function  $\rho(y, g(x)) \in \mathbb{R}$  over a prediction function  $g(x)$ .

The optimal prediction function  $g^*$  is defined by:

$$g^* = \operatorname{argmin}_g \mathbb{E}_{Y, X} [\rho(y, g(x^T))]. \quad (3.5.3)$$

In practice, dealing with realisations  $(y_i, \mathbf{x}_i^T)$ ,  $i = 1, \dots, n$  of  $(y, \mathbf{x}^T)$  and the expectation given above is therefore unknown. As a result, instead of minimising the expected value, boosting algorithms minimise the observed mean such that:

$$\mathcal{R} := \frac{1}{n} \sum_{i=1}^n \rho(y_i, g(x_i)) \quad (3.5.4)$$

called the *empirical risk* (Hofner et. al., 2014).

Consider the prediction function equation (3.1.3). The empirical risk  $\mathcal{R}$  in equation (3.5.4) can be taken into consideration for the estimation of  $g^*$  with boosting.

Gradient boosting algorithm is given as follows:

1. Initialise  $\hat{g}^{[0]}$  with an offset value. Usual options are

$$\hat{g}^{[0]}(.) \equiv \operatorname{arg}_c \min \frac{1}{n} \sum_{i=1}^n \rho(Y_i, c) \quad (3.5.5)$$

or  $\hat{g}^{[0]}(.) \equiv 0$ . Set the iteration counter  $m$  to zero i.e  $m=0$

2. Augment  $m$  by 1. Determine the negative gradient  $-\frac{\partial}{\partial g} \rho(Y, g)$  and evaluate at  $\hat{g}^{[m-1]}(X_i)$ :

$$U_i = -\frac{\partial}{\partial g} \rho(Y_i, g) \Big|_{g=\hat{g}^{[m-1]}(X_i)} \quad (3.5.6)$$

for  $i= 1, . . . , n$

3. Fit the negative gradient vector  $U_1, . . . , U_n$  to  $X_1, . . . , X_n$  by the real-valued base procedure (For example, regression)

$$(X_i, U_i) \rightarrow \hat{f}^{[m]}(.) .$$

Thus,  $\hat{f}^{[m]}(.)$  can be regarded as an approximation of the negative gradient vector.

4. Update  $\hat{g}^{[m]}(.) = \hat{g}^{[m-1]}(.) + \nu \cdot \hat{f}^{[m]}(.)$  where  $\nu \in (0,1]$  is a step-length factor, that is, proceed along an estimate of the negative gradient vector.

5. Iterate steps 2 to 4 until  $m = m_{stop}$  for some stopping iteration  $m_{stop}$ .

The principal tuning parameter  $m_{stop}$  can be set on by cross-validation or AIC-based techniques. Boosting algorithms must not run until convergence. Large stopping iteration, overfitting resulting in a suboptimal out-of prediction accuracy are likely to happen (Bühlmann and Hothorn, 2007). Large stopping iteration  $m_{stop}$  leads to a high number on non-informative covariates incorporated in the estimation of  $g^*$  hence the algorithm should be stopped early (before convergence) and to select  $m_{stop}$  thus out-of sample prediction accuracy becomes optimal. However, choosing the step length factor  $\nu$  has proven to be of less significant for the predictive performance of a boosting algorithm provided it is “minimal” (For example,  $\nu=0.1$ , see Schmid and Hothorn , 2008). Small step length warrants that estimated effect increase steadily in the course of the boosting algorithm, and the estimates stop increasing once the

optimal stopping iteration  $m_{stop}$  is arrived at. It ensures that the algorithm will not exceed the minimal of the empirical risk  $\mathbb{R}$ .

Having gradient boosting algorithm, the most appropriate discrete hazard model loss function has to be specified. The empirical risk,  $\mathbb{R}$ , is equated to the negative binomial log-likelihood function (Tutz and Schmid, 2016) such that

$$\mathbb{R} = - \sum_{i=1}^n \sum_{s=1}^{t_i} y_{is} \log \lambda(s|x_i) + (1 - y_{is}) \log(1 - \lambda(s|x_i)), \quad (3.5.7)$$

where  $y_{is}$  codes the move to the next period. The linear predictor  $f(\mathbf{x}_{it}) = \mathbf{x}_{it}^T \beta$  with  $\mathbf{x}_{it}^T = (0, \dots, 0, 1, \dots, 0, \mathbf{x}_i^T)$  and  $\beta^T = (\gamma_{01}, \gamma_{02}, \dots, \gamma_{0q}, \gamma^T)$  are used. Finally, the boosting algorithm is applied to the augmented data. The boosting algorithm can be implemented using *gamboost* function from the package *mboost*. The function is flexible and powerful in fitting linear or non-linear additive models through component-wise boosting (Hofner et. al., 2014). The package *mboost* can handle random effects as random effects base-learners are specified with a call to *brandom*. The *cvrisk* function is used to determine optimal stopping iteration.

## 3.6 Inferences on predictors

Considering the discrete survival model in equation (3.1.3), the hypothesis about  $\gamma$ , the effects of explanatory variables on survival or not, Wald's, the likelihood ratio and score tests are applied. These test the null hypothesis,  $H_0: \gamma_i = 0$  versus alternative,  $H_1: \gamma_i \neq 0$ .

### 3.6.1 Wald's test

The test statistic is given by:

$$w = \hat{\gamma}^T I(\hat{\gamma})^{-1} \hat{\gamma}. \quad (3.6.1)$$

### 3.6.2 Likelihood ratio test

The test statistic used in the likelihood ratio test is given by:

$$L_r = -2\{l(\tilde{\gamma}) - l(\bar{\gamma})\} \quad (3.6.2)$$

$L_r$  follows a  $\chi^2$ - distribution.

### 3.6.3 Score test

The score statistic has the form

$$u = s^T(\tilde{\gamma}) F^{-1}(\tilde{\gamma}) s(\tilde{\gamma}), \quad (3.6.3)$$

where  $s(\tilde{\gamma})$  being the score function and  $F^{-1}$  denoting the Fisher information matrix.

The score statistic like the likelihood ratio and Wald statistics, also follows a  $\chi^2$ - distribution.

## 3.7 Goodness-of fit of the model

To assess the goodness of fit of a model, several methods are employed. The methods measure how well fitted failure probabilities match with the corresponding observed proportions.

For models not nested, the AIC and BIC aim to penalise the log-likelihood for the



number of explanatory variables in the model. For a model with  $q$  parameters, the AIC has the form:

$$AIC = -2\log(L_m) + 2q, \quad (3.7.1)$$

where  $L_m$  is the maximum log-likelihood.

The BIC is given by

$$BIC = -2\log(L_m) + q\log(n), \quad (3.7.2)$$

with sample size  $n$ .

Small values of the criteria (AIC or BIC) indicate a better fitted model i.e. the model that has the minimal AIC or BIC is chosen as the best model.

Alternatively, model fitness can be assessed by using martingale residuals. Martingale residuals take into account censoring and make a comparison between the observed number of events up to time  $t_i$  for each individual and the expected number of events up to  $t_i$  (Tutz and Schmid, 2016). Martingale residuals are given by

$$r_i = \sum_{s=j}^{t_i} (y_{ij} - \hat{\lambda}_{ij}), i = 1, \dots, n \quad (3.7.3)$$

for binary variables with  $(y_{i1}, \dots, y_{it}) = (0, \dots, \delta_i)$ .

If the martingale residuals are random and there is no correlation with the covariate values then model fits well.

### 3.8 Model Prediction Accuracy

It is vital to evaluate the performance of estimated survival models. Below are ways to quantify the predictive performance of survival models. For all the approaches, the assumption is that training data  $(t_i, \delta_i, x_i)$ ,  $i = 1, \dots, n$  is used to fit the model

and test data set  $(t_i^\tau, \delta_i^\tau, x_i^\tau)$ ,  $i = j, \dots, n^\tau$  used to assess predictive performance of the model. The learning data and test data should have the same distribution.

### 3.8.1 Predictive Deviance

Predictive deviance is a predictive accuracy measure based on likelihood and evaluated on the test data. For discrete survival times, it has the form:

$$D = -2 \sum_{i=1}^{n^\tau} (\delta_i^\tau \log(\hat{P}^L(T_i^\tau = t_i^\tau)) + (1 - \delta_i^\tau) \log(\hat{P}^L(T_i^\tau > t_i^\tau))) \quad (3.8.1)$$

$$= -2 \sum_{i=1}^{n^\tau} \sum_{t=1}^{t_i^\tau} ((y_{it}^\tau \log(\hat{\lambda}^L(t|x_i^\tau)) + (1 - y_{it}^\tau) \log(1 - \hat{\lambda}^L(t|x_i^\tau))), \quad (3.8.2)$$

where  $y_{i1}, \dots, y_{it}$  is the binary transitions over time periods and  $\hat{\lambda}^L$  shows the fitting of the hazard rate using learning data and evaluation based on test data (Tutz and Schmid, 2016). Predictive deviance is equated to the negative log-likelihood of a binomial regression model evaluated on the test data. A model has a large predictive performance if the given deviance is small.

### 3.8.2 Concordance Index

The concordance index is a method to evaluate prediction accuracy based on discrimination measures. The measures uses the predictor  $\eta_{it} = x_i^T \gamma$ . In discrimination measures, responses are considered as time-dependent binary outcomes with classes namely event at  $t$  (cases) and event after  $t$  (controls). Prediction performance is high if  $\eta_{it}$  has a high discriminative power i.e. correctly specifying cases in the test data. Concordance Index is a popular discrimination measure used with its roots in Receiver Operating Characteristics (ROC) methodology.

For binary responses, the discriminative power can be summarised through time-dependent sensitivity and the time-dependent specificity (Heagerty and Zheng, 2005) defined by

$$sens(c, t) = P(\eta_{it} > c | T = t) \quad (3.8.3)$$

and

$$spec(c, t) = P(\eta_{it} \leq c | T > t) \quad (3.8.4)$$

where  $c$  is a threshold of the linear predictor  $\eta_{it}$  (Tutz and Schmid, 2006).

A time-dependent ROC curve results from summarising sensitivity and specificity. The ROC curve is given by:

$$ROC(c, t) = \{1 - spec(c, t), sens(c, t)\} c \in \mathbb{R} \quad (3.8.5)$$

The ROC curve plots sensitivity against false positive rates and usually concave in shape. If the ROC curve is closely concave and the area under it is large then the linear predictor  $\eta_{it}$  has a high discriminative power.

Computing the area under the ROC which for each time point  $t$  yields a time dependent Area Under the Curve (AUC) curve. Time-dependent AUC curve measures the discriminative power of a  $\eta_{it}$  at each time under consideration. Values of the AUC curve should be greater than 0.5

Uno et. al. (2007) and Schmid et. al. (2014) stated that sensitivity and specificity can be estimated by

$$sens(c, t) = \frac{\sum_{i=1} \delta_i^T I(\hat{\eta}_{it}^T > c \cap t_i^T = t) / \hat{G}^L(t_i^T - 1 | \mathbf{x}_i^T)}{\sum_{i=1} \delta_i^T I(t_i^T = t) / \hat{G}^L(t_i^T - 1 | \mathbf{x}_i^T)} \quad (3.8.6)$$

$$spec(c, t) = \frac{\sum_{i=1} I(\hat{\eta}_{it}^T > c \cap t_i^T > t)}{\sum_{i=1} I(t_i^T > t)} \quad (3.8.7)$$

where  $\hat{\eta}_{it}^{\tau}$ ,  $i= 1, \dots, n^{\tau}$  is the estimated linear predictor using observations from the test data while estimated parameters are from the learning data.

Numerical integration of the estimated ROC curve results to estimates of  $AUC(t)$ . However, by means of the time-dependent estimated area under the curve  $\widehat{AUC}(t)$ , the concordance index for discrete is given by

$$C^* = \sum_t \frac{\hat{P}^L(T = t) \cdot \hat{P}^L(T > t)}{\sum_t \hat{P}^L(T = t) \cdot \hat{P}^L(T > t)} \widehat{AUC}(t) \quad (3.8.8)$$

Concordance Index computes the likelihood that observations with large values of the linear predictor have shorter survival times than the observations with small values of  $n$ . A Concordance Index of one ( $C^*=1$ ) shows a perfect discriminating linear predictor.

## 3.9 Application: Age at first marriage

### 3.9.1 Data description

The study uses data drawn from the Zimbabwe Demographic Healthy (ZDHS) conducted in 2015-16. This includes data on adult and child health, mortality, fertility, marriage, contraceptive use and maternal health (ZIMSTATS and ICF International, 2012). The 2015-16 ZDHS is a nationwide representative survey of 9955 women aged 15-49 and men aged 15-54. The three questionnaires used in the survey are the household questionnaire, man's questionnaire and woman's questionnaire. The 9955 women interviewed from the 9 provinces is the sample in this study. Cases with missing data were eliminated and a sample of size 2000 to ensure timely convergence as in model building. Glmmlasso had convergence problems on a sample of more than 2 000 observations. These women were either married or ever-lived with a man as a

wife irrespective of the legality or otherwise of their union. A woman's age, in years at the time she started to live with a man is her age at first marriage. Women who had never been in a union or lived with a man were censored at their age on the date of the interview single hence no information given on their age at first marriage.

### 3.9.2 Model building

The response variable in the study is age at first marriage. It is defined as the age in years at which the respondent started to live with a man.

The explanatory variables considered in the study are given in Table 3.1 below.

Table 3.1: Table of names and descriptions of variables used in the study

Variable	Description
Place of residence	The type of place where the respondent was interviewed; categorical: Rural or Urban, Urban (reference)
Employment status	Whether the respondent is currently working or not; categorical: Yes or No, No (reference)
Region	The province in which the participant was interviewed; categorical: Mashonaland, Manicaland, Matebeleland, Masvingo, Midlands, Manicaland (reference)
Religion	The religious belief of the respondent; categorical: Christian, Muslim, Traditional, Apostolic faith, None, Apostolic faith (reference)
Birth Cohort	The year interval the respondent was born; categorical: <1970, 1970 - 1979, 1980 - 1990, >1990, <1970 (reference)
Education	Highest educational level attended by the respondent; categorical: No education, Primary, Secondary, Higher, No education (reference)
Cluster	The cluster number identifying the sample point; categorical

Having the response variable and covariates, two discrete survival models are fitted, one without random effect and the other one with random effect. The models are of the form:

$$\begin{aligned} \lambda(t|\mathbf{x}_{it}, \mathbf{b}_i) = & h(\gamma_{0t} + \gamma_1 Place + \gamma_2 Education + \gamma_3 BirthCohort \\ & + \gamma_4 Region + \gamma_5 Religion + \gamma_6 EmploymentStatus + b_i), \end{aligned} \quad (3.9.1)$$

where  $h(\cdot)$  is the logistic link function,  $\gamma_{is}$  for  $i = 1, \dots, 6$  the weight of the covariates place, education, birth cohort, region, religion and employment status and  $b_i$ , the random effect. This study models age at first marriage using community random effect on the hazard model, which permits the dependence observations in the same community into the model.

A simulation study is to be conducted to evaluate and compare the performance of lasso and gradient boosting with and without unobserved heterogeneity. More details are given in the next chapter.

# Chapter 4

## Simulation study

To be able to study how well the two variable selection methods perform, we will use simulated data. This chapter presents simulation set up employed and results obtained from the simulation exercise.

### 4.1 Simulation approach

This section describes how data was simulated and measures applied to evaluate the results from the simulations.

The simulation study was carried out to compare the performance between penalised and boosting variable selection methods. The data generating process used here was adopted from the simulation schemes employed by Groll and Tutz (2014). Data was simulated from a logistic hazard function of the form:

$$\lambda(t|\mathbf{x}_i, b_i) = \frac{\exp(\gamma_{0t} + x_1 + x_2 + \dots + x_p + b_i)}{1 + \exp(\gamma_{0t} + x_1 + x_2 + \dots + x_p + b_i)} \quad (4.1.1)$$

with  $p = \{15, 50, 100\}$

The time-varying baseline hazard is given by  $\gamma_{0t} = 2\xi_t - 2.3$  where  $\xi_t = f_{\Gamma}(t-2)$  where  $f_{\Gamma}(t)$  being a Gamma distribution density  $\Gamma(\zeta, \theta)$  with shape parameter  $\zeta = 5$  and

scale parameter  $\theta = 1$ . This results in a baseline function with a moderate hump, as shown in Figure 4.1 below. B-splines approach is used for  $\gamma_{0t}$ . Right-censoring was considered for all duration above 12.

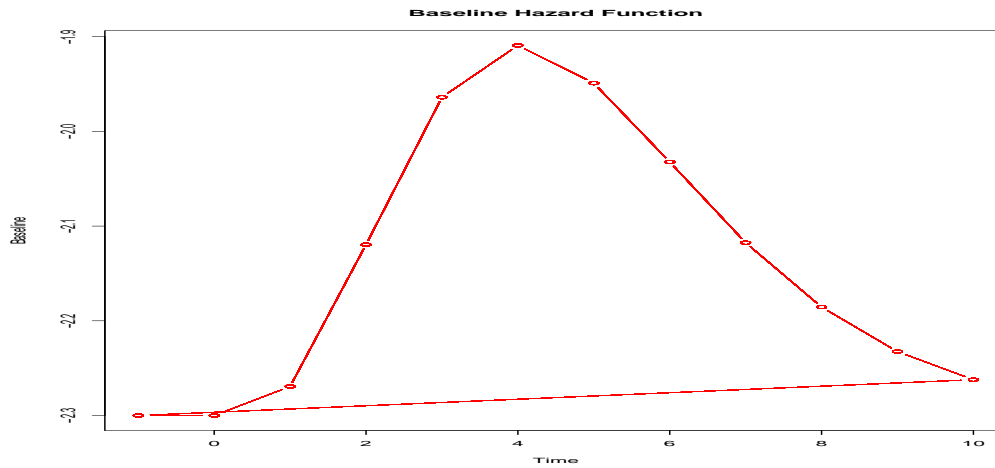


Figure 4.1: Discrete baseline hazard function

The settings of 100 observations ( $n = 100$ ) for  $p$  covariates,  $X_{i1}, X_{i2}, \dots$ , and  $X_{ip}$  are drawn independently and follow a uniform distribution within the interval  $[0, 1]$  i.e.  $X_{i2}, \dots, X_{ip} \sim U(0, 1)$ . The linear effects  $\gamma_1, \dots, \gamma_5$  are non-zero while  $\gamma_j = 0$  for  $j = 6, \dots, p$ . The random effects  $b_i$  are simulated from a normal distribution i.e.  $b_i \sim N(0, \sigma_b^2)$  with  $\sigma_b = \{0, 1, 2\}$ . Censoring times are simulated from the binomial distribution i.e.  $C \sim \text{bin}(1, \pi_{CN})$  where  $\pi_{CN} = 0.05$ . Binary response is simulated from the binomial distribution i.e.  $y_{it} \sim \text{bin}(1, \lambda(t | x_i))$  where  $\lambda(t | x_i)$  is the hazard rate and  $t = \{1, 2, \dots, 12\}$ . The values "0" and "1" are used for  $y_{it}$ ,  $y_{it} = 1$  showed that the subject experienced the event. Logistic link is used and the simulation scheme is replicated 100 times.



On *lasso*, finding a good value for the tuning parameter ( $\lambda$ ) is the main concern. This study used the BIC as a criterion for choosing the optimal tuning parameter. The method involves sequentially fitting models each with a different value of the tuning parameter value and selecting the parameter value with the lowest BIC. To ensure fast convergence, the parameter estimation of the previous iteration was used as starting values on each iteration. A  $\lambda$  sequence starting at a big value ensures that all covariates are shrunk to zero. To ensure smooth estimates of the baseline hazard, an additional penalty,  $\nu_s$  is included. In this study  $\lambda$  sequence has started at 40 and additional penalty value of 10 used (Groll and Tutz, 2014). The cluster number (ID) was used as the random effect.

The distinct variable selection methods' performance will be evaluated using a number of different measures. Performance of estimators was assessed for both the structural components and the variance. After the estimation of the coefficients  $\gamma_j$ , the results are compared to the true parameter. For every simulation run, mean squared error ( $MSE_\gamma$ ) is calculated where;

$$MSE_\gamma = \|\gamma - \hat{\gamma}\|^2 \quad (4.1.2)$$

False positive rates (FPR) and False Negative Rates (FNR) are also accounted for each run. False positive denotes a single parameter value that is exactly zero is set to not zero . Conversely, false negative means that a single non-zero parameter is set to zero.

$$FPR = \frac{\text{Number of exactly zero set to not-zero}}{\text{Number of exactly zero}} \quad (4.1.3)$$

$$FNR = \frac{\text{Number of not-zero set to zero}}{\text{Number of truly not zero}} \quad (4.1.4)$$

The average of the mean squared errors, false positive and false negative are computed and presented in Table 4.1 and 4.2 below.

## 4.2 Simulation Results

Table 4.1: Results for  $MSE_\gamma$  for *glmLasso* and *boosting*,  $\pi_{\text{censor}}=0.05$  and  $n = 100$

$\sigma_b$	$p$	<i>glmLasso</i>	<i>glmLasso</i>	<i>boosting with-</i> <i>out random</i> <i>effect</i>	<i>boosting with</i> <i>random effect</i>
0	15	4.72	8.64	44.16	81.32
0	50	6.76	9.88	50.56	81.34
0	100	10.18	14.84	58.69	91.4
1	15	13.81	8.47	85.11	52.08
1	50	17.06	9.81	85.21	55.83
1	100	35.44	34.06	86.02	61.53
2	15	26.87	21.95	86.21	70.6
2	50	37.88	36.74	90.9	70.28
2	100	41.89	40.21	94.8	78.69

The summary of the mean square error referring to covariates ( $MSE_\gamma$ ) are shown in Table 4.1. For all values of  $\sigma_b = \{0, 1, 2\}$ , Lasso methods outperform boosting. Lasso methods with and without random effect, i.e. *glmLasso* and *glmLasso* respectively have smaller mean squared errors than those of boosting. For example, for a frailty Lasso model with 100 covariates ( $p = 100$ ) and  $\sigma_b = 1$ , the  $MSE_\gamma$  is 34.06 compared to 61.53 for boosting. Gradient boosting showed poor results for  $MSE_\gamma$ . *GlmLasso* works best for small  $p$ . When the number of covariates in a model increases,  $MSE_\gamma$  increases. For example, for *glmLasso* with  $\sigma_b = 1$ ,  $MSE_\gamma$  for  $p = 15$  and  $p = 100$  are 8.47 and 34.06 respectively.

When there are no random effects i.e.  $\sigma_b = 0$ , *glmLasso* outperforms *glmmLasso*. For example, for  $\sigma_b = 0$  and  $p = 100$ ,  $MSE_\gamma$  for *glmLasso* model is 10.18 compared to 14.84, the  $MSE_\gamma$  for the model with the random effect. However, for non-zero  $\sigma_b$  i.e.  $\sigma_b = \{1, 2\}$ , the methods that account for heterogeneity outperform those that do not include the random effect. For example, in boosting, for  $\sigma_b = 2$  and  $p = 50$ ,  $MSE_\gamma$  for boosting with random effect is 70.28 compared to 90.9 the  $MSE_\gamma$  for boosting model that does not account for heterogeneity. Overall, for the estimate of covariates covariance, *Lasso* yields better than *gradient boosting*.

In terms of  $MSE_\gamma$ , the results are different from Groll and Tutz (2014) as procedures with frailty (*glmmlasso*) perform better than those without (*glmlasso*).

Table 4.2: Results for false positive (f.p) and false negative (f.n),  $\pi_{\text{censor}}=0.05$  and  $n = 100$

$\sigma_b$	$p$	glmLasso		glmmLasso		boosting		boosting (frailty)	
		$f.p$	$f.n$	$f.p$	$f.n$	$f.p$	$f.n$	$f.p$	$f.n$
0	15	0.14	1.99	0.32	1.96	6.7	0.5	0.3	1
0	50	0.62	1.97	1.08	2.08	18	1	1.6	1.1
0	100	0.77	2.04	2.11	2.14	23	1	2.2	1.1
1	15	0.19	1.98	0.27	2.13	5.6	0	0.51	1
1	50	1.06	2.01	1.33	2.24	17	0.5	1.2	1.4
1	100	5.57	2.14	1.53	2.46	30	0.9	4.4	1
2	15	0.19	2.28	0.34	2.57	5	0.3	1	1.7
2	50	1.28	2.37	1.55	2.7	18.5	0.45	2.3	1.5
2	100	4.36	2.72	4.11	2.76	29	0.5	6	1.1

The performance of *lasso* and *gradient boosting* algorithms were evaluated using false positive and false negative. The average over 100 simulations of the number of truly zero variables set to non-zero ( $f.p$ ) and the number of truly non-zero variables set to zero ( $f.n$ ) are computed. From Table 4.2, lasso methods perform better than boosting methods in terms of false positives as shown by their smaller values of  $f.p$ . For  $\sigma_b = 1$  and  $p = 100$  *glmmLasso* recorded mean false positive of 1.53 compared to 4.4 for boosting with a random effect. A big  $\sigma_b$  results in large false positive compared to small values such as  $\sigma_b = 0$ . Furthermore, an increase in the number of variables in a model results in large false positives. Methods that account for heterogeneity have small mean false positives. However, in terms of false negative, boosting methods have lower values than *lasso* methods.

### 4.3 Conclusion Remarks

Evidence from the Mean Squared Error, False Positives and False Negatives from the simulation study suggest that the lasso variable selection outperformed gradient boosting regardless of whether the random effect was included or not in the model. Lasso methods (*glmLasso* and *glmmLasso*) recorded smaller values of MSEs and False Positives than gradient boosting. In addition, the two approaches improve on their performance when few covariates are present, and  $\sigma_b$  is kept small.

# Chapter 5

## Empirical data analysis results

### 5.1 Descriptive Results

This section presents both the univariate and bivariate results in the form of tables and graphs.

#### 5.1.1 Univariate analysis

Table 5.1 below shows that most of the respondents lived in rural areas and account for about 58.5% of the sample population as compared to 41.5% found in urban areas.

In terms of educational level attained, Table 5.1 shows that the majority of the respondents attained secondary education and account for 61.6% of the sample population. The percentage of women reported having no education is 0.8. Table 5.1 also shows that 28.7% and 8.9% of the sample had attained primary and higher education, respectively.

Table 5.1: Summary statistics on independent variables considered for analysis

<i>Variable</i>	<i>Category</i>	<i>Frequency (sample proportion)</i>
Place of residence	Urban	829 (41.5%)
	Rural	1171 (58.5%)
Education	No education	16 (0.8%)
	Primary	575 (28.7%)
	Secondary	1 231 (61.6%)
	Higher	178 (8.9%)
Employment status	No	1095 (54.7%)
	Yes	905 (45.3%)
Birth cohort	<1970	124 (6.2%)
	1970-1979	536 (26.8%)
	1980-1990	851 (42.5%)
	>1990	489 (24.5%)
Religion	Christian	1035 (51.8%)
	Muslim	8 (0.4%)
	Tradition	14 (0.7%)
	Apostolic Faith	827 (41.3%)
	Other/None	116 (5.8%)
Region	Manicaland	215 (10.8%)
	Mashonaland	877 (43.9%)
	Masvingo	205 (10.2%)
	Matabeleland	470 (23.5%)
	Midlands	233 (11.6%)

From Table 5.1 it is seen that more women from the sample population were unemployed (54.7%). Approximately 67% of the respondents from the sample were born after 1980, while 6.2% comprised of those born before 1970. More than half (51.8%) of the respondents in the study believed in Christianity with the smallest percentage being Muslims (0.4%). The table also shows that 41.3% and 5.8% of the sampled respondents are in Apostolic Faith and other or no religious groups, respectively.

According to Table 5.1, more women of the sampled population are from Mashonaland (43.9%). The region with the least number of women in the sample is Masvingo accounting for 10.2%. Matabeleland and Midlands regions account for 23.5% and 11.6% respectively.

### **5.1.2 Bivariate analysis**

The Kaplan-Meier plots are used to describe how the risk of age at first marriage is distributed across the strata of the chosen covariates. The Kaplan-Meier survival estimator is used for comparison of survival curves in two or more groups (Kleinbaum et. al., 2008).



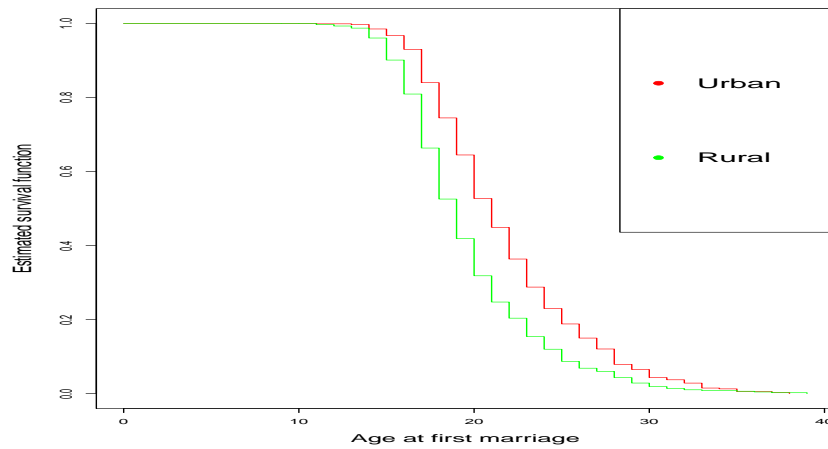


Figure 5.1: Kaplan-Meier survivorship for age at first marriage by place of residence

Figure 5.1 above seems to suggest that no woman respondent married under the age of 15 years. However, women living in rural areas have a higher risk of first marriage in comparison to those living in urban areas.

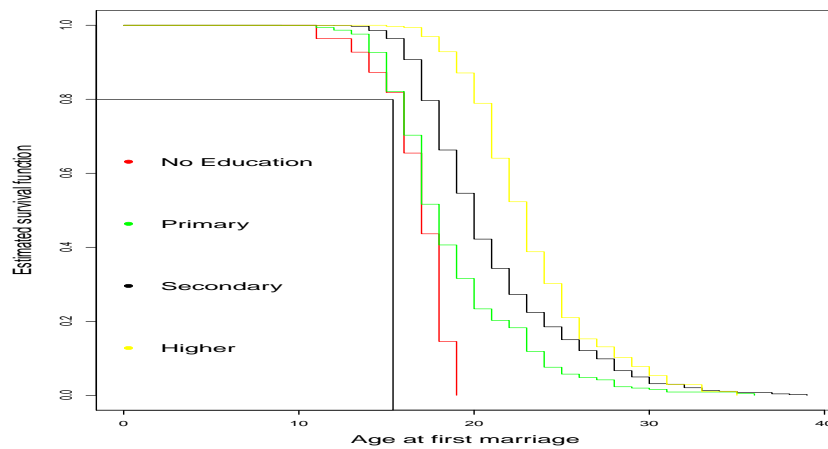


Figure 5.2: Kaplan-Meier survivorship for age at first marriage by education

Figure 5.2 presents survival probability on the educational level attained. From the plot, it seems to show that women who are highly educated are associated with

lower probabilities of early marriages. Women without education or attained primary education have a high likelihood of first marriage in comparison to those who reached secondary and/or higher education. At the age of approximately 19 years, all the respondents that reported to have no education in the sample population had already entered into first marriage. Conversely, at the same age i.e. 19 years, about 80% of the women that had higher education survived early first marriage.

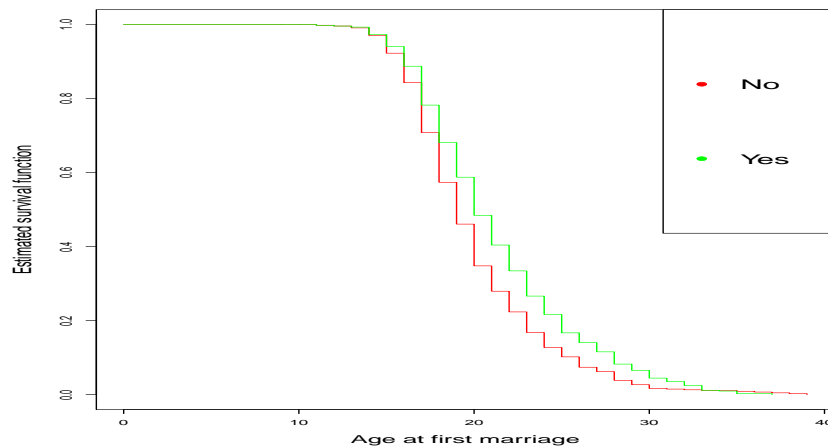


Figure 5.3: Kaplan-Meier survivorship for age at first marriage by employment status

Survivorship probabilities for age at first marriage by employment status are presented in Figure 5.3 above. It is suggested that unemployment of a woman increases the chances of early first marriage in Zimbabwe. Employed women were not highly exposed to the risk of having first early marriage compared to the unemployed ones. For example, from the plot, at the age of 20, about 50% of the women employed survived first marriage compared to 30% for those who are working.

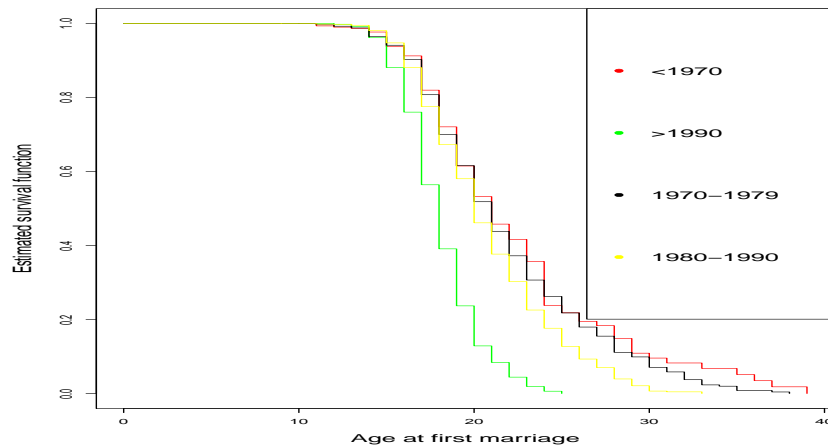


Figure 5.4: Kaplan-Meier survivorship for age at first marriage by birth cohort

Figure 5.4 seems to reveal that those women in the study who were born after 1990 were related to a higher probability of early first marriage. Those born before 1970 had a lower risk of first marriage. Between the ages of 20 and 30 years women born before 1970 and those born 1970-1980 had almost the same risks of entering into marriage. For example, nearly 42% of women in these cohorts would have survived first marriage at the age of 22. However, beyond 30 years, those born before 1970 have the highest survival rate.

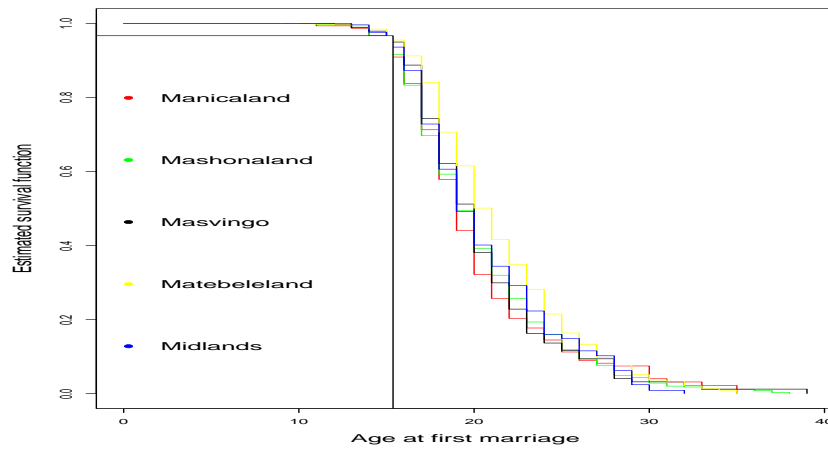


Figure 5.5: Kaplan-Meier survivorship for age at first marriage by region

Figure 5.5 below seems to show that there is a slight difference in the risks of first marriage between regions. However, women from the Matabeleland region had a low risk of first marriage when compared to those from other regions such as Manicaland and Mashonaland. At 20 years, approximately 30% of the women would have survived marriage in Manicaland compare to nearly 50% in Matabeleland.

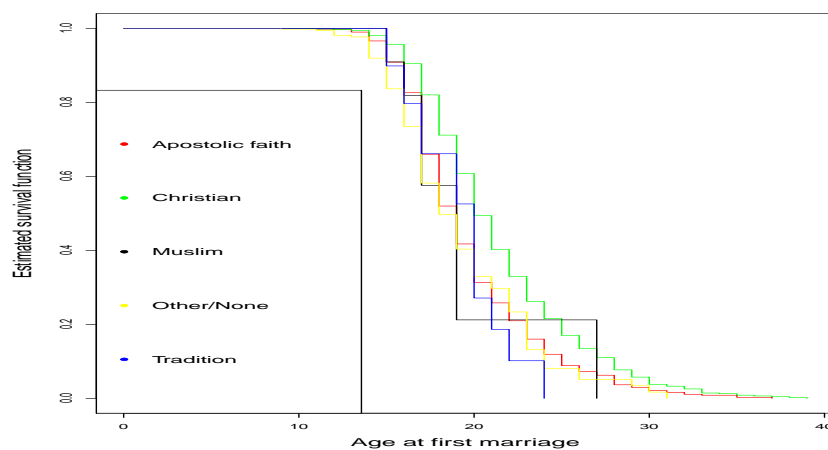


Figure 5.6: Kaplan-Meier survivorship for age at first marriage by religion

Figure 5.6 presents survival probabilities by religious beliefs of respondents in a sample population. The diagram appears to reveal that the risks of first marriage for those from Muslim and traditional religious groups are higher compared to Christians. Before the age of approximately 25 years, all the women that reported to be Muslim or believe in traditions would have entered into their first marriage. At the age of 20 years about 20% of Muslim women would have survived early first marriage.

## 5.2 Estimated linear effects

### 5.2.1 Estimation from *lasso* and *gradient boosting* without random effect

Table 5.2: Predictors' effects without random effect

<i>Variable</i>	<i>Category</i>	<i>Lasso without random effect</i>	<i>gradient boosting</i>
Place of residence	Rural	0.2207	0.0449
Education	(i)Primary	0.1512	-0.0814
	(ii)secondary	-0.4631	-0.4041
	(iii)higher	-0.9253	-0.6692
Employment	Yes	0.0000	-
Age Cohort	(i)>1990	1.0792	0.4753
	(ii)1970-1979	-0.1525	-0.1139
	(iii)1980-1990	0.2819	0.0422
Region	(i)Mashonaland	0.0000	0.0121
	(ii) Masvingo	0.0000	-0.1248
	(iii)Matebeleland	0.0000	-0.1832
	(iv)Midlands	0.0000	-0.0096
Religion	(i)Christian	0.0000	-0.0819
	(ii)Muslim	0.0000	0.1189
	(iii)Other/None	0.0000	0.1661
	(iv)Traditional	0.0000	-0.1112

Table 5.2 gives the estimates of regression coefficients from lasso and boosting without random effect. In *glmLasso*, the regression coefficients of employment, region and religion including all their categories were shrunk until the complete variables were removed from the model. Place of residence, education and age cohort were selected into the model as their coefficients are non-zero. With boosting only employment was not selected into the model.

Gradient boosting retained region and religion into the model though their coefficients are small. This is not surprising in light of higher false positives observed for boosting in the previous chapter. However, the variables with small coefficients were not selected by lasso.

### 5.2.2 Estimation from *lasso* and *gradient boosting* with random effect

Table 5.3 gives the regression coefficients of predictors from the lasso and boosting models that account for heterogeneity. In *glmmLasso*, region and religion are not selected into the model. However, with gradient boosting two variables, place of residence and employment status were not included in the model. Gradient boosting retained more variables into the model compared to lasso.

Table 5.3: Predictors' effects with random effect

<i>Variable</i>	<i>Category</i>	<i>Lasso with random effect</i>	<i>gradient boosting</i>
Place of residence	Rural	0.1407	-
Education	(i)Primary	0.0764	0.1208
	(ii)secondary	-0.5387	-0.1692
	(iii)higher	-1.0921	-0.4445
Employment	Yes	-0.1052	-
Age Cohort	(i)>1990	1.1866	0.2594
	(ii)1970-1979	0.2205	-0.2062
	(iii)1980-1990	0.5851	-0.0973
Region	(i)Mashonaland	0.0000	0.0159
	(ii) Masvingo	0.0000	-0.0579
	(iii)Matebeleland	0.0000	-0.1042
	(iv)Midlands	0.0000	0.020
Religion	(i)Christian	0.0000	-0.0393
	(ii)Muslim	0.0000	-0.1661
	(iii)Other/None	0.0000	0.0536
	(iv)Traditional	0.0000	-0.1287

### 5.2.3 Boosting iterations

Most covariates can be used for fitting a model though no extensive variable selection being performed. Early stopping support in gradient boosting helps to find the least number of iterations which is sufficient to build a model. Early stopping avoid overfitting. In this study, cross-validation was employed in determining the optimal stopping iterations. Figure 5.7 and Figure 5.8 below are the graphs showing the optimal number of boosting iterations for the two boosting models, without and with random effect.

From Figure 5.7 and Figure 5.8 boosting with random effect has less number of

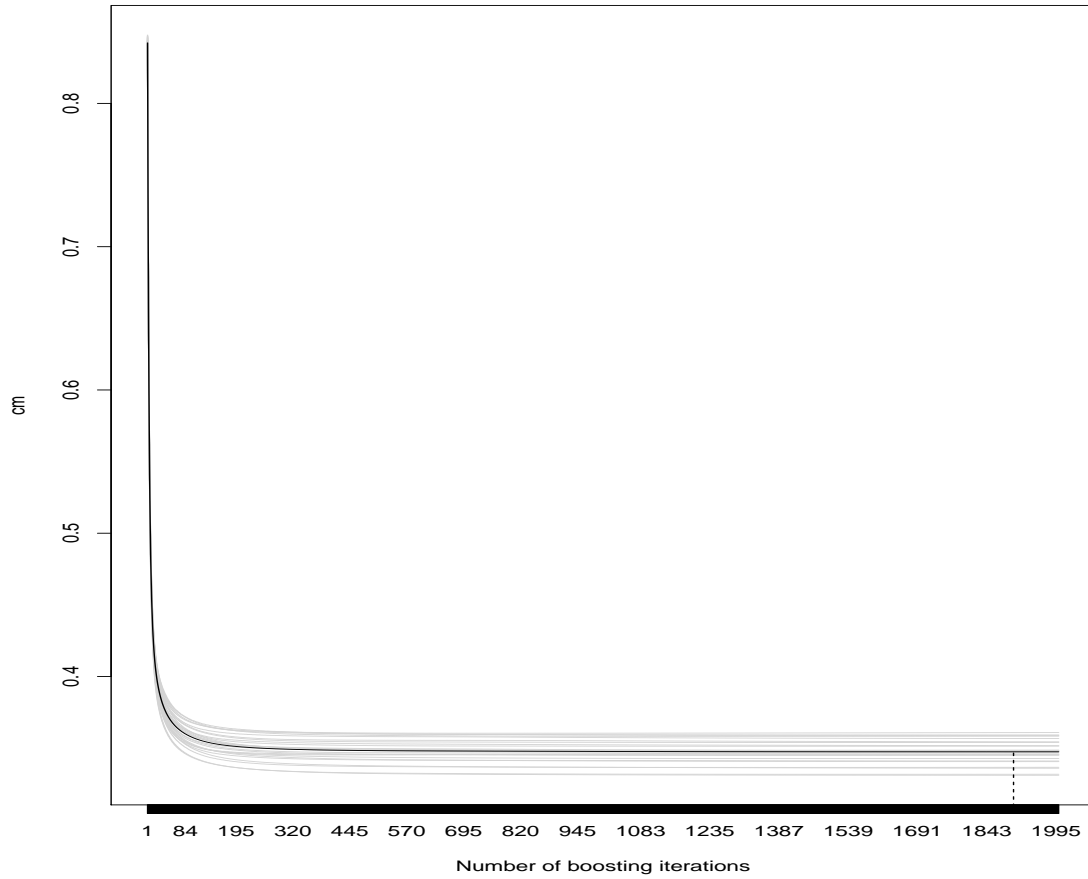


Figure 5.7: Optimal number of boosting iterations= 1887

boosting iterations (951) compared to the boosting model that excludes the random effect (1887).



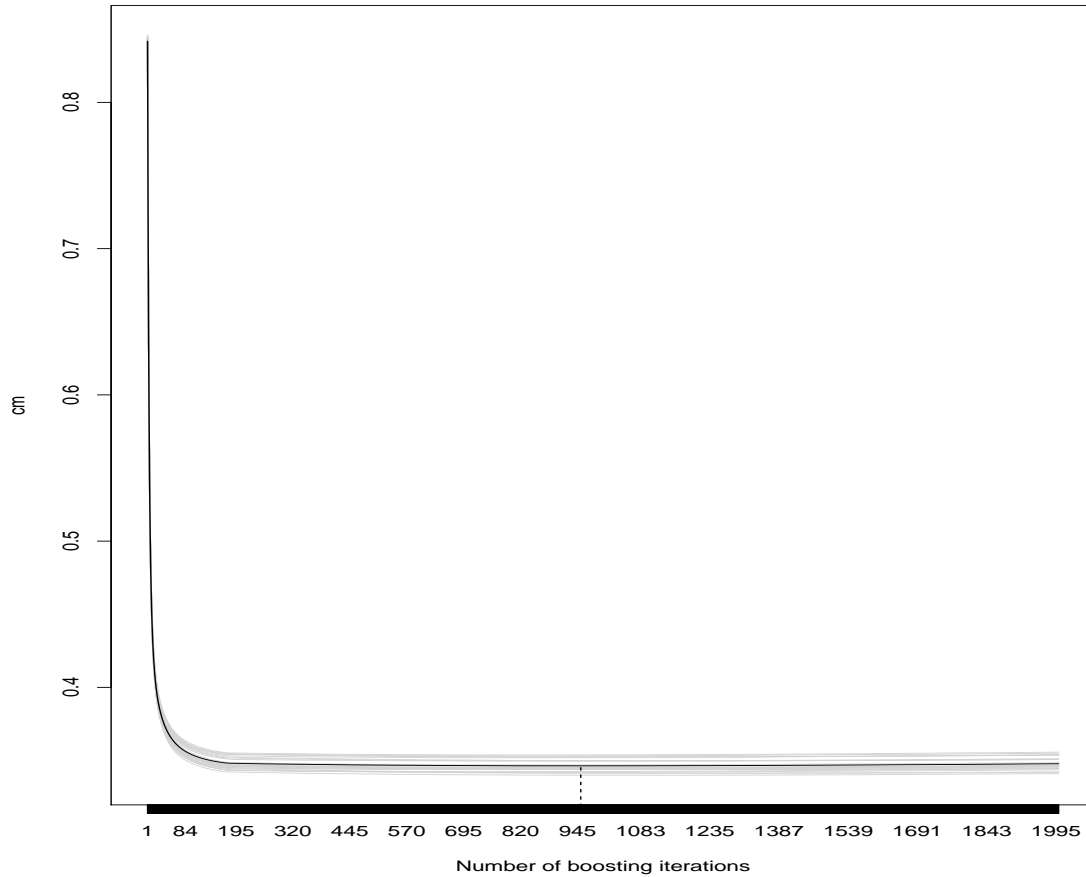


Figure 5.8: Optimal number of boosting iterations= 951

## 5.3 Model diagnostics

### 5.3.1 Residual analysis

Figure 5.9 and Figure 5.10 show the normal quantile plots of the adjusted residuals obtained from boosting and lasso models with and without random effect. From the plots in Figure 5.9 and Figure 5.10, it can be deduced that all the models performed moderately well as points are forming a line that is roughly straight except at the

tails.

### 5.3.2 Concordance Index

Concordance index evaluates prediction accuracy based on discrimination measures. The results from the concordance indices show models with boosting variable selection performing better than lasso models. The concordance indices for boosting with and without random effect are 0.5734 and 0.5453, respectively.  $C^*$  is larger than 0.5 showing better prediction (Tutz and Groll, 2016). For lasso models with and without the concordance indices are 0.2170 and 0.2424, respectively. Lasso indices are far less than 0.5 showing low prediction. Boosting models have higher  $C^*$  probably because they have retained more covariates than lasso.

## 5.4 Model comparison

Table 5.4: Anova table comparing Lasso models with and without the random effect

Model	AIC	BIC	Deviance
<i>glmLasso</i>	9492.3	9549.1	9476.3
<i>glmmLasso</i>	9707.4	9778.5	9687.4

Table 5.5: Testing the significance of the random effect

Group	Name	Variance	Std.Dev
Cluster	(intercept)	0.0000	0.0000
Residual		0.1724	0.4152

Table 5.4 contains the AIC, BIC and Deviance from lasso model without random effect and the lasso frailty model. The model without the random effect has AIC, BIC

and Deviance values of 9 492.3, 9 549.1 and 9 476.3 respectively lower than those of the frailty model. The frailty model model has AIC, BIC and Deviance values of 9 707.4, 9 778.5 and 9687.4 respectively. From Table 5.5, the random effect (cluster) has a variance of 0.0000 indicating that the level of between-group variance is not sufficient to warrant incorporating random effect in the model hence the interpretation of results in this study based on lasso model without random effect.

## 5.5 Interpretation of the results based on the model with *lasso* variable selection

Table 5.6: Results of Discrete-Time Logit Model without random effect

<i>Predictor</i>	<i>Estimate</i>	<i>se</i>	<i>P-value</i>	<i>Hazard ratio</i>	<i>95% C.I</i>
Place(Rural)	0.2207	0.0664	0.0009	1.2469	1.0947; 1.4202
Edu(Primary)	0.1512	0.0767	0.05	1.1632	1.00; 1.3519
Edu(Secondary)	-0.4631	0.1359	0.0007	0.6229	0.4772; 0.813
Educ (Higher)	-0.9253	0.5595	0.0982	0.3964	0.1324; 1.1868
Employment status	0.0000	-	-	1	-
Cohort(>1990)	1.0792	0.0732	2.2e-16	2.9423	2.549; 3.3962
Cohort(1970- 1979)	-0.1525	0.0811	0.0599	0.8586	0.7324; 1.006
Cohort(1980- 1990)	0.2819	0.2142	0.1880	1.3256	0.8711; 2.0172

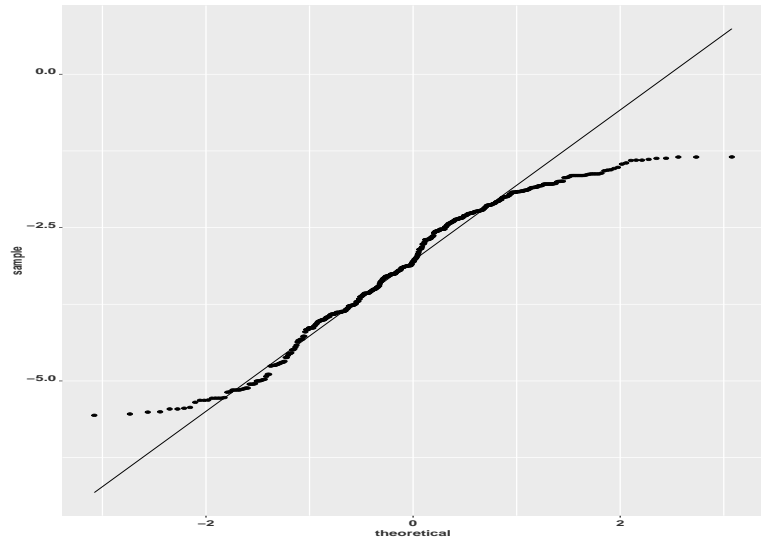
From Table 5.6, the effect of each predictor in the model is determined by its hazard ratio. The hazard ratio measures the increase or decrease in the hazard of age at first marriage due to the influence of the respective risk factor. The significance of each predictor is determined by its respective *p – value*.

As observed from Table 5.6, place of residence by the respondents was found to be significant in describing the hazard for age at first marriage. Rural residence is associated with high risk of first marriage. The hazard ratio for this group was 1.2469 indicating that Zimbabwean women residing in rural areas are 24.69% more likely to enter into first marriage than their urban counterparts. At 95% confidence level, the hazard for rural women ranges from 9.47% to 42.02% higher than that for women residing in urban areas.

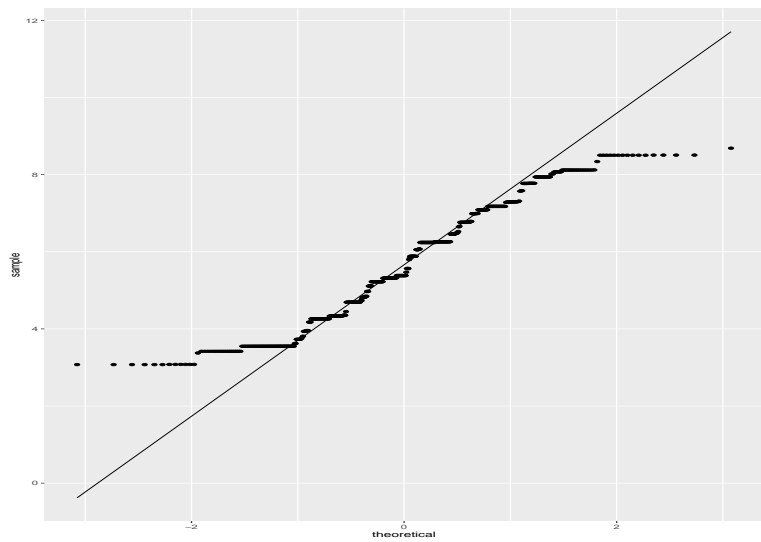
From Table 5.6, it is also observed that educational level attained by the respondents had a significant association with the risk of first marriage. It is observed that the hazard of having early first marriage is significantly ( $p < 0.05$ ) lower (37.71%) for respondents with secondary education than their counterparts with no education. We are 95% confident that for this category, the hazard for first marriage is between 18.2% and 52.28% lower than those women with no education. However, primary and higher educational levels were not associated with the probability of first marriage as shown by their  $p$  – values greater than 0.05 and 95% confidence intervals of hazard ratios (1.00; 1.3519) and (0.1324; 1.1868) respectively containing 1.

Age cohort was found to have a significant effect on age at first marriage in Zimbabwe. Women in the study who were born after 1990 were associated with a higher probability of early first marriage. The group was 194.23% more probably to marry earlier than women born before 1970. At 95% confidence level, the hazard for cohort(>1990) women ranges from 154.9% to 239.69% higher than that for women born before 1970.

Finally, the age cohorts 1970-1979 and 1980-1990 for Zimbabwean women in the study were observed not to have an association with the timing of first marriage as shown by  $p$  – values of 0.0599 and 0.1880 respectively. These  $p$  – values are greater than 0.05. Their corresponding interval of hazard ratios (0.7324; 1.0065) and (0.8711; 2.0172) respectively contain 1.

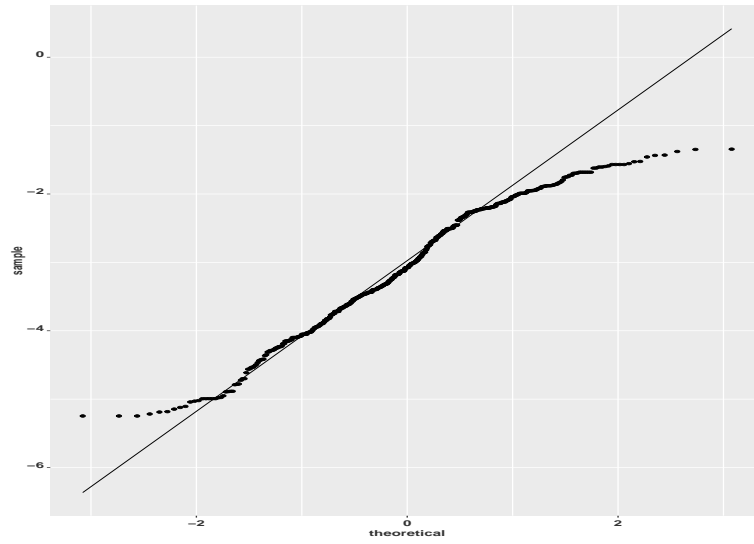


(a) Boosting without random effect

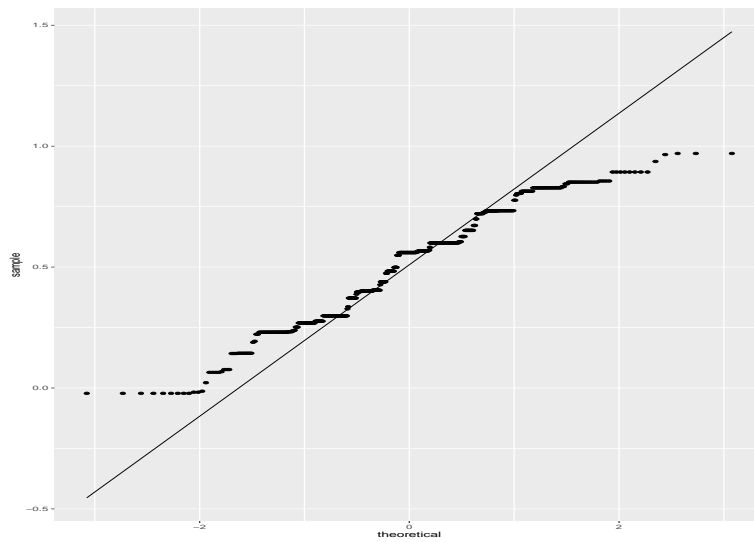


(b) Lasso without random effect

Figure 5.9: Normal Q-Q (Quantile to Quantile) plots of the adjusted residuals for models without random effect



(a) Boosting with random effect



(b) Lasso with random effect

Figure 5.10: Normal Quantile-Quantile plots of the adjusted residuals for models with random effect

# Chapter 6

## Conclusion

### 6.1 Introduction

This chapter covers a summary of the significant findings of the study and limitation

### 6.2 Conclusion

The study was set to evaluate and compare the effectiveness of lasso and gradient boosting variable selection methods in discrete survival models. The effect of a random effect in a model was investigated.

A simulation study was conducted for the two-variable selection methods. Mean Squared Errors, False Positives and False Negatives were used to evaluate the performance of the two methods. From the simulation study, lasso performed better than gradient boosting method, especially with small number of covariates in the model as shown by small MSEs and False Positives.



The methods mentioned above were used to build models on age at first marriage for women in Zimbabwe using the 2015-16 Demographic Health Survey (DHS) data. Discrete Logit Models with and without random effects were fitted.

It was found that the two methods, lasso and gradient boosting do not select exactly the same variables. The results based on lasso revealed that place of residence, highest educational level attended and birth cohort had a significant impact on the probability of age at first marriage. Women from rural areas and those born after 1990 were significantly more likely to get married earlier. However, women with secondary education were significantly less likely to enter into early first marriage. Region, religion and employment status of a woman were not significant determinants of age at first marriage.

### **6.3 Limitations of the study**

This study focused on variable selection in discrete survival models with smooth baseline hazard and a logistic family of link functions. Perhaps as future research, a comparison between discrete survival models with smooth and discrete baseline hazards and/or flexible link function can be investigated.

## 6.4

### References

1. Anderson, D. A., and Aitkin, M. (1985). Variance component models with binary response: interviewer variability . *JR Stat Soc Ser B*, 47, 203-210.
2. Austin, P. C., and Lees, D.S. (2016). Introduction to the analysis of survival data in the presence of competing risk. *Pubmed*, 133(6), 601-9
3. Axinn, W. G., and Thornton, A. (1992). The influence of parental resources on the timing of the transition to marriage. *Soc. Sci. Res*, 21, 61-85.
4. Biggeri, L. M., Bini, M. A., and Grill, L. (2001). The transition from the university to work: a multilevel approach to the analysis of the time to obtain the first job. *J. R. Stat. Soc. Ser. A*, 164, 293-305.
5. Bledsoe, C. H ., and Cohen, B. (1993). Social dynamic of adolescent in sub-Saharan Africa. *National Academy Press*, 43.
6. Bracher, M., and Santow, G. (1998). Economic independence and union formation in Sweden. *Populat. Stud*, 52 (3), 275-294.
7. Breheny, P., and Jian, J. (2015). Group descent algorithm for non-convex penalized linear and logistic regression models with grouped predictors. *Statistical Computing*, 2(25), 173-187.
8. Breiman, L. (1996). Bagging predictors *Annals of Statistics*, 24, 123-140.
9. Bühlmann, L., and Hothorn, T. (2007). Boosting algorithm: Regularization, predictors and model fitting (with discussion). *Statistical Science*, 22, 477-522.
10. Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34, 187-220.

11. Efrom, B. (1988). Logistic regression, survival analysis, and the kaplan-meier curve. *Journal of the American Statistical Association*, *83*, 414-425.
12. Ekman, A. (2017). Variable selection for the Cox proportional hazards models: A simulation study comparing the stepwise, lasso and bootstrap approach. *Unpublished masters thesis*.
13. Fahrmer, L. (1994). Dynamic modelling and penalized likelihood estimation for discrete time survival data. *Biometrika*, *81*, 317-330.
14. Fan, J., and Li, R. (2001). Variable selection via non-cave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348-1360.
15. Freund, Y., and Schapire, R. E. (1996). Experiments with a new boosting algorithm. *In-ICML*, *9*, 148-156.
16. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *1*(29), 1189-1232.
17. Friedman, J. H, Tibshirani, R., and Hastie, T. (2007). Pairwise Co-ordinates Optimization. *Annals of Applied Statistics*, *5*, 302-322.
18. Friedman, J. H., Tibshirani, R., and Hastie, T. (2010). Regularization paths for generalized linear models via co-ordinates descent. *Journal of Statistical Software*, *1*, 1-22.
19. Garrenne, M. (2004). Age at marriage and modernisation in Sub-Saharan Africa. *South. Afr. J. Demogra*, *9* (2), 58-80.
20. Gaskin, K. (1978). Age at first marriage in Europe before 1850: A summary of family reconstitution data. *Journal of Family History*, *3*, 23-26.

21. Goeman, J. J. (2010).  $L_1$ -penalized estimation in the Cox proportional hazards model *BiomJ*, 52, 70-84.
22. Gould, W. W., Gutierrez, R. G., and Cleves, M. A. (2008). An introduction to survival analysis using Stata *Stata Corp.*
23. Groll, G., and Tutz, G. (2014). Variable selection for generalized linear mixed models by  $l_1$ -penalized estimation. *Statistics and Computing*, 24, 137-154.
24. Groll, A., and Tutz, G. (2016). Variable selection in discrete survival models including heterogeneity. *Lifetime Data Anal* [published online].
25. Groll, A., and Tutz, G. (2017). Variable selection in discrete survival models including heterogeneity. *Lifetime Data Anal*, 23, 305-338.
26. Guo, G., and Rodriguez, G. (1992). Estimating a multivariate proportional hazards model for clustered data using EM algorithm, with an application to survival in Guatemala. *J. Am. Statist. Assoc*, 87, 969-976.
27. Hajnal, J. (1953). Age at marriage and proportions. *Population Studies*, 2, 111-136.
28. Harwood-Lejeune, A. (2001). Rising age at marriage and fertility in Southern and Eastern Africa. *European Journal of Population*, 17(3), 261-280.
29. Heagerty, P. J., and Zheng, Y. (2005). Survival model predictive accuracy and roc curves. *Biometrics*, 61, 92-105.
30. Hess, W., and Persson, M. (2010). Duration of Trade Revisited: Continuous time vs Discrete-time hazard. *Working paper*, 829.
31. Hinde, J. (1982). *Compound poisson regression models*. New York: Springer

32. Hoffertn, S. L., and Reid, L. (2002). Early childbearing and achievement's and behaviour overtime. *Persepctive Sex Reproduction Health*, 34(1), 67-84.
33. Hofner, B., Robinzonov, N., Mayr, A., and Schmid, M. (2014). Model-based boosting in r: a hands-on tutorial using rpackage mboost *Computational Statistics*, 29, 3-35.
34. Hothorn, T., Bühlmann, L., Klein, T., Schmid, M., and Hofner, B. (2015). Mboost: Model-based boosting. R package version 2.5.0. *Statistical Science*, 53.
35. Hurvich, C., and Tsai, C. (1990). The impact of model selection on inference in linear regression. *The American Statistician*, 44, 214-217.
36. Islam, M. (1999). Adolescent childbearing in Bangladesh. *Asia-Pacific Population Journal*, 14(3), 73-87.
37. Jensen, R., and Thornton, R. (2003). Early female marriage in the developing world.
38. Kalbleish, J. D., and Prentice, R. L. (2002). *The statistical analysis of failure data (2nd ed)*. New York: Wiley.
39. Kamal, S. M. (2012). Adolescent motherhood in Bangladesh: Evidence from 2007 BDHS data. *Canadian Studies in Population*, 39(1), 63-82.
40. Klein, J. P., and Goel, P. K. (1992). Survival analysis: state of art, volume 211. *Springer*, 24.
41. Kleinbaum, D., Kupper, L., Nizan, A., and Muller, K. (2008). Applied regression analysis and other multivariable methods. *Thomas Higher Education*, 1, 311-314.

42. Lesthaeghe, R., Kaufmann, G., and Meekers, D. (1989). *The Nuptiality regimes in sub-Saharan Africa*. University of California Press.
43. Lewis, S. M., and Raftery, A. E (1999). Comparing explanations of fertility decline using event history models and unobserved heterogeneity. *Social Methods Res* , 28, 35-60.
44. Lumley, L. (2017). Leaps: Regression subset selection. R Package Version 3.0 *Available online*.
45. Manda, S., and Meyer, R. (2005). Age at first marriage in Malawi: A Bayesian multilevel analysis using discrete time to event-model. *Journal of the Royal Statistical Society A*, 168, 439-455.
46. Mayr, A., Gefeller, O., Binder, H., and Schmid, M. (2014a). The evolution of boosting algorithms (with discussion). *The Methods of Information in Medicine*, 53, 419-427.
47. Mayr, A., Gefeller, O., Binder, H., and Schmid, M. (2014b). The evolution of boosting algorithms (with discussion). *The Methods of Information in Medicine*, 53, 428-435.
48. McDevitt, T. M., Fowler, T. B., Adlakha, A., and Harris-Bourne, V. (1996). Trends in adolescent fertility and contraceptive use in the developing world.
49. Meier, L., Geer, S. V. D., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 1(70), 53-71.
50. Mirowsky, J. (2005). Age at first birth, health and mortality. *Journal of Health and Social Behaviour*, 46(1), 32-50.

51. Möst, S., Tutz, G., and Pöbnecker, W. (2016). Variable selection for discrete competing risk model: Quality and Quantity. *Annals of Statistics*.
52. Palamuleni, M. E. (2011). Socioeconomic determinants of age at marriage in Malawi. *Journal of Sociology and Anthropology*, 3(7), 224-235.
53. Schmid, M., and Hothorn, T. (2008). Boosting additive models using component-wise p-splines. *Computational Statistics*, 53, 298-311.
54. Schmid, M., Kestler, H. A., and Potapov, S. (2014). On the validity of time-dependent estimators. *Bioinformatics*.
55. Singer, J. D., and Willet, J. B. (2003). *Applied Data Analysis: Modelling change and event occurrence*. Oxford University Press.
56. Singh, S., and Samara, R. (1996). Early marriage among women in developing countries *International Family Planning Perspective*, 22(4), 148-157.
57. Steyerberg, E. W., and Marius, J. C. (1999). Stepwise selection in small data sets: A simulation study of bias in logistic regression analysis. *Journal of Clinical Epidemiology*, 52, 935-942.
58. Tibshirani, R. (1996). Regression shrinkage and selection via lasso *Journal of the Royal Statistical Society, Series B: Methodology*, 58, 267-288.
59. Tutz, G., and Binder, H. (2006). Generalized additive modelling with implicit variable selection by likelihood-based boosting. *Biometrics*, 62, 961-971.
60. Tutz, G., and Schmid, M. (2016). *Modelling discrete time-to-event data*. Munich: Springer.
61. UNICEF. (2001). Early marriage. *Innocenti Digest*, 7.

62. UNICEF. (2007). Progress for Children: A world fit for children *Statistical Review*, 6.
63. Uno, H., Tian, L., Cai, L., and Wei, L. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association*, 102, 527-537.
64. Zimbabwe National Statistics Agency and ICF International. (2012). *Zimbabwe Demographic Health Survey 2010-2011*.
65. Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418-1429.
66. Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67, 301-320.



## 6.5 Appendix

```
rm(list = ls())
library(glmLasso)
library(foreign)
  library(survival)
  library(discSurv)
  library(mgcv)
  library(gamlss.mx)
  library(mgcv)
  library(MASS);
  library(nlme)
  library(Hmisc)
  library(mboost)
  library(car)
  library(Ecdat)
  source("C:/Users/Mabvu/Desktop/Discret/SurvivalSimulation/bs.design.r")
MYFUNC <- function(M)
y<-NULL
CN<-NULL
for (i in 1:nrow(M))
y[i] <-rbinom(1,1,M[i,3])
if (y[i]==1)
v<-c(rep(0,i-1),1)
P<-M[1:i,] oup<-as.data.frame(cbind(P,v))
Error1<-NULL
Error2<-NULL
```

```
Sigma<-NULL
FN<-NULL
FP<-NULL
ni<-100
for (ii in 1:ni)
print(paste("ii ", ii,sep=""))
df<-matrix(NA,100*12,15)
colnames(df)< colnames
(df) <- - paste0('x', 1:(ncol(df)))
print(df)
set.seed(23)
Xs< -c(6,-4,4,0.2,-6, rep(0,10))
tm <-c(1:12)
baseline< -dgamma(tm-2, shape=5, scale = 1)*2-2.3
beta< -c(baseline,Xs)
cvs< -df
cvs < -rowSums(cvs)
bi <-rep(rnorm(100, mean = 0, sd = 1),each=12)
t <- rep(1:12, times=100)
t <- rep(1:12, times=100)-2
ID <- as.factor(rep(1:100, each=12))
bh <-2*dgamma(t, shape=5, scale= 1)-2.3
beta <-2*dgamma(c(1:12), shape=5, scale= 1)-2.3
h <-exp(bh+cvs+bi)/(1+exp(bh+cvs+bi))
D2 <-NULL
D3 <-NULL
D4 <-NULL
D5 <-NULL
```

```
D6 <-NULL
D7 <-NULL
D8 <-NULL
D9 <-NULL
D10 <-NULL
D11 <-NULL
D12 <-NULL
for (i in 1:1200)
  D2[i]<-ifelse(t[i]==2,1,0)
  D3[i]<-ifelse(t[i]==3,1,0)
  D4[i]<-ifelse(t[i]==4,1,0)
  D5[i]<-ifelse(t[i]==5,1,0)
  D6[i]<-ifelse(t[i]==6,1,0)
  D7[i]<-ifelse(t[i]==7,1,0)
  D8[i]<-ifelse(t[i]==8,1,0)
  D9[i]<-ifelse(t[i]==9,1,0)
  D10[i]<-ifelse(t[i]==10,1,0)
  D11[i]<-ifelse(t[i]==11,1,0)
  D12[i]<-ifelse(t[i]==12,1,0)
for i DF<-data.frame(cbind(ID,t,h,D2,D3,D4,D5,D6,D7,D8,D9,D10,D11,D12,df))
CM<- data.frame(matrix(ncol = 30, nrow = 0))
sq<-seq(0,1200,12)
w<-length(sq)-1
for (j in 1:w)
  L<-sq[j]+1
  U<-sq[j+1]
  DT<-DF[L:U,]
  DD<-MYFUNC(DT)
```

```

CM<-data.frame(rbind(CM,DD))
CMID <- as.factor(CMID)
formula.1 <- as.formula(paste("v ",paste("x",(1:15),sep=" ",collapse=" +"),sep=""))
mean.vec<-apply(CM[,c(15:29)],2,mean)
sigma.vec<-apply(CM[,c(15:29)],2,sd)
CM[,c(15:29)]<-scale(CM[,c(15:29)])
family = binomial(link = logit)
lambda <- seq(40,0,by=-2)
BICvec<-rep(Inf,length(lambda))
Delta.start<-as.matrix(t(rep(0,100+16)))
Q.start<-0.1
for(j in 1:length(lambda))
print(paste("Iteration ", j,sep=" "))
glm3 <- glmLasso(formula.1 ,rnd = list(ID= 1), family = family, data = CM,
lambda=lambda[j], switch.NR=F,final.re=TRUE, control = list(smooth=list(formula= -
1 + t, nbasis=7,spline.degree=3, diff.ord=2, penal=10), start=Delta.start[j,],qstart=Q.start[j]))

print(colnames(glm3Deltamatrix)[2:16][glm3Deltamatrix[glm3conv.step,2:16]!=0])
BICvec[j]<-glm3bic
Delta.start<-rbind(Delta.start,glm3Deltamatrix[glm3conv.step,])
Q.start<-c(Q.start,glm3Qlong[[glm3conv.step+1]])
opt3<-which.min(BICvec)
glm3final <- glmLasso(formula.1, rnd =list(ID= 1) , family = family, data =
CM, lambda=lambda[opt3], switch.NR=F,final.re=TRUE, control = list(smooth=list(formula= -
1 + t, nbasis=7,spline.degree=3, diff.ord=2, penal=10),start=Delta.start[opt3,],qstart=Q.start[opt3
summary(glm3final)
spline.ma<-bs.design(1:12, diff.ord=glm3finaldiff.ord,
Design<-cbind(spline.maX[,-1],spline.maZ)

```

```

final.coef<-glm3finalcoef/c(1,sigma.vec[1:15])
final.coef[1]<-final.coef[1]-(mean.vec[1:15]/sigma.vec[1:15])
Error2[ii]<-sum((beta[1:12]-t(Design)
Sigma[ii]<-glm3finalStdDev
FN[ii]<-sum(final.coef[2:7]==0)
FP[ii]<-sum(final.coef[8:16]!=0)
Error1<-Error1[Error1<500]
Error2<-Error2[Error2<500]
Sigma<-Sigma[Sigma<2.5]
print(mean(Error1))
print(mean(Error2))
print(mean(Sigma))
print(mean(FN))
print(mean(FP))

dataset = read.spss("D:/zimwomen.sav", to.data.frame=TRUE,use.value.labels=TRUE)
dataset1<-data.frame(cbind(dataset[,c(3,4,13,15,29,45,51,66,3634,3684)]))
colnames(dataset1)[1] <- "CL"
colnames(dataset1)[3] <- "BirthYear"
colnames(dataset1)[5] <- "Pres"
colnames(dataset1)[10] <- "EmplmntStatus"
attach(dataset1)
AFI<-NULL
CNS<-NULL
Ctime<-NULL
Cohort<-NULL
Religionm<-NULL
Regionm<-NULL

```

```
for (i in 1:length(AFM))
#print(MaritalS[i]==1)
AFI[i]<-ifelse(is.na(AFM[i])==TRUE,Age[i],AFM[i])
CNS[i]<-ifelse(is.na(AFM[i])==TRUE,0,1)
#Time
if (AFI[i]>=8 & AFI[i]<11)
Ctime[i]<-1
else if(AFI[i]>=11 & AFI[i]<14)
Ctime[i]<-2
else if(AFI[i]>=14 & AFI[i]<17)
Ctime[i]<-3
else if(AFI[i]>=17 & AFI[i]<20)
Ctime[i]<-4
else if(AFI[i]>=20 & AFI[i]<23)
Ctime[i]<-5
else if(AFI[i]>=23 & AFI[i]<26)
Ctime[i]<-6
else if(AFI[i]>=26& AFI[i]<29)
Ctime[i]<-7
else if(AFI[i]>=29 & AFI[i]<32)
Ctime[i]<-8
else if(AFI[i]>=32& AFI[i]<35)
Ctime[i]<-9
else if(AFI[i]>=35 & AFI[i]<38)
Ctime[i]<-10
else if(AFI[i]>=38 & AFI[i]<41)
Ctime[i]<-11
else if(AFI[i]>=41 & AFI[i]<44)
```

```

Ctime[i]<-12
else
Ctime[i] <-13
#BirthCohort
if(BirthYear[i]<1970)
Cohort[i] <-"< 1970"
else if(BirthYear[i]<1980 & BirthYear[i]>=1970)
Cohort[i] <->1970-1979"
else if(BirthYear[i]<1990 & BirthYear[i]>=1980)
Cohort[i] <->1980-1990"
else
Cohort[i] <->1990"
#Religion
if(Religion[i]==>Traditional")
Religionm[i]<->Traditional"
else if(Religion[i]==>Muslim")
Religionm[i]<->Muslim"
else if (Religion[i]==>Other" — Religion[i]==>None")
Religionm[i]<->Other/None"
else if(Religion[i]==>Apostolic sect")
Religionm[i]<->Apostolic Faith"
else
Religionm[i]<->Christian"
#Region
if(Region[i]==>Bulawayo" —Region[i]==>Matabeleland North" —Region[i]==>Matabeleland
South")
Regionm[i]<->Matabelaland"
else if (Region[i]==>Masvingo")

```

```

Regionm[i]<-"Masvingo"
else if (Region[i]=="Midlands")
Regionm[i]<-"Midlands"
else if (Region[i]=="Manicaland")
Regionm[i]<-"Manicaland"
else
Regionm[i]<-"Mashonaland"
dataset1<-data.frame(cbind(dataset1,AFI,CNS,Ctime,Cohort,Religionm,Regionm))
attach(dataset1)
dataset1<-dataset1[complete.cases(dataset1), ]
dataset1<-dataset1[sample(nrow(dataset1), 2000), ]
dtLong <-as.data.frame(dataLong (dataSet=dataset1, timeColumn=" Ctime",
censColumn=" CNS"))
head(dtLong)
dtLong$CL<-as.factor(dtLong$CL)
dtLong$timeInt<-as.numeric(dtLong$timeInt)
dtLong$timeInt<-dtLong$timeInt-mean(dtLong$timeInt)
formula.1<-y ~as.factor(Pres)+as.factor(Educ)+as.factor(EmplmntStatus)+as.factor(Cohort)+a
family<-binomial(link = "logit")
m1<-gamboost(as.factor(dtLong$y)~bbs(timeInt,center=TRUE,knots=20,degree=2,differences=
bols(Cohort,intercept=FALSE)+ bols(Regionm,intercept=FALSE)+bols(Religionm,intercept=FALSE)
,data = dtLong,family = Binomial(),boost control(mstop = 100,stopintern =
TRUE,nu = 0.09,trace = TRUE,center = TRUE),offset = mean(dtLong$y))
m1[2000]
set.seed(23)
cvm <-cvrisk(m1)
print(cvm)
plot(cvm)

```



```

bn<-mstop(cvm)
vv<-m1[mstop(cvm)]
coef(vv)
IDsample <- sample(1:dim(dataset1)[1], 600)
sub <- dataset1 [IDsample, ]
names(sub)<-names(dataset1)
datLong <-as.data.frame(dataLong (dataSet = sub,
timeColumn=" Ctime",censColumn=" CNS"))
datLong$timeInt<-as.numeric(datLong$timeInt)
#datLong$timeInt<-datLong$timeInt-mean(datLong$timeInt)
m2<-gamboost(as.factor(dtLong$y) bbs(timeInt,center=TRUE,knots=20,degree=2,differences=
bols(Cohort,intercept=FALSE)+ bols(Regionm,intercept=FALSE)+bols(Religionm,intercept=FALSE)
,data = dtLong,family = Binomial(),boostcontrol(mstop = 100,stopintern = TRUE,nu =
0.09,trace = TRUE,center = TRUE),offset = mean(dtLong$y))
m2[2000]
set.seed(23)
cvm<-cvrisk(m2)
print(cvm)
plot(cvm)
mstop(cvm)
vv<-m2[mstop(cvm)]
coef(vv)
#hazPreds <- predict(gamFit, type="response")
hazPreds <- predict(vv, type = "response", newdata = datLong)
adj <- adjDevResidShort ( dataSet = datLong , hazards = hazPreds )
$Output$ AdjDevResid l <- quantile (adj ,0.1)
u <- quantile (adj ,0.90)
tadj<-adj[adj>l & adj <u]

```

```

df <- data.frame(y = tadj)
p <- ggplot(df, aes(sample = y)) p + stat qq() + stat qq line()
tadjq <-data.frame (y = tadj )
pq <- ggplot (tadjq , aes( sample = y)) pq + stat qq () + stat qq line()+
theme ( axis.text.x = element text ( colour =" grey20 ", size = 12 , angle = 0,
hjust = .5 , vjust = .5 , face =" bold"), axis.text.y = element text(colour =" grey20
",size = 12 , angle = 0, hjust = 1, vjust = 0, face =" bold"),
axis.title.x = element text ( colour =" grey20 ", size = 12 , angle = 0, hjust =
.5 , vjust = 0, face =" bold"), axis.title.y = element text( colour =" grey20 ",
size = 12 , angle = 90 , hjust = .5 , vjust = .5 , face =" bold"))
# Extract subset of data set.seed (3) IDsample <-sample(1:dim(dataset1)[1] ,
600)
Sub <- dataset1 [IDsample , ]
set.seed(-7)
TrainingSample<-sample(1:600 , 400)
Train <- Sub[TrainingSample, ]
Test<-Sub [-TrainingSample, ]
# Convert to long format
TrainL<-dataLong(Train,timeColumn =" Ctime", censColumn =" CNS")
TrainL$timeInt<-as.numeric(TrainL$timeInt)
#TrainL$timeInt1<-TrainL$timeInt-mean(TrainL$timeInt)
#Test$Ctime<-Test$Ctime-mean(TrainL$timeInt)
m3<-gamboost(as.factor(dtLong$y) bbs(timeInt,center=TRUE,knots=20,degree=2,differences=
bols(Cohort,intercept=FALSE)+ bols(Regionm,intercept=FALSE)+bols(Religionm,intercept=FA
,data = dtLong,family = Binomial()),boost control(mstop = 100,stopintern =
TRUE,nu = 0.09,trace = TRUE,center = TRUE),offset = mean(dtLong$y))
m3[2000]
set.seed(23)

```

```
cvm<-cvrisk(m3)
print(cvm)
plot(cvm)
mstop(cvm)
vv<-m3[mstop(cvm)]
#coef(vv)
gamFitPreds2 <- predict (vv , newdata = cbind (Test , timeInt = Test $Ctime))
fprGamFit <- fprUnoShort (timepoint =6, marker = gamFitPreds2,newTime =
Test$Ctime)
tprGamFit <- tprUnoShort (timepoint =6 , marker = gamFitPreds2 , newTime
= Test$Ctime, newEvent = Test$CNS , trainTime = Train$Ctime , trainEvent =
Train$CNS)
tryAUC <- aucUno ( tprObj = tprGamFit , fprObj = fprGamFit )
tryAUC
plot (tryAUC)
tryConcorIndex <- concorIndex (tryAUC, printTimePoints=TRUE) tryConcorIndex
summary(tryConcorIndex)
```