University of Venda

A COMPARISON OF SOME METHODS OF MODELING BASELINE HAZARD FUNCTION IN DISCRETE SURVIVAL MODELS

By

Mashabela Mahlageng Retang (Student No. 11636207)

SUBMITTED IN FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
AT
UNIVERSITY OF VENDA
THOHOYANDOU, VENDA
AUGUST 2019

UNIVERSITY OF VENDA
DEPARTMENT OF
STATISTICS

The undersigned hereby certify that they have read and recommend to the Faculty of Mathematics and Natural Science for acceptance a dissertation entitled "**A COMPARISON OF SOME METHODS OF MODELING BASELINE HAZARD FUNCTION IN DISCRETE SURVIVAL MODELS**" by **Mashabela Mahlageng Retang (Student No. 11636207)** in fulfillment of the requirements for the degree of **Master of Science**.

Dated: August 2019

Supervisor: _____
Dr. Alphonce Bere

Co-supervisor: _____
Dr. Caston Sigauke

_____

Declaration

I, **Mashabela Mahlageng Retang** (student no: 11636207), declare that this dissertation titled "**A comparison of some methods of modeling the baseline hazard function in discrete survival models**" hereby submitted for qualification of Masters degree in Statistics at the University of Venda is the original record of my research and not been submitted for any degree or professional qualification in any institution. This study is almost completely my own work and all the references have been provided on all the supporting literatures and resources.

Signature: _____ Date: _____

To God, my Father in heaven, without You I can do nothing.

And to Thato my son, I love you very much my boy.

Table of contents

List of Tables

List of Figures

Abstract

The baseline parameter vector in a discrete-time survival model is determined by the number of time points. The larger the number of the time points, the higher the dimension of the baseline parameter vector which often leads to biased maximum likelihood estimates. One of the ways to overcome this problem is to use a simpler parametrization that contains fewer parameters. A simulation approach was used to compare the accuracy of three variants of penalised regression spline methods in smoothing the baseline hazard function. Root mean squared error (RMSE) analysis suggests that generally all the smoothing methods performed better than the model with a discrete baseline hazard function. No single smoothing method outperformed the other smoothing methods. These methods were also applied to data on age at first alcohol intake in Thohoyandou. The results from real data application suggest that there were no significant differences amongst the estimated models. Consumption of other drugs, having a parent who drinks, being a male and having been abused in life are associated with high chances of drinking alcohol very early in life.

Keywords: *discrete-time survival model, hazard function, baseline hazard function, smoothing splines, penalised regression splines, RMSE*

Acknowledgements

I am greatly indebted to the God who has been with me since the inception of this project through to the end. It has been the longest thorny journey I'd traveled in my life but I am profoundly grateful for His unceasing presence that has been my constant strength. "The Joy of the LORD is my strength".

Dr. Alphonce Bere, words fail to express my gratitude for your invaluable contribution to this research. Your unwavering patience, guidance and support was crucial to the successful completion of this work.

Many thanks to Hulisani Sithuba for generously availing his data for analysis in this study. May God richly bless your generosity. I dare not forget the invaluable encouragement from Dr. Caston Sigauke, who refused to allow me to give up on this programme.

My family, more especially my son Thato, thank you so much for enduring and understanding my frequent absence from home. Many birthdays celebrated over the phone. I appreciate your amazing love and support. Mama you always discouraged the thought of me giving up. Thank you so much.

To all my colleagues, thank you very much for your timely and much needed help each time I solicited for it. I am grateful.

Last but not least I wish to thank my pastor Dr. K.A Kyei and my sister Vhodaho Nevondo for their prayers and encouragement; my brother Israel Kehinde Ekanade, you never stopped praying for me. You always encouraged me to seek strength in the Lord; Happiness Sizakele Mathenjwa you have been a pillar of strength all throughout, always cheering me on and praying for me; my best friends Tebogo Mocheki and Hlayisani Pataka, thank you so much for your love and support. And to all I forgot to mention, your support was very valuable to me and I will be eternally grateful for that.

List of abbreviations

CRS	Penalised cubic regression splines
AIC	Akaike information criterion
BIC	Bayesian information criterion
CV	Cross validation
GVC	Generalised cross validation
GLMM	Generalised Linear Mixed Models
Lasso	Least absolute shrinkage and selection operator
MSE	Mean square error
OCV	Ordinary cross validation
PIRLS	Penalised iteratively reweighted least squares
PS	P-spline
RMSE	Root mean square error
REML	Restricted maximum likelihood
ROC	Receiver operating curves
TPS	Thin plate regression spline

Chapter 1

Introduction

1.1 Introduction

Suppose a researcher is interested in analysing the time T that it takes for an event of interest to occur as opposed to investigating whether or not an event has occurred. Or suppose the probability of surviving beyond a certain point in time is of more interest than the expected time of an event. Both cases typify the analysis of time-to-event data which is also known as survival analysis.

Survival analysis is a collection of techniques used to study the duration until an event of interest occurs, where duration is observed from the time at which a subject is exposed to the risk of experiencing the event (Langova, 2008). For example, a respondent is at risk for divorce from the time they marry. However, in some cases, it is not obvious when the risk period begins and the researcher must decide on the appropriate origin. For instance, in the study of age at sexual debut one may choose age at menarche as the start of the risk period. And in societies where pre-marital sex is a taboo, age at first marriage would be a natural origin.

Survival data possesses two types of features, namely, censoring and time varying explanatory variables. Censoring arises when there are subjects who have incomplete information in the study. The most common form of incomplete observation is right censoring, where the individual either drops out of the study or the study ends before he/she experiences the event of interest. Suppose T_i and C_i denote the event time and censoring time respectively for the i^{th} person of a sample of n and let $(t_i = \min\{T_i, C_i\})$. The censoring indicator δ_i is defined as:

$$\delta_i = \begin{cases} 1, & \text{if } T_i \leq C_i \text{ i.e. if a subject experiences the event at time } t; \\ 0, & \text{if } T_i > C_i \text{ i.e. if a subject survives beyond } t. \end{cases}$$

The other form of censoring is called left censoring or left truncation, where a respondent has already experienced the event of interest before the observation period. Interval censoring occurs when the event time T is not known precisely but known to fall within an interval.

Since survival data is collected over a finite period, the values of the covariates have the propensity to change. When studying age at first alcohol intake, we might be interested in the relationship between a respondent's probability of initiating alcohol intake at time t and their education status. In this case education status is an example of a time-varying covariate because the level of education might change with time.

The use of orthodox statistical techniques such as regression analysis to model survival data may be inappropriate as they may produce large biases or they may lead to a loss of information as they do not possess the versatility to handle censoring and time-variant covariates. Survival methods are more appropriate to use because they can account for censoring and also incorporate time-variant covariates in the models.

Hazard functions and survival functions are two of the most common summaries used to examine the distribution of events in survival analysis. The hazard at time $T = t$ is defined as

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}, \tag{1.1.1}$$

where the numerator represents the probability that an event occurs during the interval $(t, t + \Delta t)$ and Δt is the width of the interval which gets infinitely small. The hazard function measures the probability that an event occurs in a very small interval of time given that the outcome event has not occurred before t. This function is also known as the instantaneous rate, transition rate, hazard rate or failure rate. The survivor function on the other hand considers the probability that an individual lives longer than a specified time period t since the beginning of the study and is defined as:

$$S(t) = Pr(T > t). \tag{1.1.2}$$

Respondents who have not experienced the event of interest until the observation period ends are said to have 'survived'. The complement of a survival function is known as cumulative distribution function and it measures the probability that an outcome event occurs before t. It is formulated as follows:

$$F(t) = 1 - S(t) = Pr(T \leq t). \tag{1.1.3}$$

A number of techniques used for analysing event-history data assume time to be measured on a continuous scale, that is, events can occur at any given point in time. Although time can be viewed as a continuum, in practice time is a discrete measurement. For example, most surveys ask only the month or year in which an individual experienced the event of interest rather than the exact date. In such cases, survival methods that consider time as a continuous variable may be inappropriate.

Discrete-time models have more desirable features than continuous-time models. Firstly, the hazard functions in discrete-time models are defined as conditional probabilities. This makes interpretation easier. Although the hazard rate in continuous-time models is thought of as a probability, in actual fact it is not since it may be greater than 1.

Another problem with continuous-time models is that they assume that only one event can occur at any given point in time. The use of continuous time models may cause problems where tied events are observed. It is not so with the discrete-time models.

Discrete-time models can be embedded into the generalised linear model (GLM) framework. Therefore estimation can be achieved by using estimation procedures for GLMs (Tutz and Schmid, 2016; Allison, 1982).

The most widely used model for continuous-time survival data is the Cox's proportional hazards model in which the hazard at time t is expressed as a function of the covariates through the equation:

$$\lambda_c(t|\mathbf{x}) = \lambda_0(t)\exp(\mathbf{x}^T\boldsymbol{\beta}), \tag{1.1.4}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function, $\mathbf{x}^T$ is a $n \times p$ matrix of explanatory variables and $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of parameter estimates (Cox, 1972). The Cox's model is termed a proportional hazard model because it has the implicit assumption that the ratio of the hazard rates for any two individuals is constant over time. This implies that the effects of the explanatory variables may be constant across duration time. This assumption is unrealistic because in practice the effect of the explanatory variables, such as the academic level of a respondent, on the occurrence of an event may vary over time. With discrete-time survival models the restrictive assumption of proportional hazards can be easily avoided (Hess and Persson, 2010 and Singer and Willet,1993).

Considering the advantages of using discrete survival models described above, this study aims to employ discrete-time survival techniques to analyse age at first alcohol at two tertiary institutions in Thohoyandou.

1.2 Problem Statement

For n given observations, $i = 1, 2, ..., n$, let T_i denote event time and C_i the censoring time for individual i. The event and censoring times, T_i and C_i, take values in {1,...,k} and are assumed to be independent. For discretised time, $i = 1, ..., k$ refers to $[0, a_1), [a_1, a_2), ..., [a_{k-1}, \infty)$ where $T_i = t$ means that the event has occurred in $[a_{t-1}, a_t)$. For right censored data, the observed time defined by $t_i = \min(T_i, C_i)$ denotes the true event time if $T_i < C_i$.

One of the fundamental tools to model discrete-time event data is the hazard function. For any given vector of explanatory variables $\mathbf{x}$'s, the hazard function is formulated as:

$$\lambda(t|\mathbf{x}_i) = Pr(T = t|T \geq t, \mathbf{x}_i), \qquad t = 1, ..., k \tag{1.2.1}$$

and it captures the probability that a subject will experience an event in interval (a_{t-1}, a_t) given that the subject has not experienced the event before the interval.

A parametric discrete-time survival model has the form:

$$\lambda(t|\mathbf{x}) = h(\lambda_0(t) + \mathbf{x}^T\boldsymbol{\beta}), \tag{1.2.2}$$

where $\lambda(t|\mathbf{x}_i)'$s stands for a discrete-time hazard which is the conditional probability of an event occurring within a specified interval given that the event has never occurred before, $h(\cdot)$ is a monotonically increasing distribution function, $\lambda_0(t)$ denotes the baseline hazard function which is defined as the hazard function obtained when all the explanatory variables are set to 0, $\mathbf{x}^T$ is a $(1 \times p)$ vector of explanatory variables and $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of unknown parameters. The parameters of the baseline hazard function are given by $\lambda_0(1), \lambda_0(2), ..., \lambda_0(q)$ where q represents the number of time points.

The most commonly used model in discrete-time survival modeling is the discrete version of the proportional hazards model developed by Cox (1972) and it takes the form:

$$\lambda(t|\mathbf{x}) = 1 - \exp[-\exp(\lambda_0(t) + \mathbf{x}^T\boldsymbol{\beta})], \tag{1.2.3}$$

where $\lambda(t|\mathbf{x})$

Another popular model is the logistic discrete hazard model or the continuation proportional

ratio model which can be written as

$$\lambda(t|\mathbf{x}) = \frac{\exp(\lambda_0(t) + \mathbf{x}^T\boldsymbol{\beta})}{1 + \exp(\lambda_0(t) + \mathbf{x}^T\boldsymbol{\beta})}. \tag{1.2.4}$$

The models of type (1.2.3) and (1.2.4) are based on the assumption each predictor has a linear effect on the transformed hazard. This assumption is too prohibitive since sometimes in practice the relationship between the predictor and response may be nonlinear. Most relevant assumption to this research is the commonly used assumption that the baseline hazard is represented by a separate intercept coefficient for each t. In cases where the number of time intervals is large relative to the size of the sample, the number of parameters will be large as well resulting in erroneous maximum likelihood estimates (Schmid and Berger, 2018 and Fahrmeir and Wagenpfeil, 1996).

To avoid this problem one may choose to use a simpler parametrization that contains a few parameters (Tutz and Schmid, 2016). To reduce the number of parameters in the model, Mantel and Hankey (1978) used a polynomial to replace the parameters of the baseline hazard function. Efron (1988) proposed the use of a cubic-linear spline to reduce the number of parameters in a model. Manda and Meyer (2015) modelled the baseline using a dynamic model with a random walk prior. The same model is discussed in Farhmeir and Knorr-held (1997). This study explores the use of penalised regression splines to flexibly model the baseline hazard function.

1.3 Aim and objectives

The main aim is to evaluate and compare the accuracy of alternative ways of modelling the baseline hazard function. To achieve the aim the following objectives are proposed:

- To compare the accuracy of penalised regression splines and the dummy variable approach in discrete survival models through a simulation study.
- To develop and evaluate discrete survival model for age at first alcohol intake with the baseline hazard function modeled using penalised regression splines and dummy variables.

1.4 Layout of the project

The rest of the dissertation is organised as follows. Chapter 2 presents a discussion on the different ways of modelling the baseline hazard function, commonly used performance measures

for assessing discrete survival models and a short review on the timing of the onset of alcohol among the students residing in South Africa. The chapter that follows, Chapter 3, describes the theoretical background of the models considered for data analysis. Chapters 4 and 5 cover a discussion of the results obtained from the simulation study and real data application, respectively. And in the final chapter, Chapter 6, the conclusions derived from the study as well as the limitations of the study are given.

Chapter 2

Literature Survey

2.1 Introduction

In this chapter the researcher surveys relevant literature on different specifications of the baseline discrete-time hazard functions, penalised regression splines as ideal smoothers and the diagnostic tools in discrete-time survival model. A short review of duration to first alcohol intake among South African students is also presented.

2.2 Modeling of the baseline hazard function in discrete time survival models

There exists quite a number of (Bayesian and non-Bayesian) approaches that can be applied in smoothing the baseline hazard function. For example, some of the Bayesian techniques include the use of multivariate priors with a common mean (Campolieti, 2002) or Bayesian P-spline priors (Adebayo and Fahrmeir, 2005) to smooth the baseline hazard. Manda and Meyer (2005) considered a random walk prior with a variance that depends on the length of the time period to model the baseline hazard function in their study.

Within the non-Bayesian framework, piecewise polynomials and spline based approaches are commonly used in practice (Campolieti, 2002 and Rizopoulos, 2012). Mantel and Hankey (1978) and Biggeri et al. (2001) substituted the baseline hazard function by a polynomial whereas Efron (1988) and Berger and Schmid (2018) considered different classes of the spline approach. In particular, Efron (1988) employed a cubic regression spline while Berger and Schmid (2018) applied a p-spline, which is a type of penalised regression splines, to model the baseline hazard

function.

Splines are pieces of polynomials joined together at points referred to as knots. Regression splines are appealingly simple in practice but the choice of the number and location of the knots heavily influence the modelling results. As a result knot selection is a serious challenge in regression spline modeling (Wood and Augustin, 2002). Smoothing splines overcome this impediment by utilizing more knots which are located at every unique data point while at the same time including a penalty term that penalises wiggliness of the curve. Penalised regression splines seem to be an even better alternative in that though they are similar to smoothing splines in choosing a substantial number of knots and adding a penalty term to penalise over-smoothing in the model, they use a smaller knot set. Because the knots are not placed at every specific data value in this approach, this diminishes the problem of having many parameters in the model caused by too many knots and thereby reducing the computational cost of the fitted model (Wood, 2006).

The smoothing parameter controls the balance between goodness of fit and the smoothness of the model. A very high value for the smoothing parameter results in over smoothing of the data and if it is very low then the data will be under smoothed. Each case implies that the spline estimate is far from the actual function. Ordinary cross validation (OCV) and generalised cross validation (GCV) are some of the criteria which can be used to choose a smoothing parameter in such a way that the estimated spline function is closer to the true function. Low values of the criteria indicates a good model fit and high values imply a poor model fit.

The main idea behind an OCV is to leave out one of the data value and then fit the model to the remaining data and then calculate the squared difference between the left out datum and its predicted value. This is repeated for each data point and then the average squared difference is calculated to obtain the OCV score.

There is a more efficient way to calculate OCV without refitting the model to each of the data value that is left out. This can be obtained by writing the OCV statistic as a weighted sum of the residual of the model. A generalised cross validation (GCV) score is then acquired by replacing the individual weights in this summation by the average weight. Wood and Augustin (2002) postulate that the GCV score is better than OCV in that it is more computationally efficient and the smoothing parameters obtained by this method have the ability to minimise

the mean square prediction error in large samples. The GCV is a prediction error criterion that is used when the scale parameter is not known. In cases where the scale parameter is known, Un-Biased Risk Estimator (UBRE) or Mallow's Cp is used as criterion that selects a smoothing parameter that minimizes the expected mean square error (MSE). Restricted maximum likelihood (REML) can also be used to select smoothness parameter of a model. This approach treats the smoothing functions as random effects in order that the smoothing parameters are variance parameter to be estimated by maximizing the log-likelihood (Wood, 2011 and Reiss and Ogden, 2009).

(Wood and Augustin, 2002; Wood, 2011 and Wood, 2006) have shown that generalized additive models (GAMs) can be constructed using penalised regression splines. The smoothing methods can be implemented using the mgcv package in R-software. The smooth terms in the mgcv package are represented using penalised regression splines. Each smooth term is represented by a set of basis function and has an accompanying penalty that measures its wiggliness.

2.3 Goodness of fit measures and model diagnostics for discrete-time survival models

2.3.1 Goodness of fit measures

Typically researchers want to assess the performance of their estimated models. To do this they employ several performance measure that judge the fit of a model. If the models are nested one can simply compare the models by using likelihood ratio tests which compare the models by taking twice the absolute difference in the log-likelihoods for the models. These kind of measures can reveal whether the complicated model is significantly "better" than the simpler model. If the models are not nested, the Akaike information criteria (AIC) and Bayesian information criterion (BIC) can be used. The main idea behind these criteria is to penalise the log-likelihood for the number of independent variables in the model (Allison, 1995).

Another way to evaluate the significance of predictors in a model is to use martingale residuals. The main idea behind a martingale residual is to compare for each subject the observed number of events with the expected number of events up to time $\tilde{T}$ which is defined as $\tilde{T} = min(T_i, C_i)$

(Tutz and Schmid, 2016 and Berger and Schmid, 2018). In the study of duration of unemployment, Berger and Schmid (2018) employed martingale residuals to assess the influence of the explanatory variables on the hazard.

2.3.2 Model diagnostics

Berger and Schmid (2018) have shown that calibration can be used to assess the goodness of fit of a discrete time survival model. Calibration refers to how adequately predicted probabilities are in agreement with the observed frequency of events. A model is well calibrated when the predictions fall perfectly or closer to a 45 degree line.

Discrimination measures also provide another way to assess the performance of a model by evaluating the prediction accuracy of a model. These measures differentiate between individuals that experience an event at a particular point in time and those who have not experienced the event (Heagerty and Zheng, 2005 and Schmid et al., 2018). The discrimination measures consider the outcome as time-dependent dichotomous variable with levels defined as "event" and "no event".

The discriminative power can then be evaluated by constructing time-dependent sensitivity (true positive rate) and specificity rate (false positive) and Receiver Operating Curves (ROC curves) for each threshold of the predictor (Schmid et al., 2018). The ROC curve plots the sensitivity rates against the specificity rates for different thresholds. Computing the areas under ROC curve produces a time-dependent AUC curve which can also be employed to estimate the discriminative ability of a survival model at each time point.

Following the approach by Heagerty and Zheng (2005), Schmid et al. (2018) have shown that the AUC curve can be summarised into a global concordance index known as C-index which can also estimate the discriminative power of a model. This statistic measures the probability that the observations with large values of the predictor have shorter survival times than observations with small values of the predictor. Schmid et al. (2018) summarised the predictive performance of the model by using the time dependent AUC curve and C-index.

2.4 Age at first alcohol intake among students in South Africa

Despite different intervention strategies introduced by the South African government such as "increasing tax on alcohol" and "implementation of laws that forbid underage persons to drink alcohol", alcohol consumption among the youth in this country is on the rise, especially among high school students and university students (Ramsoomar and Morojele, 2012; Kyei and Ramagoma, 2013; Tayob and Van der Heever, 2015 and Tshitangano and Tosin, 2016).

Nkhoma and Maforah (1994) found that three quarters of students residing in a self-catering residence at the university of Cape Town were consuming alcohol. In a study conducted at the university of Venda, 65 percent of the students are reported to have been drinking alcohol and almost half of the students in the sample confessed to abusing the substance (Kyei and Ramagoma, 2013).

High level of alcohol intake has also been reported among high schoolers. Chauke et al. (2015) showed that 35% and 29.7% of male and female respondents respectively used alcohol in a study conducted at Makhosini Secondary School in Soshanguve township.

A number of existing literature on alcohol consumption in South Africa has been dedicated to estimating the prevalence and investigating contributory factors on alcohol use (Ramsoomar and Morojele, 2012; Kyei and Ramagoma, 2013; Tayob and Van der Heever, 2014; Tshitangano and Tosin, 2016 and Chauke et al., 2015). Understanding the effects of age at alcohol debut is of vital importance too. Experts have provided evidence that suggests that early onset of alcohol increases the risk of alcohol abuse and the development of serious challenges in the future including alcohol disorders (Liang and Chikritzhs, 2013 and Pitkanen et al., 2005).

A study conducted on US data from the 2010 National Survey on Drug and Health revealed that the earlier a person starts drinking is associated with an increased likelihood of heavy alcohol use in the future (Liang and Chikritzhs, 2013). DeWit et al. (2000) have shown that alcohol debut at ages 11 to 14 increases the risk of progression to the development of alcohol disorders. Some of reasons young people start drinking alcohol very early in life include seeking to be accepted by their peers and trying to solve typical challenges they experience in psychological growth (Pitkanen et al., 2005).

2.5 Concluding remarks

It is vitally important to smooth the baseline hazard function lest we obtain spurious estimates from the model results. Most of the studies on smoothing the baseline hazard by using splines focussed on just the application these splines whereas this study compares the accuracy of splines in smoothing the baseline hazard function.

To illustrate the proposed methods on real data, this study considers alcohol intake data which was also analysed by Sithuba (2017). Sithuba focussed on determining the factors that explain the variation in age at first alcohol intake, whereas in this study the main aim is to check if different results are obtained after applying the different methods of modelling the baseline hazard. In the current study, we also present model evaluation results in the form of a calibration plot, discrimination measures and adjusted residuals. These were not given in Sithuba (2017).

Chapter 3

Methodology

Introduction

This chapter covers a detailed theoretical background of the proposed methodology. In the following, the discrete time survival models and the nonlinear transformations of the baseline hazard function, in particular, the penalised regression splines are considered. We also discuss the various performance measures that assess discriminative ability and calibration of a discrete time survival model. The chapter closes with a description of the simulation scheme that will be used to assess and compare the accuracy of the splines and how these methods are going to be applied on real data.

3.1 Discrete-time hazard function

For n given observations, $i = 1, 2, ..., n$, let T_i denote event time and C_i the censoring time for individual i. The event and censoring times, T_i and C_i, take values in {1,...,k} and are assumed to be independent. For discretised time, $i = 1, ..., k$ refers to $[0, a_1), [a_1, a_2), ..., [a_{k-1}, \infty)$ where $T_i = t$ means that the event has occurred in $[a_{t-1}, a_t)$. For right censored data, the observed time defined by $t_i = \min(T_i, C_i)$ denotes the true event time if $T_i < C_i$.

The hazard function and survivor functions are two main tools to model survival data. A discrete hazard function given by

$$\lambda(t|\mathbf{x}_i) = Pr(T = t|T \geq t, \mathbf{x}_i), \qquad t = 1, ..., k \tag{3.1.1}$$

measures the probability of an event occurrence at time t provided that it has not occurred

before. The corresponding survival function defined as

$$S(t|\mathbf{x}_i) = Pr(T > t, \mathbf{x}_i) = \prod_{s=1}^{t}(1 - \lambda(s|\mathbf{x}_i)), \qquad t = 1, ..., k \tag{3.1.2}$$

quantifies the probability that an event has taken place later than time t. Conditional on $T_i \geq t$, the discrete hazard function in equation (3.1.1) can be seen as a dichotomous variable that distinguishes between the event taking place at time t or not, for a fixed time t. Therefore regression modelling strategies for a dichotomous response variable can be used to derive a model for the discrete hazard function (Berger and Schmid, 2018).

A parametric method for expressing the hazard function as a function of the covariates is of the form:

$$\lambda(t|\mathbf{x}_i) = h(\lambda_{0t} + \mathbf{x}_i^T\boldsymbol{\beta}), \tag{3.1.3}$$

where $h(\cdot)$ is a fixed response function assumed to be increasing monotonically. It links the hazard with a linear predictor $\eta_{it} = \lambda_{0t} + \mathbf{x}^T\boldsymbol{\beta}$. The λ_{0t}'s are $t = 1, 2, ..., k-1$ time dependent intercepts, $\mathbf{x}^T$ is a $(n \times p)$ matrix of covariates and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ represents a vector of parameter estimates independent of t. The set of intercepts $\lambda_{01}, ..., \lambda_{0,k-1}$, known as the baseline hazard function, describes the hazard function when the covariates are set to zero.

The most widely used models of type equation (3.1.3) are the discrete version of the Cox proportional hazard model and proportional continuation ratio model (commonly known as logistic discrete hazard model). The discretized Cox proportional hazard model is defined as

$$\lambda(t|\mathbf{x}_i) = 1 - \exp[-\exp(\lambda_0(t) + \mathbf{x}_i^T\beta)] \tag{3.1.4}$$

with baseline effects

$$\lambda_{0t} = \int_{a_0}^{\infty} \lambda_u du$$

derived from the baseline function $\lambda_0(u)$ (Kalbfleisch and Prentice, 1980). Alternative formulation of equation (3.1.4) yields a complementary log-log model:

$$\log\{\log[1 - \lambda(t|\mathbf{x}_i]\} = \lambda_{0t} + \mathbf{x}_i^T\boldsymbol{\beta}. \tag{3.1.5}$$

Another model popularly used to model the discrete hazard is the logistic model. The model is specified as

$$\lambda(t|\mathbf{x}_i) = \frac{\exp(\lambda_{0t} + \mathbf{x}_i^T\boldsymbol{\beta})}{1 + (\exp(\lambda_{0t} + \mathbf{x}_i^T\boldsymbol{\beta}))}. \tag{3.1.6}$$

Taking the log on both sides of the equation we achieve the following:

$$\log\left(\frac{P(T_i = t|\mathbf{x}_i)}{P(T_i > t|\mathbf{x}_i)}\right) = \lambda_{0t} + \mathbf{x}_i\boldsymbol{\beta}, \tag{3.1.7}$$

where the ratio $\frac{P(T_i=t|\mathbf{X}_i)}{P(T_i>t|\mathbf{X}_i)}$ measures a comparison of the probability of an event at time t to the probability of an event later than t is quantified by the ratio.

These models assume that the coefficients of the baseline hazard function and parameters are fixed effects. This assumption is only appropriate if the number of intervals is small. It should be noted that the number of parameters in the model is determined by the number of intervals. Consequently if the number of intervals is large relative to the sample size then the number of parameters will be very high usually leading to spurious maximum likelihood estimates.

To avoid numerical problems associated with the estimation of the baseline hazard, one may consider using an additive model that has the form:

$$\lambda(t|\mathbf{x}_{it}) = h(s_0(t) + \mathbf{x}_i^T\boldsymbol{\beta}), \tag{3.1.8}$$

where $s_0(t)$ is a smooth function of time. A model with fewer number of parameters can be achieved by relating the values of the baseline hazard at neighbouring time points through $s_0(t)$ and the problem of low counts of events at some time points is circumvented (Berger and Schmid, 2018 and Tutz and Schmid, 2016). One of the popular ways to represent the smooth function in t is to use splines, which are defined by a weighted sum of m basis. This study focuses on flexible spline functions that can be procured by selecting a relatively large number of m basis functions and simultaneously adding a penalty term to the model objective to prevent the estimates from being too wiggly. This approach is referred to as the penalised regression spline approach.

3.2 Penalised regression splines

The smoothing methods explored in this study are penalised regression smoothers which are constructed using basis functions. Prior to defining penalised regression smoothing methods, a description of what basis functions are is given. To define a basis function $c(x)$, suppose that $g(x)$ represents a univariate function and $\{c_i(x) : i = 1, ..., d\}$ set of functions, then $g(x)$ can be defined as

$$g(x) = \sum_{i=1}^{d} \alpha_i c_i(x), \tag{3.2.1}$$

where α_i's are d unknown parameters. Then $g(x)$ is comprised of a linear combination of the basis functions $c_i(x)$. The most commonly used bases are polynomial bases and spline bases. Polynomial bases are appealing to use in practice but their theoretic approximation and numerical stability properties are poor (Wood and Augustin, 2002). For example, suppose that g is a fourth order polynomial, then the basis functions are given by: $c_1(x) = 1$, $c_2(x) = x$, $c_3 = x^2$, $c_4(x) = x^3$ and $c_5(x) = x^4$. Then Equation (3.2.1) becomes

$$g(x) = \sum_{i=1}^{5} \alpha_i c_i(x), \tag{3.2.2}$$

which is a sophisticated way of writing

$$g(x) = \alpha_1 + \alpha_2 x + \alpha_3 x^2 + \alpha_4 x^3 + \alpha_5 x^4.$$

A major drawback of these types of bases is that an increase in basis dimension increases the collinearity of the basis functions, consequently leading to parameter estimators that are highly correlated, usually resulting in high estimator variance and numerical challenges (Wood and Augustin, 2002; Wood, 2011).

Our focus in this study is cubic spline bases which have proved to have good theoretical and numerical properties (Wood, 2011). Consider again, the problem of estimating $g(x)$ and suppose that $\{x_i^\star : i = 1, ..., d\}$ denotes a set of points referred to as knots. Cubic splines are defined as piecewise polynomials joined at the knots and should also be continuous to second derivative at the knots. These types of bases are popular because they offer greater flexibility for fitting data and are visually smooth because of the first and second derivatives (Durrleman and Simon, 1989). A simple mathematical representation of a cubic spline basis is achieved by denoting $c_m(x) = |x - x_i^\star|^3$ for $i = 1, 2, ..., d$, $c_{i+1}(x) = 1$, $c_{i+2}(x) = x$ and therefore $g(x)$ is defined as

$$g(x) = \sum_{i=1}^{d+2} \alpha_i c_i(x).$$

Generally spline based approaches can be classified into regression splines, smoothing splines and penalised regression splines. The regression splines are joined at predetermined set of knots usually placed at quantiles of data. This type of spline is relatively simple in practice, however, it is highly sensitive to the number as well as the location of the knots.

A conventional approach of avoiding problems associated with knot placement is to use smoothing splines or penalised regression splines. A smoothing spline approach uses a relatively large

number of knots that are placed at every data point but at the same time a penalty term that penalises curvature is incorporated in the model fit objective function. This approach seems to be somewhat of an ideal smoother. However, the only fundamental problem is that they have as many parameters as the data that are to be smoothed. As a result it is computationally expensive to estimate this type of spline.

Penalised regression splines are on the other hand a better alternative in that they retain the good properties of smoothing splines but with a reduced knot set to lower down the computational expense (Roshani and Ghaderi, 2016; Govindarajulu et al., 2007; Lian and Luan, 2002; Gray, 2016; Eisen et al., 2004; Wood and Augustin, 2002). In the subsequent sections, three variants of spline based penalised regression smoothers are discussed. These techniques are implemented in the R package mgcv (Wood, 2011). The *mgcv* package contains different options of basis functions.

3.2.1 Penalised cubic regression splines (CRS)

Consider modeling data using the following model

$$\lambda(t|\mathbf{x}) = h(s_0(t) + \mathbf{x}^T\boldsymbol{\beta})$$

where $s_0(\cdot)$ is a smooth function which can be represented using penalised cubic regression splines so that

$$s_0(t) = \sum_{s=1}^{m} \lambda_{0h} \cdot \gamma_h(t). \tag{3.2.3}$$

where $\gamma_h(t)$'s are the cubic spline basis and λ_{0h}'s are knot coefficients. Penalised cubic regression splines are knot based approximations on the knots $t_1 < ... < t_{m_s}$ where $m_s = m - 1$ which minimize

$$\sum (s_{0t} - \hat{s}_{0t})^2 + \theta \int \hat{s}''_{0t} dt. \tag{3.2.4}$$

where θ is a smoothing parameter and $\hat{s}_{0t}$ is the estimated spline. To select penalised cubic regression spline to represent the baseline hazard we use s(bs = "cr", k), where k represents the dimension of the basis.

3.2.2 P-splines (PS)

P-splines are low rank polynomial splines that are conventionally defined on equally spaced knots in which a difference penalty is directly applied to the parameters to control wiggliness

of the function (Wood, 2003; 2011 and Eisen et al, 2004).

P-splines can be implemented using polynomial splines in the form of truncated power series basis or B-splines. Suppose that the parameters of $\lambda_0(t)$ are approximated by a weighted sum of m bases. Polynomial splines are obtained by dividing the time domain into continuous intervals $[k_i, k_{i+1}]$. The polynomial splines are joined smoothly at the knots $k_1 < k_2 < ... < k_{m_s}$, where $m_s = m - 1$, which determine the boundaries of the intervals. A polynomial spline of degree d is the truncated power series defined as

$$s_0(t) = \lambda_{00} + \lambda_{01}t + ... + \lambda_{0d}t^d + \sum_{i=1}^{K} \lambda_{s+i}(t - k_K)_+^d \tag{3.2.5}$$

where $(t - k_i)_+^d$ are truncated power functions defined by:

$$(t-k)_+^d = \begin{cases} (t-k)^d & \text{if } t \geq k; \\ 0 & \text{if } t < k. \end{cases}$$

where $1, t, ..., t^d; (t - k_1)_+^d + ... + (t - k_K)_+^d$ are the basis functions used in equation (3.2.5).

Another representation of the polynomial splines is by B-spline basis functions. To define a B-spline basis with d parameters, we divide the time domain into $d + m + 1$ knots, such that, $t_1 < ... < t_{d+m+1}$ and the spline will be evaluated within the interval $[t_{m+2}, t_d]$. A spline of order $(m + 1)$ is given by:

$$s_0(t) = \sum_{i=1}^{d} \lambda_{0h}^m(t)\gamma_h(t). \tag{3.2.6}$$

where the B-spline basis functions are defined recursively as

$$\lambda_{0h}^m(t) = \frac{t - t_h}{t_{m_h+m+1} - t_h}\lambda_{0h}^{m-1}(t) + \frac{t_{h+m+2} - t}{t_{h+m+2} - t_{h+1}}\lambda_{h+1}^{m-1}(t); \quad i = 1, ..., d,$$

where

$$\lambda_{0h}^{-1} = \begin{cases} 1, & \text{if } t_h \leq t \leq t_{h+1}; \\ 0, & \text{otherwise.} \end{cases}$$

is a B-spline of order 1, and a second difference penalty given by:

$$\rho = \sum_{h=1}^{d-1} (\gamma_{h+1} - \gamma_h(t))^2$$

is applied directly to the knot coefficients to control wiggliness (Eilers and Marx, 1996); Wood, 2011). B-splines are more popular in the application of smoothing techniques since they allow for a more local fit to data than polynomials (Roshani and Ghaderi, 2016; Wood, 2003; Eisen et al., 2004). The locality of B-spline function is more suitable because it leads to a more increased

numerical stability (Tutz and Schmid, 2016; Eisen et al., 2004). The P-spline has shown some desirable properties such as the ability to conserve moments and they can fit polynomial data very well (Eilers and Marx, 1996).

3.2.3 Thin plate regression spline (TPRS)

The basis functions considered in the previous section are one dimensional basis. In this section we present a multi dimensional basis, more specifically the thin plate (regression) spline. Prior to describing a thin plate regression spline, we first have to explain what the thine plate spline. The thin plate regression spline is a special form of a thin plate spline. Suppose that the knots are represented by $\{t_h^\star : h = 1, ..., n\}$, a thin plate spline based on knot approximation can be calculated as

$$s_0(t) = \sum_{i=1}^{n} \xi_i \gamma_h(t) + \sum_{i=1}^{M} \alpha_i \omega_h(t) \tag{3.2.7}$$

where $\gamma_h(t) = \eta_{md}(||\mathbf{t} - \mathbf{t}_h^\star||)$ and $\omega_h(\cdot)$ are known basis functions and where ξ_i and α_i are unknown parameter estimates and $\boldsymbol{\xi}$ is subject to the constraint $\mathbf{T}^T\boldsymbol{\xi} = \mathbf{0}$, where $n \times M$ matrix and $\mathbf{T}$ has elements $T_{ih} = \omega_h(\mathbf{t}_i)$.

Suppose that the matrix $\mathbf{E}$ by $E_{ih} = \eta_{md}(||\mathbf{t}_i^\star - \mathbf{t}_h^\star||)$, therefore the thin plate spline fitting objective becomes

$$\text{minimize}||\boldsymbol{\lambda}_{0t} - \mathbf{E}\boldsymbol{\xi} - \mathbf{T}\boldsymbol{\alpha}||^2 + \boldsymbol{\theta}\boldsymbol{\xi}\mathbf{E}^T\boldsymbol{\xi} \tag{3.2.8}$$

subject to the constraint $\mathbf{T}^T\boldsymbol{\xi} = \mathbf{0}$, $\mathbf{E}$ with respect to $\boldsymbol{\xi}$ and $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ is a smoothing parameter.

This type of spline is something of an ideal smoother in that it does not have problems related to choosing knot positions and basis functions and it also has the capability of handling multiple predictors and allow the researcher to choose the order of derivatives used in the measure of the wiggliness function. However, it is relatively computationally expensive to estimate because they have as many unknown parameters as there are data (Wood, 2006).

Thin plate regression splines are built from basis and penalty in the full thin plate splines and optimally truncate the basis to achieve a low rank smoother. Set $\mathbf{E} = \mathbf{BUB}^T$ to represent a spectral decomposition of $\mathbf{E}$ such that $\mathbf{U}$ is a diagonal matrix of the eigenvalues of $\mathbf{E}$ so that $|U_{i,i}| \geq |U_{i-1}, U_{i-1}|$ and the columns of $\mathbf{B}$ are the eigenvectors. Suppose that $\mathbf{B}_k$ is the matrix that contains the first k columns of $\mathbf{B}$ and $\mathbf{U}_k$ represents the $k \times k$ submatrix of $\mathbf{U}$. Now, we

restrict $\boldsymbol{\xi}$ to the column space of $\mathbf{B}_k$ by writing $\boldsymbol{\xi} = \mathbf{B}_k\boldsymbol{\xi}_k$, this implies that equation (3.2.8) is

$$\text{minimize}||\boldsymbol{\lambda}_{0t} - \mathbf{EBU}\boldsymbol{\xi}_k - \mathbf{T}\boldsymbol{\alpha}||^2 + \boldsymbol{\theta}\boldsymbol{\xi}_k^T\mathbf{H}\boldsymbol{\xi}_k \tag{3.2.9}$$

subject to $\mathbf{TB}\boldsymbol{\xi}_k = \mathbf{0}$ with respect to $\boldsymbol{\xi}_k$ and $\boldsymbol{\alpha}$.

To rewrite Equation (3.2.9) in an unconstrained way, we firstly find an orthogonal column basis $\mathbf{H}_k$ such that $\mathbf{T}^T\mathbf{B}_k\mathbf{H}_k = \mathbf{0}$. This can be achieved by performing a **QR** decomposition of $\mathbf{B}^T\mathbf{T}$. Writing $\boldsymbol{\xi} = \mathbf{H}_k\tilde{\boldsymbol{\xi}}$ restricts $\boldsymbol{\xi}_k$ to the space and this results in a model that is not constrained and can be solved to fit the rank k approximation to the spline:

$$\text{minimize}||\boldsymbol{\lambda}_{0t} - \mathbf{B}_k\mathbf{U}_k\mathbf{H}_k\tilde{\boldsymbol{\xi}} - \mathbf{T}\boldsymbol{\alpha}||^2 + \theta\tilde{\boldsymbol{\xi}}^T\mathbf{H}_k^T\mathrm{U}_k\mathbf{H}_k\tilde{\boldsymbol{\delta}}$$

with respect to $\tilde{\boldsymbol{\xi}}$ and $\boldsymbol{\alpha}$. The eigenvalue decomposition of $\mathbf{E}$ to approximate the TPRS uses the Lanczos iteration (Wood, 2011;2012). This yields a lower computational cost than that of a full thine spline approach. Truncation and the changing of basis in the TPS can be seen as substituting $\mathbf{E}$ in the norm in Equation (3.2.7) by $\hat{\mathbf{E}} = \mathbf{EB}_k\mathbf{B}_k^T$, also replacing the $\mathbf{E}$ in the term that penalises curvature by replaced $\tilde{\mathbf{E}} = \mathbf{B}_k^T\mathbf{B}_k\mathbf{EB}_k\mathbf{B}_k^T$. The fitted values in the splines are estimated by $(\mathbf{E}\boldsymbol{\xi} + \hat{\mathbf{T}}\boldsymbol{\alpha})$ and so the quantity

$$\hat{e}_k = \max_{\boldsymbol{\xi}\neq\mathbf{0}} \frac{||(\mathbf{E} - \hat{\mathbf{E}}_k)\boldsymbol{\xi}||}{||\boldsymbol{\xi}||^2}.$$

describes the poorest possible deviations in the fitted values. The weight $||\boldsymbol{\xi}||^2$ is needed became the euclidean norm in the numerator is a maximum at infinity.

In similar fashion, a more fitting quantity that measures the poorest possible deviations in the curvature of the spline arising from truncation is

$$\tilde{e}_k = \max_{\boldsymbol{\xi}\neq\mathbf{0}} \frac{\boldsymbol{\xi}^T(\mathbf{E} - \tilde{\mathbf{E}}_k)\boldsymbol{\xi}}{||\boldsymbol{\xi}||^2}.$$

Minimising $\hat{e}_k$ and $\tilde{e}_k$ at the same time depends on the choice of $\mathbf{B}_k$ as the truncated basis for $\boldsymbol{\delta}$.

In a nutshell, thin plate regression splines are built from utilising the basis and the penalty term in the thin plate spline and then optimally truncate the basis to achieve a low rank smoother. One key advantage of this approach is that it avoids problems associated with placement of knots inherent in the classical regression spline models and also has the advantage of smooths of lower rank. This spline is fitted as $s(..., bs = "tp", ...)$ in R.

3.2.4 Choosing the basis function

As part of the model building process the researcher has to choose the basis dimension. The basis dimension should be large enough that the basis is more flexible than it is expected to be. Kim and Gu (2004) demonstrated that the size of the basis should scale as $n^{\frac{2}{9}}$, where n is the size of the sample. Based on their results from a simulation study they proposed that $10n^{\frac{2}{9}}$ can be used as the basis dimension.

The exact size of the basis of the dimension is not of critical importance because the basis dimension only sets an upper bound on the flexibility of a term. The actual effective degrees of freedom are controlled by the smoothing parameter. Consequently the fit of the spline is not sensitive to the model basis dimension given that it is not set prohibitively too low for the application concerned (Wood, 2011).

3.3 Estimation of parameters

3.3.1 Discrete time case

Let each observation be given by $(t_i, \delta_i, x_i), i = 1, 2, ..., n$ where $t_i = \min\{T_i, C_i\}$ denotes the minimum of the survival T_i and censoring C_i, with the censoring indicator variable defined as:

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

In the case where random censoring is assumed, T_i and C_i are independent random variables and the likelihood function of each observation i can be obtained as follows:

$$P(t_i, \delta_i|\mathbf{x}_i) = P(T_i = t)^{\delta_i} P(T_i > t_i)^{1-\delta_i} P(C_i \leq t_i)^{\delta_i} P(C_i = t_i)^{1-\delta_i}. \tag{3.3.1}$$

Assuming that the censoring contributions involving C_i are independent of parameters determining the survival time, referred to as non informative censoring, one obtains the following formula:

$$L_i = c_i P(T = t_i)^{\delta_i} P(T_i > t_i)^{1-\delta_i}, \tag{3.3.2}$$

where $c_i := P(C_i \geq t_i)^{\delta_i} P(C_i = t_i)^(1 - \delta_i)$. Dropping the terms involving C_i and using the discrete hazard and also incorporating the covariates yield the following likelihood function:

$$L_i = c_i \lambda(t_i|\mathbf{x}_i)^{\delta_i} (1 - \lambda(t_i|\mathbf{x}_i))^{t_i - \delta_i} \prod_{j=1}^{t_i - 1} (1 - \lambda(j|\mathbf{x}_i)) \tag{3.3.3}$$

With closer inspection, Equation (3.3.2) appears to be equivalent to the likelihood of a dichotomous response model that is defined by the following outcome variables

$$(y_{i1}, ..., y_{it_i}) = \begin{cases} (0, ..., 1), & \text{if } \delta_i = 1; \\ (0, ..., 0), & \text{if } \delta_i = 0. \end{cases}$$

For observations that experience the event of interest the observation vector is $(0, ..., 0, 1)$ of length t_i and for the censored observations the sequence becomes $(0, ..., 0)$. Consequently, the total log-likelihood of the proportional continuation ratio model is given by

$$\ell \propto \sum_{i=1}^{n} \sum_{s=1}^{t_i} y_{is} \log(\lambda(s|\mathbf{x}_i)) + (1 - y_{is}) \log(1 - \lambda(s|\mathbf{x}_i)). \tag{3.3.4}$$

The log-likelihood in Equation (3.3.4) is equal to the log-likelihood of a logistic regression model with the predictor $\eta_{it} = h(\mathbf{x}_{ij}^T \boldsymbol{\beta})$ where the values of the binary responses can be described as binary decisions for the transition from interval $[a_{s-1}, a_s)$ to $[a_s, a_{s+1})$(Tutz and Schmid, 2016 ; Schmid et al., 2018).

3.3.2 Smoothing case

Suppose that $\lambda_0(t)$ is expanded in m basis functions and denote the linear predictor as:

$$\eta_{it} = \lambda_0(t) + \mathbf{x}_i^T \boldsymbol{\beta},$$

expanding the baseline hazard as a sum of basis functions yields:

$$\eta_{it} = \sum_{s=1}^{m} \lambda_{0s} \phi_s(t) + \mathbf{x}_{it}^T \boldsymbol{\beta}, \tag{3.3.5}$$

where $\phi_1(\cdot), ..., \phi_m(\cdot)$ are basis functions of order q.

If a small number of basis functions are used then the fitting procedure for binary regression models can be used based on the log-likelihood specification in Equation (3.3.4). However if a great number of basis functions are used to achieve increased flexibility, estimation is based on the penalised log-likelihood approaches. The penalised likelihood estimation is given by:

$$l_p(\boldsymbol{\gamma}) = l(\boldsymbol{\gamma}) - \frac{1}{2} \theta J(\boldsymbol{\gamma}) \tag{3.3.6}$$

where $l(\boldsymbol{\gamma})$ is the log-likelihood function, θ represents the smoothing parameter and $J(\boldsymbol{\gamma})$ is a term that puts restrictions on the $\boldsymbol{\gamma}$ vector that collects parameter estimates in the model.

The restrictions placed on $\boldsymbol{\gamma}$ controls the 'wiggliness' of a smooth function by preventing the elements in $\boldsymbol{\gamma}$ from having relatively large deviances from the neighbouring elements (Tutz and Schmid, 2016).

Assuming that the basis functions are represented by B-spline functions that are defined on an equidistant grid, then in discrete time hazard models with smooth baseline hazard $\lambda_0(t)$ the parameter vector $\boldsymbol{\gamma}^T = (\lambda_{01}, ..., \lambda_{0m}; \boldsymbol{\beta}^T)$ contains the parameters corresponding to the baseline parameter $\lambda_0(t)$ and $\boldsymbol{\beta}$ respectively. A commonly used penalty is the difference penalty

$$J_\delta = \sum_{J=\delta+1}^{m} (\Delta^\delta \lambda_{0j})^2 \tag{3.3.7}$$

where Δ is the difference penalty on adjacent coefficients of the B-spline, i.e, $\Delta\lambda_{0j} = \lambda_{0j} - \lambda_{0j-1}$ and $\Delta^2\lambda_{0j} = \Delta(\lambda_{0j} - \lambda_{0j-1}) = \lambda_{0j} - 2\lambda_{0j-1} + \lambda_{0j+2}$. This difference penalty estimates the smooth parameters and the degree of smoothing is controlled by the smoothing parameter (θ).

3.3.3 Estimation criteria for smoothing parameter (θ)

Penalised likelihood maximization methods can only be used to the model coefficients $\boldsymbol{\gamma}$ given the smoothing parameter θ. The smoothing parameter θ controls the compromise between the goodness of fit of a model and the smoothness. Generally there are two classes of smoothing selection criteria. The first class optimizes the criteria by trying to minimize the prediction error of the model. Within this class, two techniques are useful in estimating θ: if the scale parameter is known then minimizing the expected mean square error leads to estimation by Mallow's C_p or Un-biased risk estimator (UBRE) and if the scale parameter is not known, trying to minimize prediction error will lead to cross validation or generalised cross validation (GCV).

The second class considers the smooth functions as random effects in such a way that the smoothing parameters θ's are treated as variance parameters and can be estimated using restricted maximum likelihood (REML) or marginal likelihood (ML) (Wood, 2006; Wood, 2017 and Wood et al., 2016).

1. **Known scale parameter: UBRE**

   Smoothing parameters should be chosen in such a way that the estimate ($\hat{\mu}$)is as close as possible to the parameter ($\mu \equiv E(y)$). A more fitting measure to summarise the distance

between $\hat{\mu}$ and the parameter μ is the expected Mean square error (MSE) of the model $\mathbb{E}(M)$:

$$\mathbb{E}(M) = E\left(\frac{||\mu - \mathbf{x}\hat{\boldsymbol{\beta}}||^2}{n}\right) = \frac{E(||\mathbf{y} - \mathbf{A}\mathbf{y})||)^2}{n - \sigma^2} + \frac{2tr(\sigma^2)}{n}. \tag{3.3.8}$$

Choosing a smoothing parameter which minimises the expected MSE implies that we want to minimise the un-biased Risk Estimator:

$$V_\mu(\theta) = \frac{||\mathbf{y} - \mathbf{A}\mathbf{y}||^2}{n - \sigma^2} + 2tr(\mathbf{A})\frac{\sigma^2}{n} \tag{3.3.9}$$

which is also known as Mallow's Cp (Wood, 2011). The right hand side of the Equation depends on the smoothing parameter through the influence matrix $\mathbf{A} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \sum_i \theta_i\mathbf{H})^{-1}\mathbf{X}^T$, where $\mathbf{H}$ is the penalty matrix and $tr(\mathbf{A})$ is the estimated degrees of freedom of the model.

2. **Unknown scale parameter: Cross Validation (CV)**

Equation (3.3.9) works well if σ^2 is known. In cases where the scale parameter is not known, estimation of the smoothing parameter is based on the mean square prediction error instead of the mean square error. The mean square prediction can be defined as the mean square error in predicting a new observation using the fitted model and is given by:

$$P = \sigma^2 + M. \tag{3.3.10}$$

To estimate P, cross validation (CV) can be used. The main idea of a CV is to leave out a datum, then fit a model to the remaining data and then compute the squared difference between the datum that was left out and the fitted model. This is repeated for each datum to obtain the squared difference between the missing data and the model fitted to the remaining data. The average squared differences obtained is referred to as the ordinary cross validation score (OCV)

$$V_o = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_i^{[-i]})^2 \tag{3.3.11}$$

where $\hat{\mu_i}^{-i}$ represents the prediction $\mathbb{E}(y_i)$ which is achieved from the model fitted with all the data without y_i.

Ordinary cross validation can be performed without refitting the model for each datum that was left out. This can be achieved by writing the OCV score as a weighted sum of

model residuals, where the weights can be calculated directly from the initial fit of all the data:

$$V_o = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{(1 - A_{ii})^2} \tag{3.3.12}$$

where A_{ii} is the diagonal element of the hat matrix $\mathbf{A}$. The OCV suffers from two major impediments; it is computationally costly to minimize in the additive case where there may be a number of smoothing parameters and also has a problem of lack of invariance (Wood, 2011).

Replacing the individual weights in the denominator of V_o by the average weights yields a score of the form

$$V_g = \frac{n||\mathbf{y} - \hat{\boldsymbol{\mu}}||^2}{[n - tr(\mathbf{A})]^2}. \tag{3.3.13}$$

This score is known as the generalized cross validation (GVC) (Wood, 2011).

In the general additive model fitting objective, the model is written in terms of the deviance

$$D(\beta) + \sum_{j=1}^{m}\theta_j\boldsymbol{\beta}^T\mathbf{S}_j\boldsymbol{\beta} \tag{3.3.14}$$

which is minimised with respect to $\boldsymbol{\beta}$. Given $\boldsymbol{\theta}$, Equation (3.3.14) can be approximated quadratically by:

$$||\sqrt{\mathbf{W}}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})||^2 + \sum_{j=1}^{m}\theta_j\boldsymbol{\beta}^T\mathbf{S}_j\boldsymbol{\beta} \tag{3.3.15}$$

where the approximation captures the dependence of the penalised deviance on $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ and $\mathbf{S}_j$ is the diagonal block in matrix containing only zero entries so that $\theta_j\boldsymbol{\beta}^T\mathbf{S}_j\boldsymbol{\beta}$ is the penalty matrix. Rewriting Equation (3.3.15), the GVC score for the smoothing parameter selection becomes

$$V_g^w = \frac{n||\sqrt{\mathbf{W}}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})||^2}{[n - tr(\mathbf{A})]^2} \tag{3.3.16}$$

where $\mathbf{W}$ and $\mathbf{z}$ are the iteratively re-weighted least squares weight and test data evaluated at convergence.

3. **Restricted maximum likelihood (REML)**

Several studies suggest the REML is a more preferable criterion for choosing smoothing parameters (Reiss and Ogden, 2009 ; Wood, 2006; Wood, 2017; Wood et al., 2016). This criterion have shown to penalize oversmoothing more severely than the prediction error

criteria. It is less prone to developing multiple minima for finite sample sizes and has a lesser tendency to produce variable θ estimates.

The REML method of estimation selects the smoothing parameter that optimizes the log of the joint density of the data and the model coefficients (Bayesian marginal likelihood)

$$V_{reml}(\boldsymbol{\theta}) = \log \int g(\mathbf{y}|\boldsymbol{\beta})g(\boldsymbol{\beta})d(\boldsymbol{\beta}).$$

Evaluating this integral by using Laplace approximation results in

$$V_{reml}(\boldsymbol{\theta}) \simeq l(\hat{\boldsymbol{\beta}}) - \frac{\hat{\boldsymbol{\beta}}^T \mathbf{S}_\theta \hat{\boldsymbol{\beta}}}{2\phi} - \frac{\log|\frac{\mathbf{S}_\theta}{\phi}|_+}{2} - \frac{\log|\frac{\mathbf{X}^T\mathbf{W}\mathbf{X}}{\phi} + \frac{\mathbf{S}_\theta}{\phi}|}{2} + \frac{M}{2}\log(2\pi), \qquad (3.3.17)$$

where l is the log likelihood, $\mathbf{W}$ diagonal matrix of Newton weights at convergence of the performance iteratively reweighted least squares (P-IRLS) iteration, M is the dimension of the null space of $\mathbf{S}_\theta$.

3.4 Model building

3.4.1 Variable selection

In cases where there are many predictor variables in a statistical model, the parameter estimates obtained are more likely to be unstable. Therefore it is imperative to employ variable selection methods to select important predictors to include in the final estimated model. To choose a set of relevant candidate models, we employ an algorithm for Generalized Linear Mixed Models (GLMMs) using L_1-penalisation developed by Schelldorfer et al. (2014). The approach is a combination of a gradient ascent optimisation and Fischer scoring algorithm and also incorporates an L_1 or least absolute shrinkage and selection operator (lasso) type penalty term. The L_1 penalty term has the ability to perform variable selection and shrinkage at the same time. The main idea behind the GLMMlasso technique is to maximize the log-likelihood of the model while simultaneously constraining the L_1-norm of the vector of parameter estimates. This results in all the coefficients shrinking to zero and some coefficients are actually set to zero. The coefficients that are set to zero imply that they have no effect on the dependant variable and thus a less complex model is obtained (Groll and Tutz, 2014).

3.4.2 Model selection criteria

Allison (1995) asserts that in cases where the models are nested, one can simply use likelihood ratio tests which compare the models by taking twice the absolute difference in the log-likelihoods for the models. These kind of statistics can reveal whether the complicated model is significantly "better" than the simpler model. In cases of non-nested models one can use the Akaike information criteria (AIC) and Bayesian information criterion (BIC) for model selection. The basic idea behind these criteria is to penalize the log-likelihood for the number of independent variables in the estimated models. The AIC was first introduced by Akaike (1974) and is given by:

$$\text{AIC} = -2\log(L_{max}) + 2k \tag{3.4.1}$$

where L_{max} is the maximum log-likelihood of the model and k is the number of parameters. The BIC criterion is very similar to Equation (3.4.1), however, it places a penalty on the number of estimable parameters (k). The criterion is estimated as follows:

$$\text{BIC} = -2\log(L_{max}) + k\log(n). \tag{3.4.2}$$

The "best" model is the one that has the lowest AIC or BIC scores.

3.4.3 Model diagnostics

1. **Discrimination measures**

   To define the discrimination measures, we first denote T as the event time and the predictor or risk score as $\eta = \mathbf{x}^T\boldsymbol{\beta}$. The discrimination measures evaluate the predictive accuracy of the risk score η by examining whether long survival times (i.e high values of T) are related to small values of η or vice versa (Schmid et al., 2018). Time to event response variables can be considered as time-dependent binary variables. Therefore methods used to measure predictive accuracy for binary classification rules can be used to assess the predictive ability of time-to-event outcomes (Schmid et al, 2018). Subjects that experience an event at $T = t$ are considered as cases whereas subjects that do not experience the event are considered as control for a fixed time t.

   Discriminative ability of a model can be measured by using time-invariant sensitivities (true positive rates) and specificities (false positive rates) defined as:

$$sens(c, t) := P(\eta > c|T = t), \tag{3.4.3}$$

and

$$spec(c,t) := P(\eta \leq c|T = t), \tag{3.4.4}$$

where c is the threshold of the predictor n (Heagerty and Zheng, 2005 and Tutz and Schmid, 2016 and Schmid et al., 2018). High sensitivity rates imply that the risk score (η) and the hazard ($\lambda(t|\mathbf{x})$) tend to be large which means that the cases are correctly specified. A large specificity on the other hand means that η and the hazard are small, which shows a correct identification of the control.

A time-dependent ROC (receiver operating curve) curve can be obtained by summarising the specificity and sensitivity rates for a fixed time t. The ROC curve is given by:

$$\text{ROC}(c,t) := \{1 - spec(c,t), sens(c,t)\}_{c\in\mathbb{R}} \tag{3.4.5}$$

The ROC curve plots sensitivity rates against specificity rates. A strongly concave and large areas below the curve indicate that η has a large discriminative power (Tutz and Schmid, 2016 and Schmid et al., 2018). The areas calculated under the ROC curve yields a time-dependent AUC curve of the form:

$$\text{AUC}(t) := P(\{\eta_i > \eta_j\}|\{T_i = t\} \cap \{T_j = t\}) \tag{3.4.6}$$

If η predicts better than chance, then AUC(t) should be greater than 0.5 for all time points. Heagerty and Zheng (2005) showed that integrating the AUC measure will yield a concordance index (C^*-index) that can be used to evaluate the predictive ability of the predictor for continuous event times. Following the approach by Heagerty and Zheng (2005), Tutz and Schmid (2018) derived a C-index for discrete event times given by:

$$C^* = \sum_{t=1} AUC(t) \cdot \frac{P(T = t) \cdot P(T > t)}{\sum_t P(T = t) \cdot P(T > t)}. \tag{3.4.7}$$

C^* measures the probability that observations with large values of η have shorter survival times that observations with small values of η. Similar to the AUC(t), the value of C^* should be greater than 0.5 if the risk score performs better than chance (Tutz and Schmid, 2016; Schmid et al., 2018). The discrimination measures are calculated in R by using the function *tprUno* (for time-dependent sensitivities), *fprUno* (time-dependent specificities) and *aucUno* (time-dependent AUC curves) which are readily available in *discsurv* package.

2. **Calibration**

Calibration measures diagnose a model by assessing how adequately predicted probabilities agree with their corresponding proportions of observed events. One of the ways to describe the comparison is by plotting a calibration plot. In a calibration plot the hazards ($\lambda(t|\mathbf{x}_i)$) estimated from a model are compared to the proportions of observed events ($y_{it} = 1$). Data is partition into subgroups, say D_k, where $k = 1, ..., K$ represents the number of subgroups which are determined by the by the percentiles of hazards estimated from the model. Popular choices for K are $K = 10$ or $K = 20$. Therefore, the proportion of observed events can be calculated in each group by

$$c = \sum_{\lambda(t|\mathbf{x}_i \in D_k} \frac{y_{it}}{|D_k|},, \qquad fori = 1, .., n, t = 1, ..., t_i$$

where $|D_k|$ represents the number of observations in the subset D_k.

If a model is well calibrated, the hazard obtained from the proportion of observed events should be as close as possible to the mean of estimated hazards in D_k for all the groups. On the plot, the predicted probabilities should lie on or close to 45 degree line if the model fit is satisfactory.

3.5 Application: Duration to alcohol intake

3.5.1 Data description

The data available in this study is based on the secondary data collected on the University of Venda and Vhembe Further Educatin and Training (FET) college students in 2017. The same dataset is discussed in Sithuba (2017). Both institutions are located in Thohoyandou, which is the seat of regional adminstration of Vhembe district municipality and Thulamela local municipality in the Limpopo province of South Africa. The questionnaire was administered to a total number of 744 male and female registered students at both institutions. Table (3.1) summarises the variables considered in the data collection. The outcome of variable in this study is duration to first alcohol intake, which was captured as a discrete variable. The respondents who have not consumed any alcohol are considered to be censored at their age on the date of the interview.

Table 3.1: This table contains the name and description of the variables considered in the data collection.

Predictor	Description
Stress	does the respondent have a history of stress; categorical: yes or no
Other drugs	respondents' intake of other drugs; categorical: yes or no
Drinking peers	does the respondent have drinking peers; categorical; yes or no
Negative life event	has the respondent ever experienced a negative life event; categorical: yes or no
Physically abused	was the respondent ever abused; categorical: yes or no
Parental alcohol intake	do the parents of the respondent take alcohol; categorical: yes or no
Sex	Sex of the respondent; categorical: male or female
Marital status	Is the respondent married or not, categorical: yes or no
Race	the respondent's race, categorical: African or other
Family income	the respondent's family income, categorical: $< \text{R}1000, \text{R}1000 - \text{R}4000, > \text{R}4000$
Family dysfunctional	Is the respondent's family dysfunctional; categorical: yes or no
Institution	the respondent's tertiary institution; categorical: Techniven, Univen
Relationship with adult	does the respondent have a relationship with an adult; categorical: yes or no
Siblings	the respondent's number of siblings; categorical: 0-3, 4-6,≥ 7
Parent's qualification	the qualification of the respondent's parent(s), categorical: none, matric, diploma, degree, honours
Childhood place of residence	the respondent's childhood place of residence; categorical: urban, rural, farmland
Welfare dependant	Is the respondent dependant on any social welfare; categorical: yes or no
Parental disciplinary	Is the respondent being disciplined by his/her parent; categorical: yes or no
Age at first alcohol intake	the age at which the respondent drank alcohol for the first time, continuous

3.5.2 Regression model

Prior to fitting the logistic discrete hazard model by using software for binary response data, the data has to be transformed into an augmented data matrix. To do this, the original discrete survival data is transformed into the corresponding binary representations. In R, the dataLong () function that is available in the *discsurv* package is used to convert the data. For a subject who experienced the outcome of event $\delta_i = 1$ at time t_i the augmented data matrix is of the form

$$\begin{array}{c} \textit{Binary observations} \quad \textit{Dummies} \quad \textit{Covariates} \quad \cdots \quad \textit{Covariates} \\ \begin{pmatrix} 0 & 1 & x_{i1} & \cdots & x_{ip} \\ 0 & 2 & x_{i1} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_i & x_{i1} & \cdots & x_{ip} \end{pmatrix} \end{array}$$

and when the individual has been censored ($\delta_i = 0$) the augmented matrix becomes

$$\begin{array}{c} \textit{Binary observations} \quad \textit{Dummies} \quad \textit{Covariates} \quad \cdots \quad \textit{Covariates} \\ \begin{pmatrix} 0 & 1 & x_{i1} & \cdots & x_{ip} \\ 0 & 2 & x_{i1} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & t_i & x_{i1} & \cdots & x_{ip} \end{pmatrix} \end{array}$$

The first column represents the binary responses $y_{i1}, ..., y_{it}$ and the second column is the time interval that starts from 1 to t_i. For fixed parameters λ_{0t}, this column is a nominal variable such as "dummy" variables. In this study the baseline parameter λ_{0t} is modeled using dummy variables and also using three variants of penalised regression splines. The remaining columns are the predictor effects. When the covariates do not change over time, the values from column 3 to $(p + 2)$ are similar.

3.6 Software

For all computations in this dissertation, the R statistical software which is publicly available via Cran (http://www.R-project.org/) was used in conjunction with the relevant packages. The R functions for discrete time survival modeling and various penalised regression approaches were implemented in the R - add on packages *discSurv* (Welchowski and Schmid, 2018) and *mgcv* (Wood, 2018) respectively. To transform the original data into binary data we use *dataLong* function. The gam() function available in the mgcv add-on package is used to the implement different penalised regression splines (Wood, 2015). Required arguments are type of spline smoother "*bs*", dimension of the basis k and m which represents the order of the penalty. The "*bs*" are chosen as $bs = (\text{"}cr\text{"}, \text{"}ps\text{"}, \text{"}tp\text{"})$ which is for a penalised cubic regression spline, p-spline and thin plate regression spline respectively. The *gam.check*() function in *mgcv* produces diagnostic information about the fitted penalised regression spline smoothers. The variable selection algorithm employed in this study was implemented using the R-package GLMMLasso.

Chapter 4

Simulation Results

Introduction

In this chapter, the researcher presents and discusses the simulation setup employed in the study as well as the results obtained from the simulation exercises. For the purpose of comparing the accuracy of penalised regression spline and the dummy variable techniques in modeling the baseline hazard function in a discrete survival model, this study considered two different functional forms of the baseline hazard. In the first simulation we considered a logarithmic baseline hazard and for simulation experiment 2 we used a gamma baseline hazard. For the penalised regression spline (CRS) the smoother was specified as s(bs="cr") and the chosen basis dimension was $k = 12$ for when $q = 12$, $k = 15$ for when $q = 15$ and $k = 20$ when $q = 20$. We used a p-spline (PS) specified as (bs ="ps") and a second order P-spline basis and with a third order difference penalty ($m = c(2, 3)$). The thin plate regression spline (TPS) was specified as ($bs = "tp"$) and a third order difference penalty ($m = 3$). For all the penalised regression smoothers, the optimal smoothing parameter θ was estimated by using REML. The tables presented below summarise the RMSE values obtained from the simulation experiments and the graphical comparisons of all the estimated models.

4.1 Simulation approach

This simulation study was carried out in an effort to compare the accuracy of penalised regression spline and the dummy variable approach in modelling the baseline hazard function. The data generating process used here was adopted from simulation schemes employed by Hess et al. (2014) and Tutz and Groll (2014). Data was simulated from a logistic hazard function of

the form

$$\lambda(t|\mathbf{x}_i) = \frac{\exp(\lambda_0(t) + x_1 + x_2)}{1 + \exp(\lambda_0(t) + x_1 + x_2)} \tag{4.1.1}$$

where $x_1 = x_2$ are generated as independent random variables from a standard normal distribution and standard gamma distribution respectively. For comparison purposes, the baseline hazard was specified as $\lambda_0(t) = -\log(t)$ and $\lambda_0(t) = 2\xi_t - 2.3$ where $\xi_t = f_\Gamma(t)$ with f_Γ being the density of a Gamma distribution with $\Gamma(\zeta, \theta)$. Shape and scale parameter from the gamma baseline hazard were chosen as $\zeta = 5$, $\theta = 1$ respectively. This will result in a baseline function with a hump as shown in the right hand side of Figure 4.1. All duration above $t = 12, 15, 20$ periods were considered to be right censored.

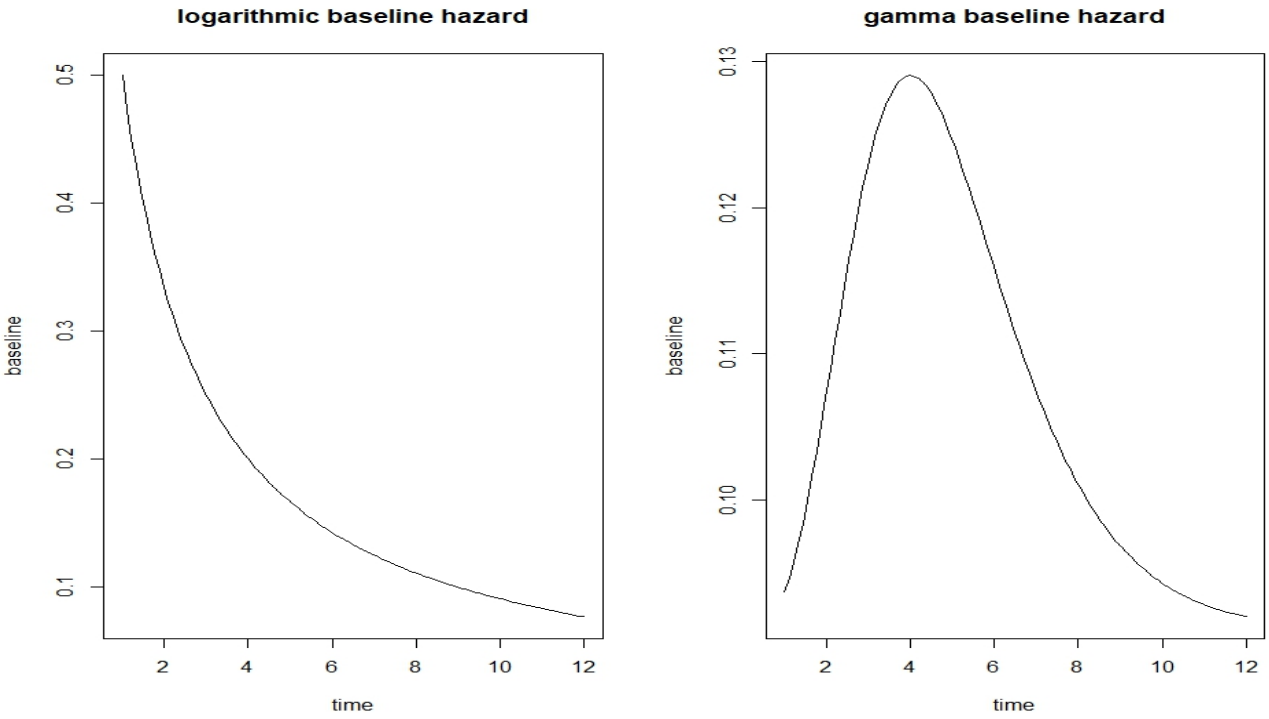


Figure 4.1: Shape of the baseline hazard function represented by $\lambda_{01}, ..., \lambda_{0q}$

To illustrate the performance of the penalised regression splines and the dummy variables in modelling the baseline hazard, the number of discrete time points (t) were varied relative to the sample size. For time period $t = 1, 2, ..., q$, where $q = 12, 15, 20$ denotes the last follow-up period for each subject, we simulated a binary variable y_{it} with values "0" and "1". When $y_{it} = 1$ in any time interval, the subject is considered to have experienced the event and every observation after that point onwards is not included in the dataset for that particular subject. A subject which survived until the last period (12, 15, 20) was considered to be censored at

the last period.

The sample sizes were varied and the maximum values of q in order to determine the effect of the ratio of the number of time periods to the sample size on the estimates. For each of the settings, 100 replicated datasets with 5 different sample sizes; $n = 50, 100, 500, 1000, 3000$ were simulated. Following the studies of Roshani and Daem (2016) and Strasak et al. (2011) where the expected mean square error was used to assess the performance of the fitted splines, we compute the square root of the mean square error (RMSE):

$$RMSE = \left[\frac{1}{n}\sum_{i=1}^{n}(\lambda_{0t} - \hat{\lambda}_{0t})^2\right]^{\frac{1}{2}}$$

to evaluate the goodness of fit for the fitted models. This measure quantifies the average departures of the estimated baseline hazard from the actual baseline hazard. Small values of RMSE indicate a "good" fit.

4.2 Results for the logarithmic baseline hazard

Table 4.1: The RMSE values for dummies, penalised cubic regression splines (CRS), P-splines (PS) and thin plate regression spline (TPS) for $n = 50$ and $q \in \{12, 15, 20\}$.

		Estimated models			
n	q	Dummies	CRS	TPS	PS
50	12	0.1170	**0.0558**	0.0813	0.0829
	15	0.1265	**0.0715**	0.0902	0.0906
	20	0.2592	**0.2019**	0.2049	0.2050

Table 4.1 presents the mean RMSE, which was used as performance criteria for the selection of the optimal model. The values in bold text represent the smallest recorded values from this setting. From this table we can deduce all the smoothing methods performed fairly better than the model where the baseline hazard was modeled using a dummy variable. The penalised cubic regression spline seemed to perform better than the other smoothing models because it recorded the smallest values of the RMSE in all the different settings. Figure 4.2 gives a graphical view of these results. The graphs give averages of 100 simulations. The blue line represents the true baseline hazard function.

Figure 4.2: Baseline hazards $n = 50$ and $q \in \{12, 15, 20\}$.

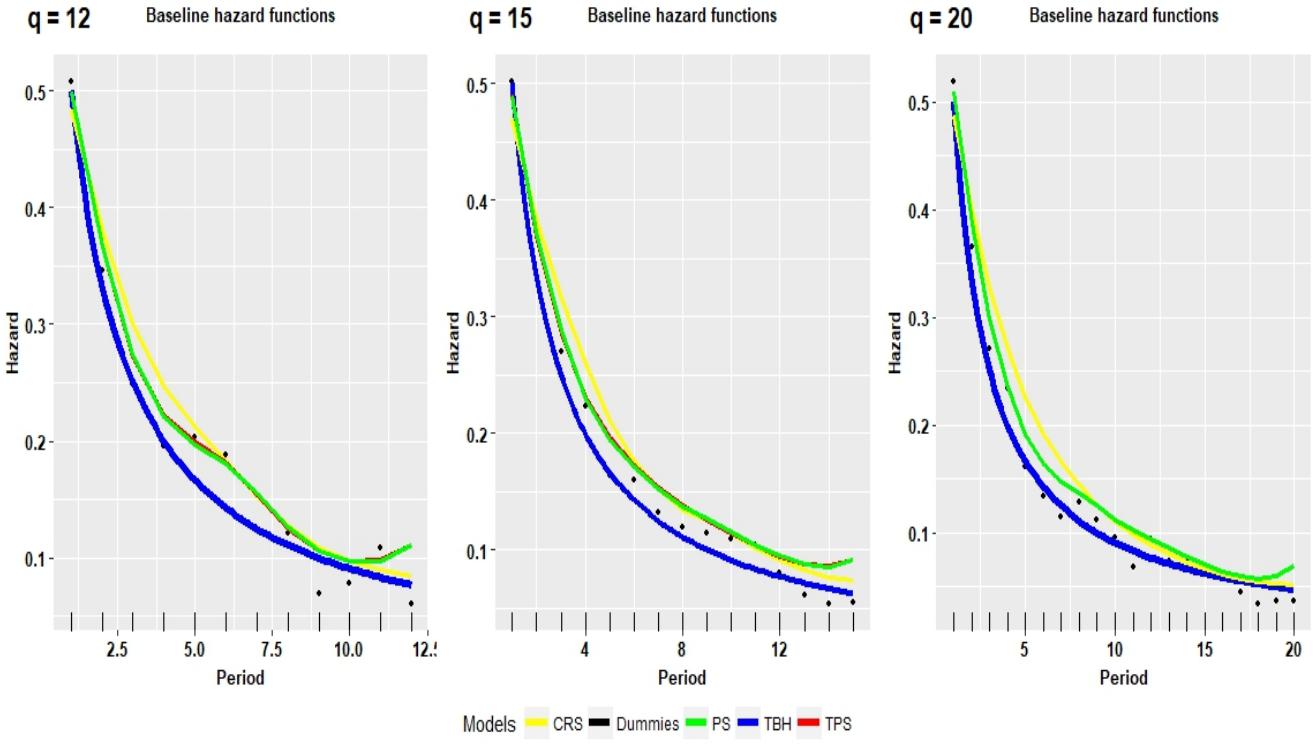


Table 4.2: The RMSE values for dummies, penalised cubic regression splines (CRS), P-splines (PS) and thin plate regression spline (TPS) for $n = 100$ $q \in \{12, 15, 20\}$.

		Estimated models			
n	q	Dummies	CRS	TPS	PS
100	12	0.1080	0.0665	**0.0419**	0.0714
	15	0.0908	**0.0505**	0.0510	0.0509
	20	0.1020	0.0666	**0.0539**	**0.0539**

From Table 4.2 above, the splines were observed to show a better performance than the model with the dummies. In this setting the thin plate regression spline showed a better fit than its smoothing counterparts, but as the ratio of the sample size to the number of dummies reduces, the results for the penalised spline become almost similar to that of the TPS. Figure 4.3 gives a visual illustration of these results.

Figure 4.3: Baseline hazards $n = 100$ and $q \in \{12, 15, 20\}$

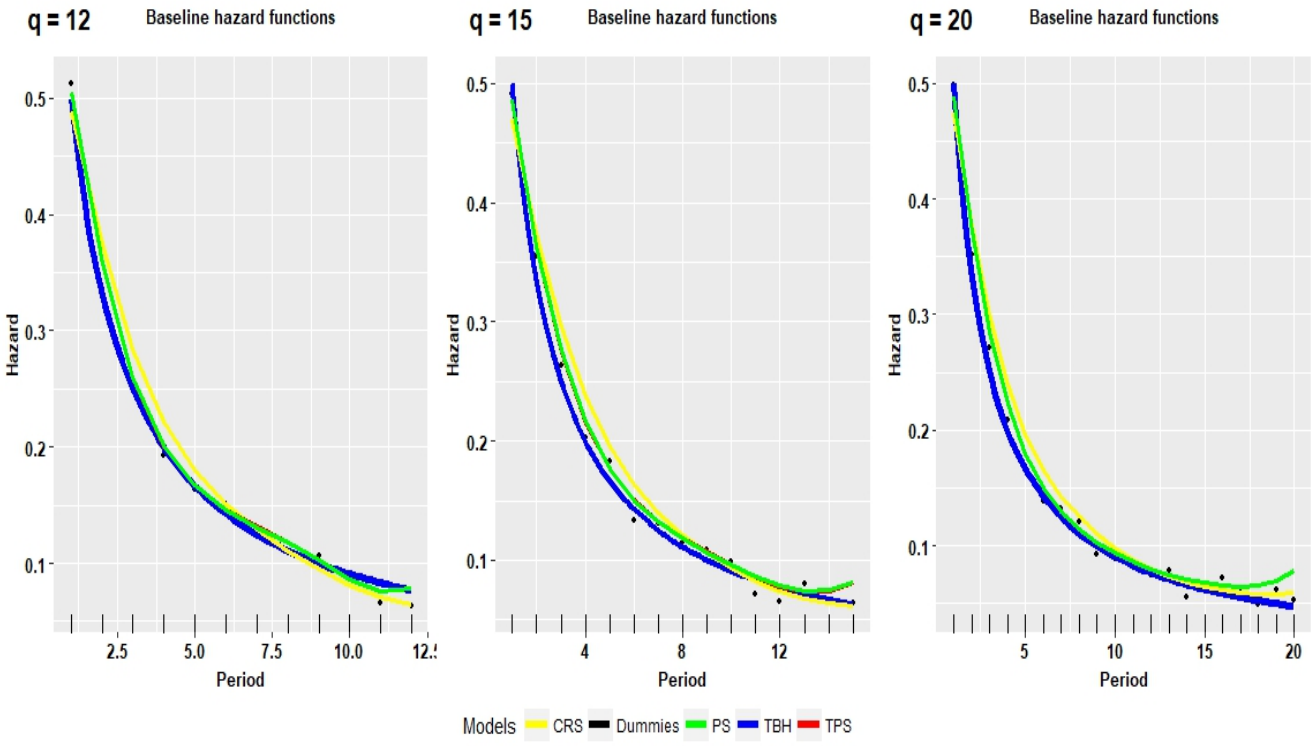


Figure 4.3 reveals that when the ratio of the sample size to the dummies gets smaller, the penalised regression smoothers seem to fit better than the dummies.

Table 4.1 presents the mean rMSE, which was used as performance criteria for the selection of the optimal model. According to the table 4.1, generally all the smoothing methods performed better than the dummies because in all the different settings, the rMSE values for the smoothing methods were smaller than the rMSE values for the model with the dummies. In all the settings, the cubic regression spline showed to have performed fairly better than the other smoothing models.

Table 4.3: The RMSE values dummies, penalised cubic regression splines (CRS), P-splines (PS) and thin plate regression spline (TPS) for $n = 500$ and $q \in \{12, 15, 20\}$.

		Estimated models			
n	q	Dummies	CRS	TPS	PS
500	12	0.0534	0.0263	**0.0236**	**0.0236**
	15	0.0347	**0.0165**	0.0171	0.0171
	20	0.0355	0.0258	**0.0244**	**0.0244**

The mean RMSE values from Table 4.3 suggest that the dummies seem to be performing poorly compared to the smoothing methods. This model recorded highest values of RMSE in this simulation setting when compared to the models with a smooth baseline hazard.

Figure 4.4: Baseline hazards $n = 500$ and $q \in \{12, 15, 20\}$.

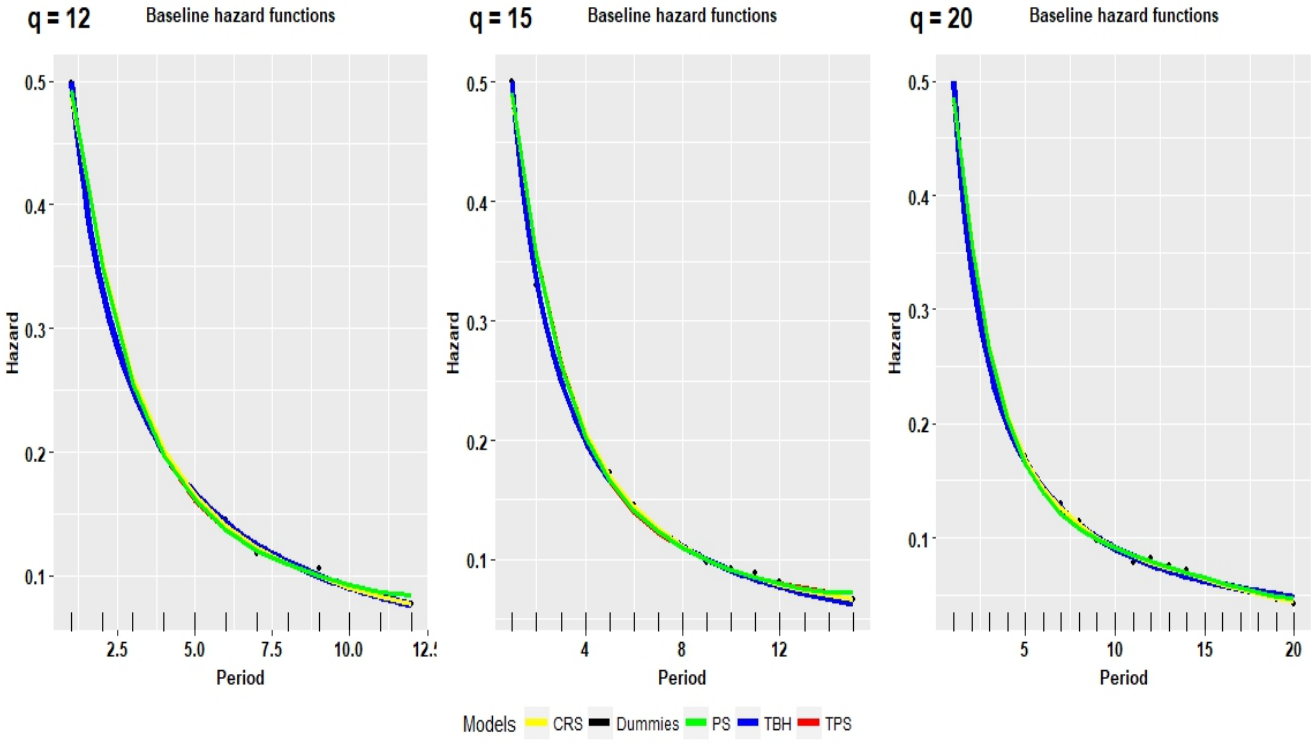


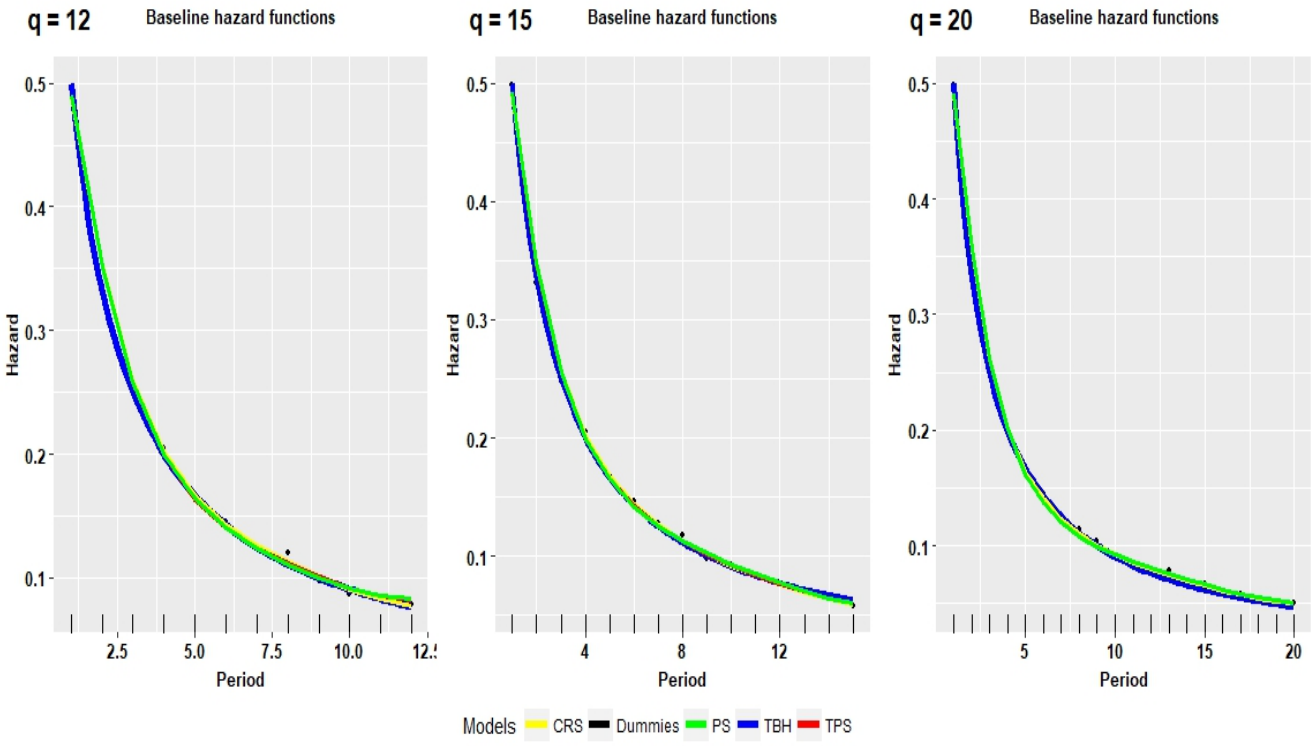
Figure 4.4 confirms the findings in Table 4.3. From this figure it is observed that as the ratio of the sample size to the number of periods decreases, the smoothing methods appeared to be performing fairly better than the model where duration was modeled using a dummy variable. However, from this plot no single smoothing method appeared to be performing better than the other smoothing methods.

Table 4.4: The RMSE values for dummies, penalised cubic regression splines (CRS), P-splines (PS) and thin plate regression spline (TPS) for $n = 1000$ and $q \in \{12, 15, 20\}$.

		Estimated models			
n	q	Dummies	CRS	TPS	PS
1000	12	0.0267	0.0224	0.0224	**0.0209**
	15	0.0149	0.0109	**0.0100**	0.0101
	20	0.0309	**0.0154**	0.0172	0.0172

Table 4.4 summarises the RMSE results obtained when the sample size was set to 1000. The average deviations of the smoothing methods from the true baseline hazard seemed to be smaller than the average deviations of the model with the dummies because the smoothing methods recorded smaller RMSE values in all the settings. This implies that the smoothing methods perform better even though no smoothing method outperformed the other.

Figure 4.5: Baseline hazards for $n = 1000$ and $q \in \{12, 15, 20\}$.



Results from figure 4.5 seem to be in agreement with the results tabled in Table 4.4. This figure shows that the model with the dummies does not perform very well in comparison with the smoothing methods. The fit of the grouped time model gets poorer when the ratio of the time periods to the sample size increases.

Table 4.5: The RMSE values for dummies, penalised cubic regression splines (CRS), P-splines (PS) and thin plate regression spline (TPS) for $n = 3000$ and $q \in \{12, 15, 20\}$.

		Estimated models			
n	q	Dummies	CRS	TPS	PS
3000	12	0.0143	**0.0110**	0.0123	0.0123
	15	0.0155	**0.0121**	0.0124	0.0125
	20	0.0160	**0.0112**	0.0114	0.0115

Table 4.5 presents the results obtained from RMSE analysis where the sample size was set to 3000. The models with the smooth baseline hazard function performed better than the model with a discrete baseline hazard. A graphical summary of these results is given in figure 4.6.

Figure 4.6: Baseline hazards for $n = 3000$ and $q \in \{12, 15, 20\}$.

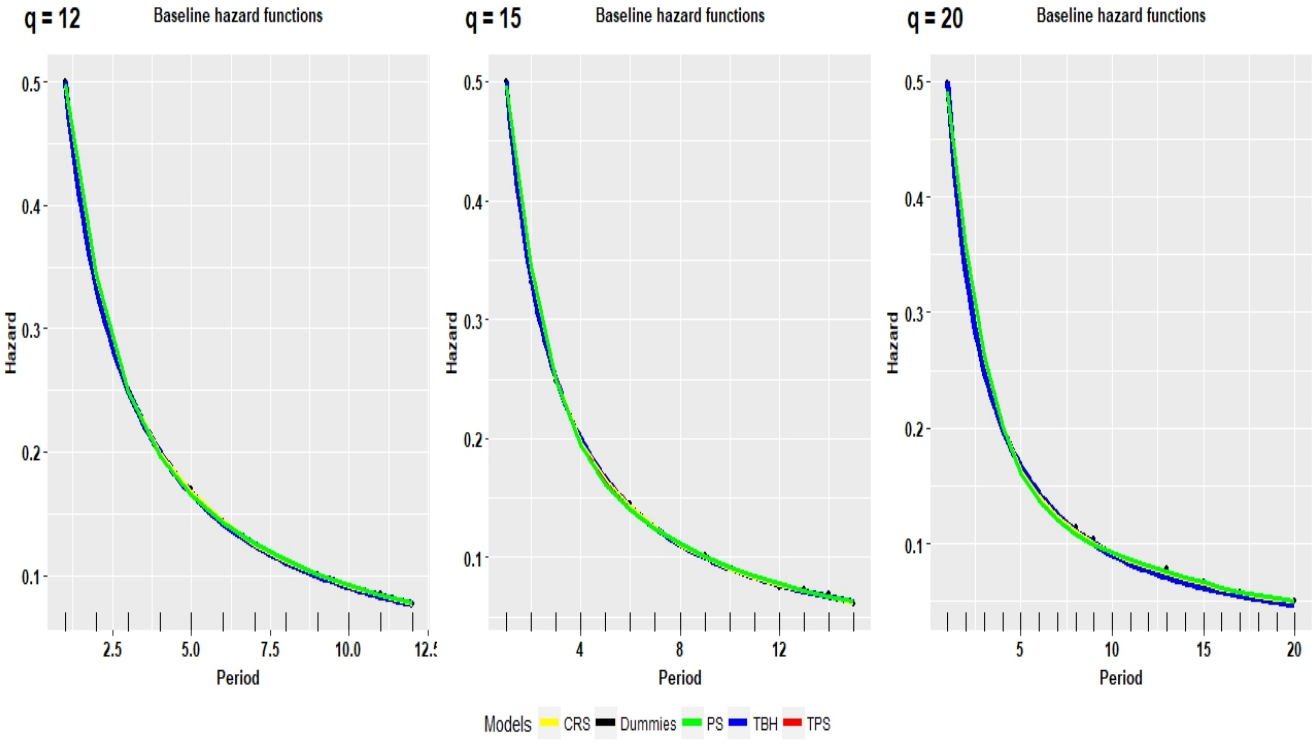
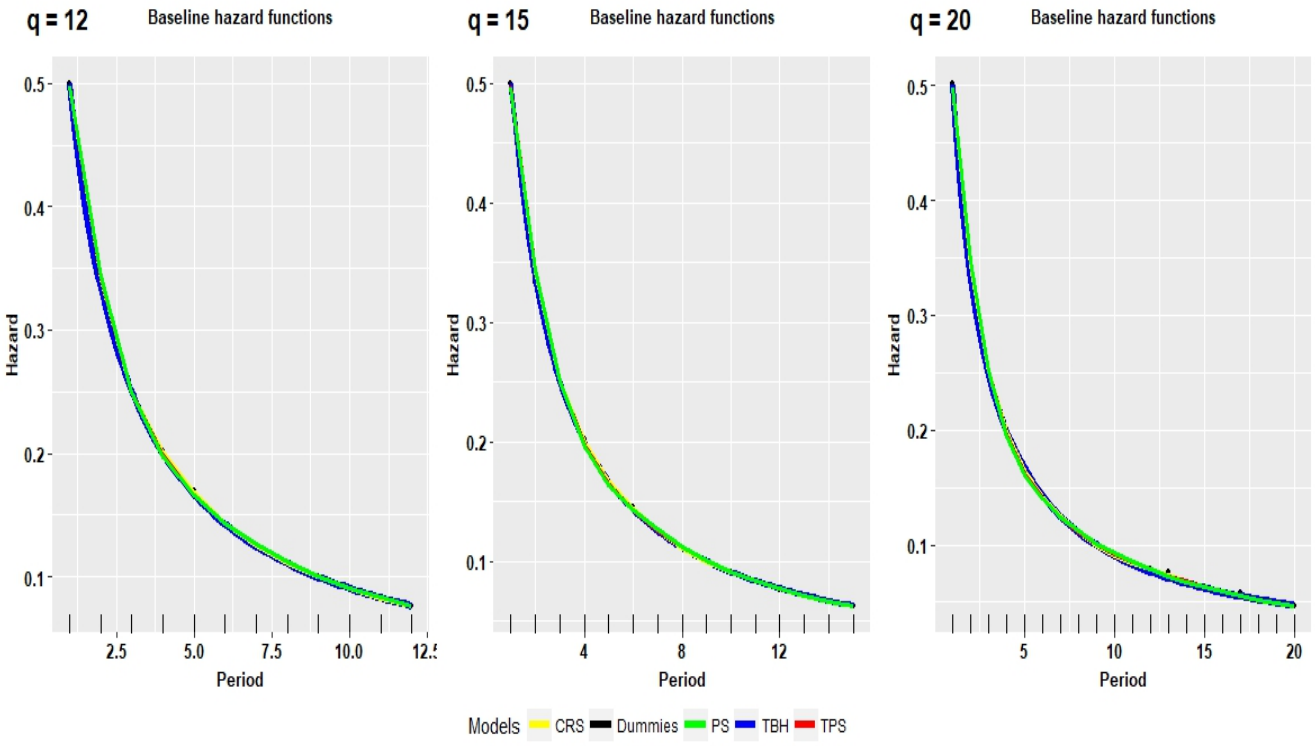


Table 4.6: The RMSE values for dummies, penalised cubic regression splines (CRS), P-splines (PS) and thin plate regression spline (TPS) for $n = 4500$ nd $q \in \{12, 15, 20\}$.

		Estimated models			
n	q	Dummies	CRS	TPS	PS
4500	12	0.1118	**0.0106**	**0.0106**	**0.0106**
	15	0.0106	**0.0058**	0.0061	0.0063
	20	0.1112	**0.1107**	**0.1107**	**01107**

The RMSE analysis presented in Table 4.6 above suggests the average distances from the estimated baseline hazard were smaller for smoothing methods than the model with the discrete baseline hazard, thus indicating that the smoothing methods work better.

Figure 4.7: Baseline hazards for $n = 4500$ and $q \in \{12, 15, 20\}$.



Observation of figure 4.7 is in concordance with the RMSE results in Table 4.6.The estimated smooth baseline hazards performed better than the dummies in this setting.

4.3 Results for the gamma baseline hazard

Table 4.7: The RMSE values for dummies, penalised cubic regression splines (CRS), P-splines (PS) and thin plate regression spline (TPS) for $n = 50$ and $q \in \{12, 15, 20\}$.

		Estimated models			
n	q	Dummies	CRS	TPS	PS
50	12	0.0789	**0.0325**	0.0545	0.05411
	15	0.0659	**0.0415**	0.0468	0.0468
	20	0.0878	**0.0194**	0.0208	0.0208

Table 4.7 is a summary of the average RMSE values obtained from fitting the estimated baseline hazard functions for when the sample size was set to 50. From these results we notice that the smoothing methods seem to be performing relatively better than the model with a discrete baseline hazard. The penalised cubic regression reported the lowest values of RMSE in all the settings, thus indicating to be performing better than the other smoothing methods.

Figure 4.8: Baseline hazards $n = 50$ and $q \in \{12, 15, 20\}$.

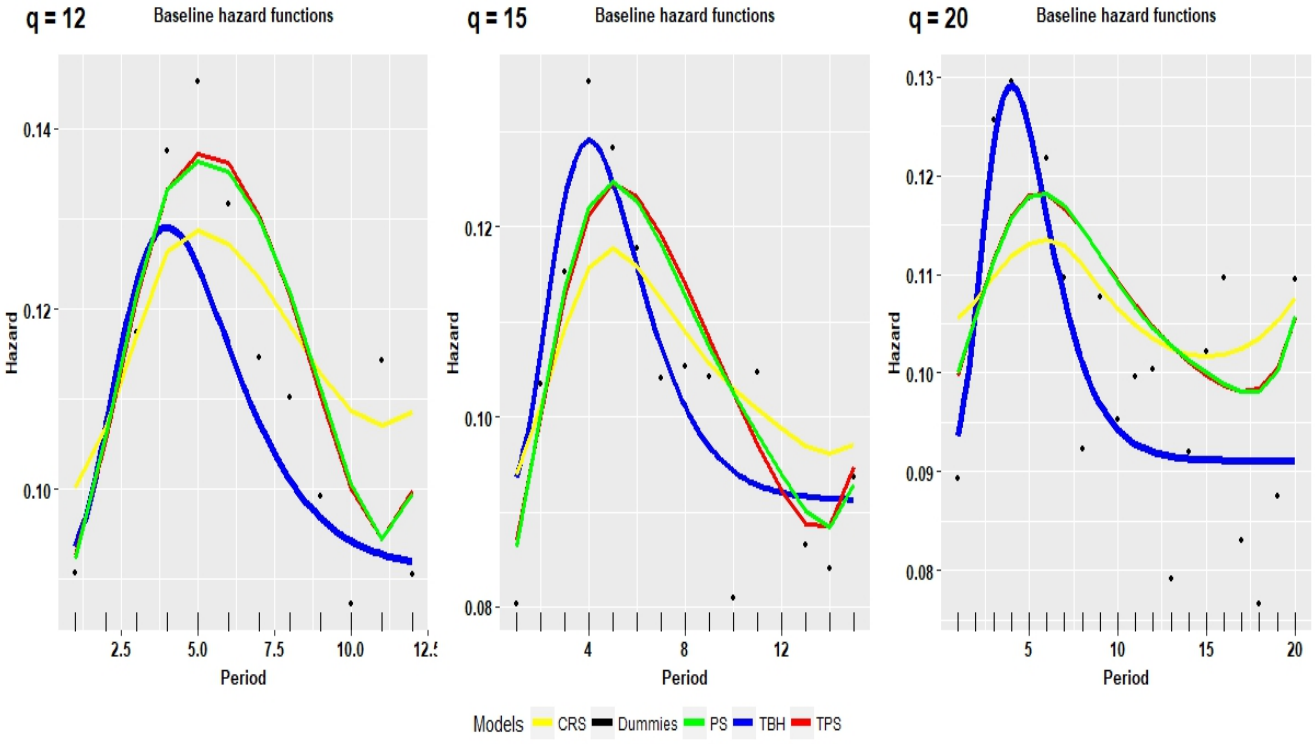


Figure 4.8 all the estimated smooth baseline hazards are appear to be performing better than the model with a discrete baseline function.

Table 4.8: The RMSE values for dummies, penalised cubic regression splines (CRS), P-splines (PS) and thin plate regression spline (TPS) for $n = 100$ and $q \in \{12, 15, 20\}$

		Estimated models			
n	q	Dummies	CRS	TPS	PS
100	12	0.0551	0.0322	0.0322	**0.0290**
	15	0.0607	**0.0297**	0.0383	0.0383
	20	0.1090	**0.0581**	0.0643	0.0643

Table 4.8 indicate that all the smoothing methods appear have a better than the model with the dummies because these methods recorded smaller average departures from the true baseline hazard. Figure 4.9 gives a visual representation of the above results.

Figure 4.9: Baseline hazards $n = 100$ and $q \in \{12, 15, 20\}$

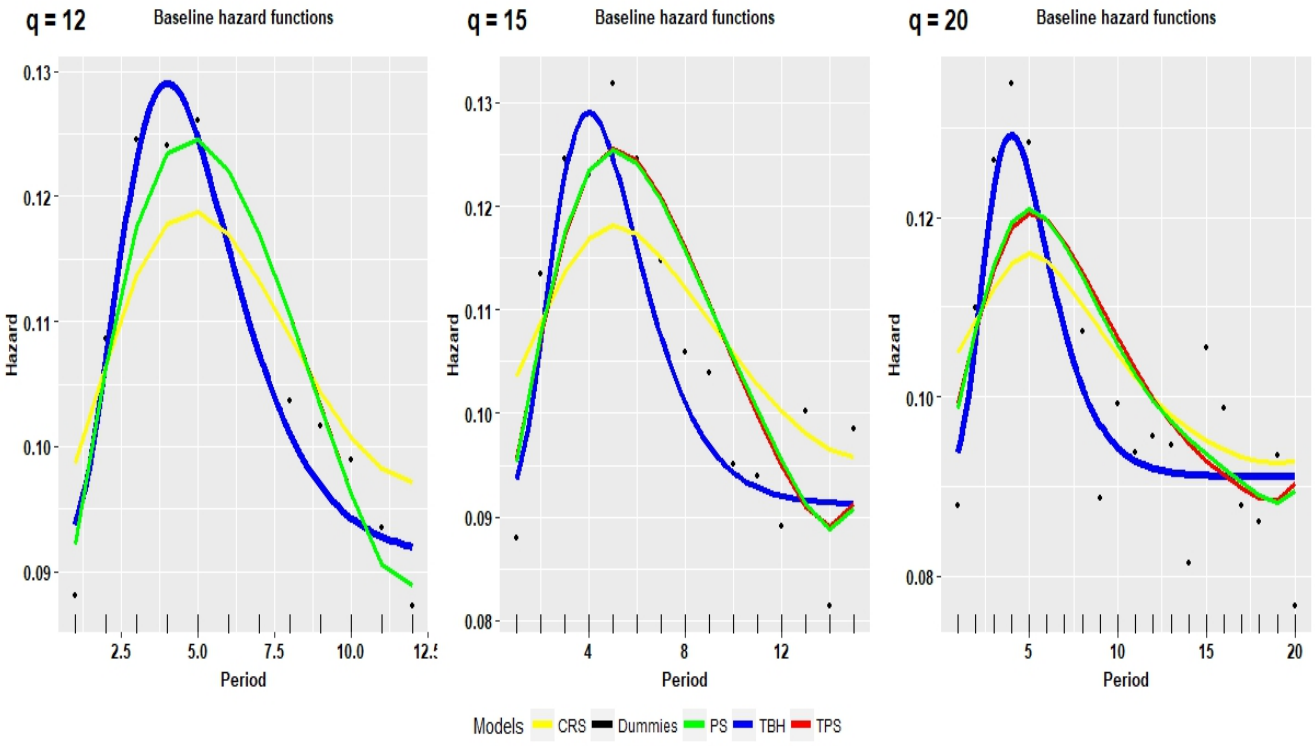
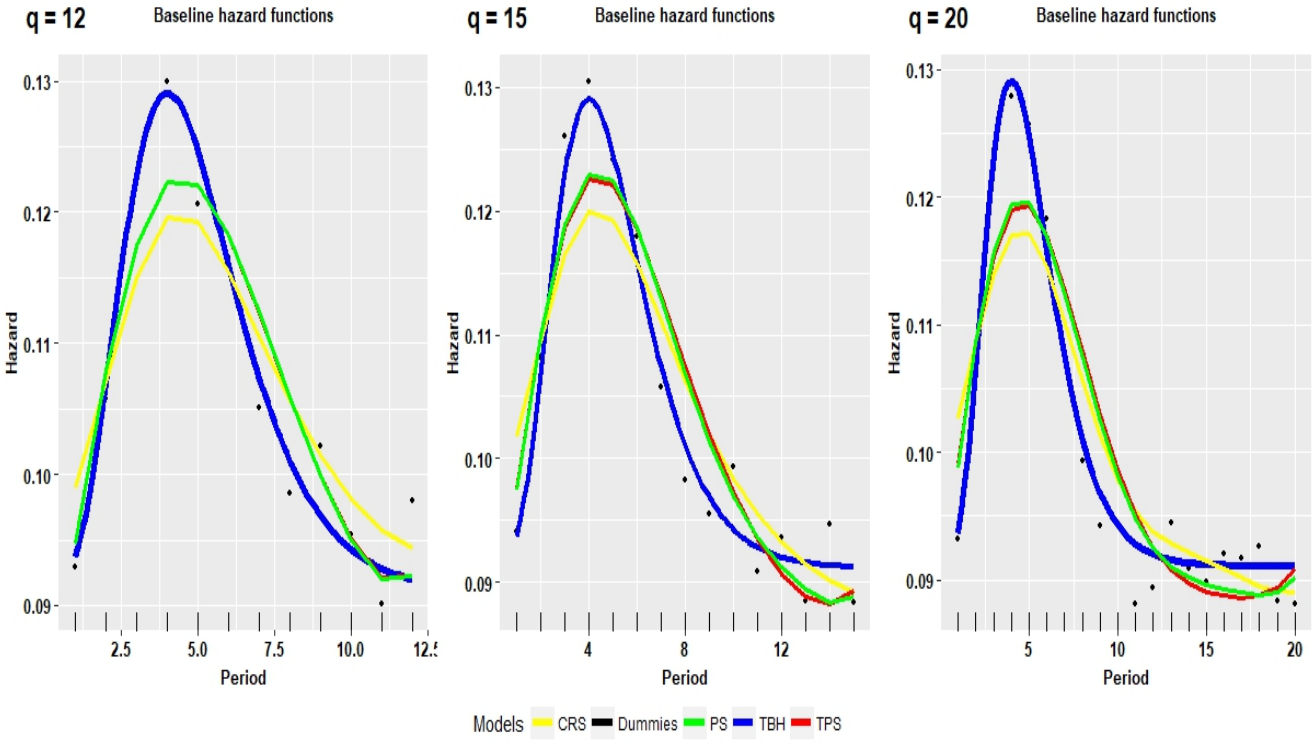


Table 4.9: The RMSE values for dummies, penalised cubic regression splines (CRS), P-splines (PS) and thin plate regression spline (TPS) for $n = 500$ and nd $q \in \{12, 15, 20\}$.

		Estimated models			
n	q	Dummies	CRS	TPS	PS
500	12	0.0263	**0.0137**	0.0156	0.0156
	15	0.0140	0.0072	**0.0060**	0.0061
	20	0.0258	**0.0091**	**0.0091**	**0.0091**

According to the RMSE analysis of the results obtained from this setting summarised in Table 4.9, the smoothing methods seem to be performing relatively better than the model with a discrete baseline hazard function.

Figure 4.10: Baseline hazards $n = 500$ and $q \in \{12, 15, 20\}$.



Generally, figure 4.10 illustrates an improved fit for all the models. However, it is observed that when the ratio of sample size to the dummies reduces, the smoothing methods appear to be closer to the true curve (TBH) than the dummies.

Table 4.10: The RMSE values for dummies, penalised cubic regression splines (CRS), P-splines (PS) and thin plate regression spline (TPS) for $n = 1000$ and $q \in \{12, 15, 20\}$.

		Estimated models			
n	q	Dummies	CRS	TPS	PS
1000	12	0.0188	**0.0078**	0.0080	0.0080
	15	0.0143	**0.0085**	0.0109	0.0109
	20	0.0216	0.0094	**0.0090**	**0.0090**

The RMSE results obtained from the simulated setting in Table 4.10 suggest that the model with the dummies performed poorer than the smoothing methods. Figure 4.11 provides a visual demonstration of these results

Figure 4.11: Baseline hazards $n = 1000$ and $q \in \{12, 15, 20\}$

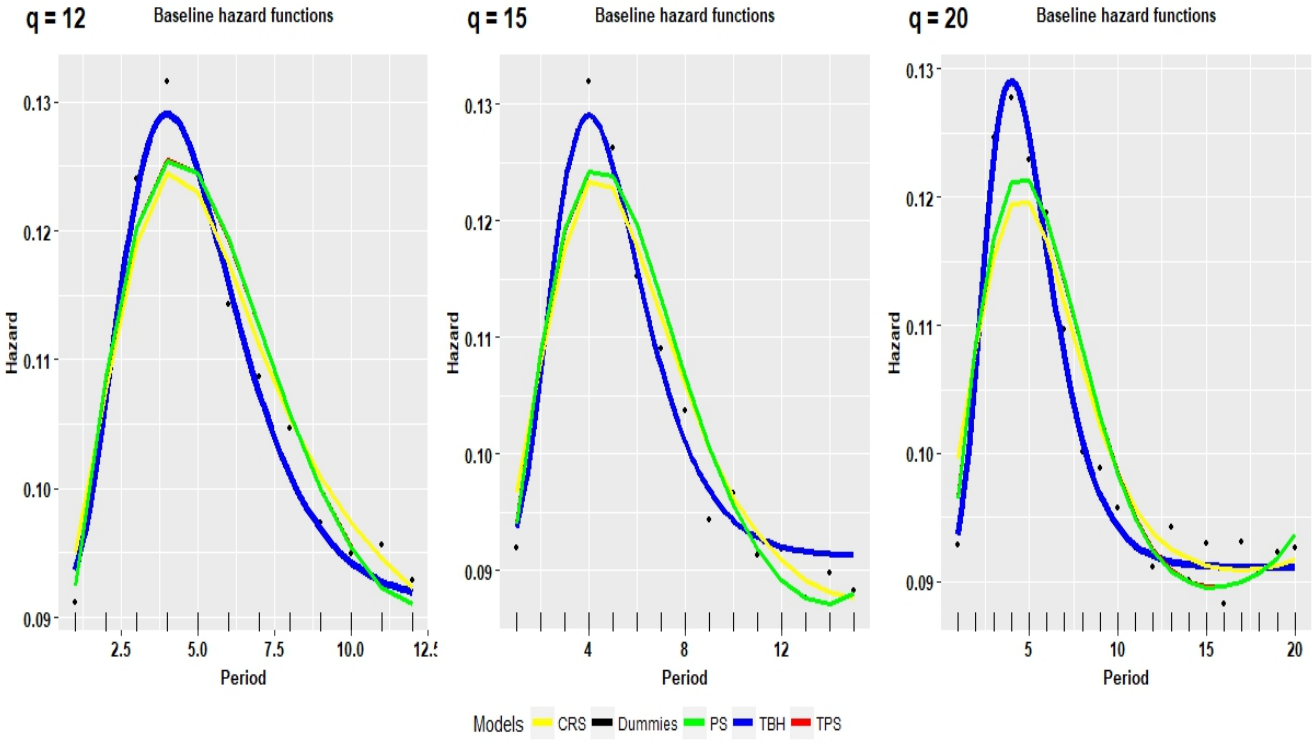


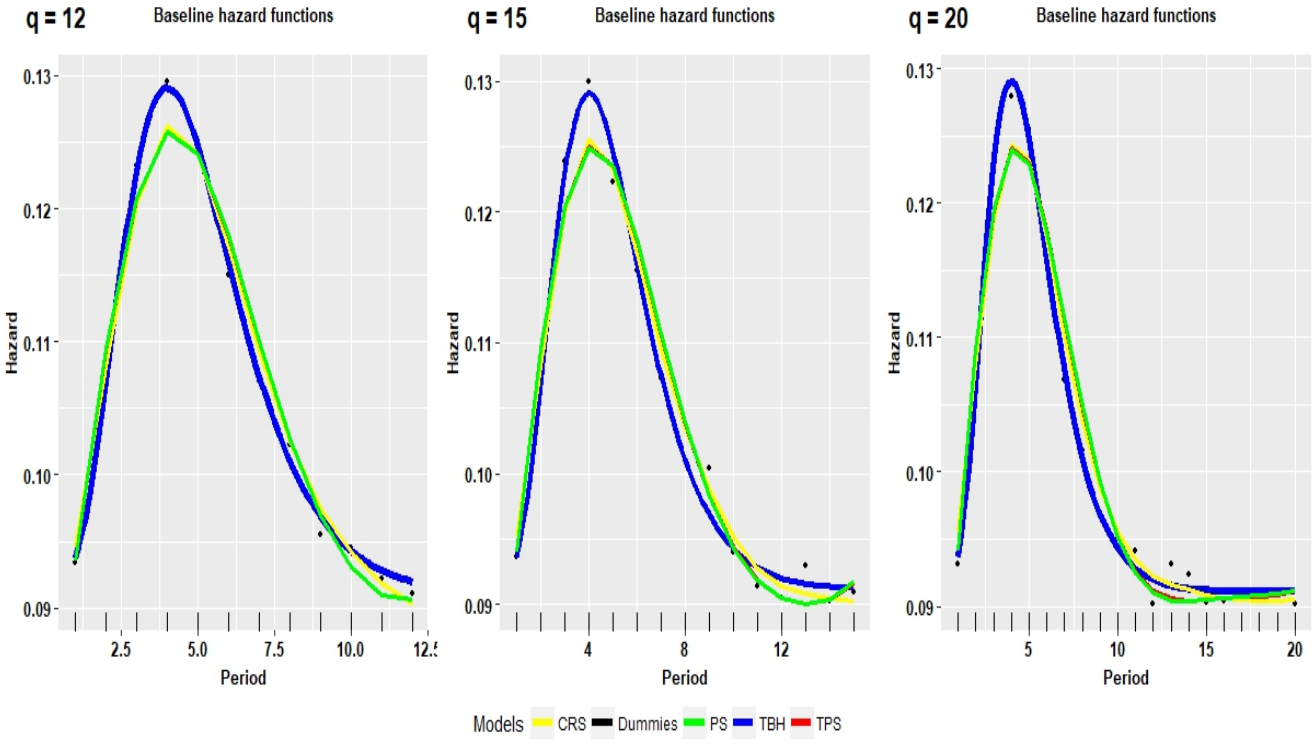
Figure 4.11 above supports the results summarised in Table 4.11. This figure depicts a seemingly better fit for the smoothing methods than the dummies.

Table 4.11: The RMSE values for dummies, penalised cubic regression splines (CRS), P-splines (PS) and thin plate regression spline (TPS) for $n = 3000$ and $q \in \{12, 15, 20\}$.

		Estimated models			
n	q	Dummies	CRS	TPS	PS
3000	12	0.0096	**0.0054**	**0.0054**	**0.0054**
	15	0.0163	**0.0091**	0.0092	0.0092
	20	0.0122	**0.0058**	0.0061	0.0061

Here, in Table 4.11, we observe that the dummies still show a poorer fit in comparison with penalised regression smoothers.

Figure 4.12: Baseline hazards for $n = 3000$ and $q \in \{12, 15, 20\}$.



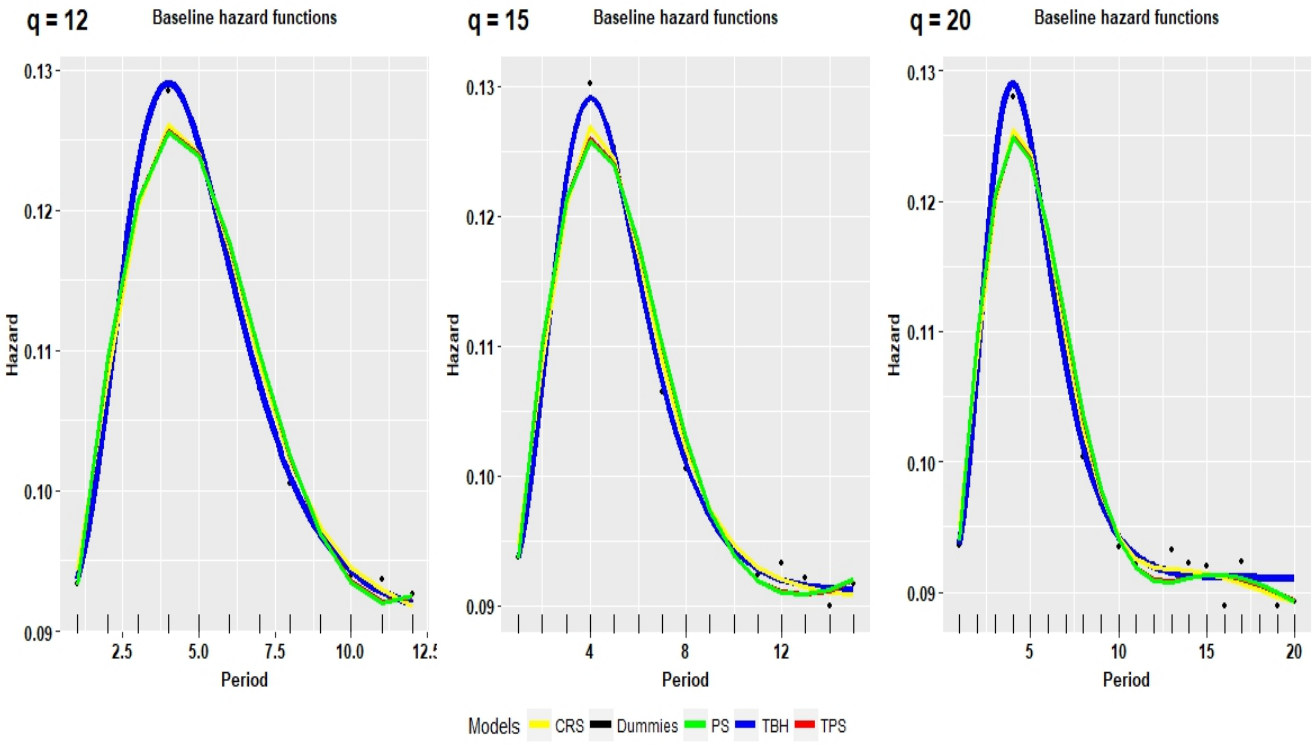
Evidence from figure 4.12 suggests that as the ratio of the sample size to the time periods decreases the smoothing methods appear to be performing better than the dummies.

Table 4.12: The RMSE values for dummies, penalised cubic regression splines (CRS), P-splines (PS) and thin plate regression spline (TPS) for $n = 4500$ and $q \in \{12, 15, 20\}$.

		Estimated models			
n	q	Dummies	CRS	TPS	PS
4500	12	0.0062	**0.0031**	**0.0031**	**0.0031**
	15	0.0078	**0.0048**	**0.0048**	**0.0048**
	20	0.1119	**0.1113**	**0.1113**	**0.1113**

For the RMSE analysis summarised in Table 4.12 reveals that the smoothing methods performed better than the model with the dummies as the ratio of the sample size to the number of time points reduces. Figure 4.13 supports these findings.

Figure 4.13: Baseline hazards for $n = 4500$ and $q \in \{12, 15, 20\}$.



From the figure 4.13 we observe that the smoothing methods seem to be performing better than the dummies as the ratio of the sample size to the dummies reduces.

4.4 Concluding remarks

Evidence from the RMSE analyses and the corresponding graphical analyses obtained from the simulation experiments suggest that the model with the discrete baseline hazard function performed relatively poor in comparison to the models with the smooth baseline hazard function when the ratio of the sample size to the number of time points becomes smaller.

The penalised cubic regression spline recorded the smallest RMSE scores in most of simulation settings indicating that it performed better than its smoothing counter- parts. There were very small differences between the thin plate regression spline and the p-spline.

Finally, there appeared to be an improvement in the fit of all the estimated models as the sample size increases.

Chapter 5

Empirical data analysis results

Introduction

This chapter presents a detailed presentation of the results obtained from the analysis of time to first alcohol intake. The first section presents the summary statistics of all explanatory variables considered in the study. Section 2 presents the results that were obtained from LASSO algorithm implemented using the R-package GLMMLasso. The subsequent section covers a comparison of the estimates that resulted from fitting a continuation proportional model with a discrete baseline hazard specification and where the baseline was also modeled flexibly by penalised regression splines. Age at first alcohol intake ranges from 9 years to 32 years.

5.1 Exploratory data analysis

5.1.1 Univariate analysis

Table 5.1: Five number summary for student's current age

Variable	Minimum	1st quartile	Median	3rd quartile	Maximum
Age (in years)	18	21	23	24	34

Table 5.1 shows the five number summary for the student's current age. From this table, it is observed that the age of the students range from 18 years and that at least fifty percent of the students were 23 years of age.

Table 5.2: Independent variables considered for the analysis of time to alcohol intake.

Variable (abbreviation)	Category	Frequency (sample proportion)
Gender	Male	396 (53.2%)
	Female	349 (46.8%)
Other drugs	Yes	108 (14.5%)
	No	637 (85.5%)
Drinking peers	Yes	318 (42.7%)
	No	427 (57.3%)
Physical abuse	Yes	60 (8.1%)
	No	685 (91.9%)
Negative life events (Neg. life ev.)	Yes	217 (29.1%)
	No	528 (70.9%)
Parents abusing drug/alcohol	Yes	344 (46.2%)
	No	401 (53.8%)
Parental discipline	Yes	441 (59.2%)
	No	304(40.8 %)
Welfare dependant	Yes	480 (64.4 %)
	No	265 (35.6 %)
Family dysfunctional	Yes	254 (34.1%)
	No	491 (65.9%)
Family income (in Rands)	<1000	134 (18.0%)
	1000-4000	405 (54.4%)
	>4000	206 (27.8%)
Faculties	Business studies	25 (3.4%)
	Engineering studies	10 (1.3 %)
	Agricultural science	22 (3.0%)
	Education	125 (16.8%)
	Environmental science	60 (8.1 %)
	Health science	48 (6.4 %)
	Law	157 (21.1%)
	Management science	63 (8.5%)
	Mathematical science	97 (13.0%)
	Human social science	138 (18.5 %)
Marital status	ever married	50 (6.7%)
	single	695 (93.3%)
Academic level	1^{st} year	127 (17.1%)
	2^{nd} year	267 (35.9%)
	3^{rd} year	219 (29.4%)
	4^{th} year	132 (17.6%)
Institutions	Univen	709 (95.3%)
	Techniven	35 (4.7%)
Relationship with adult	Yes	423 (56.8%)
	No	322 (43.2%)
Siblings	0-3	347 (46.7%)
	4-6	354 (47.5%)
	≥ 7	43 (5.8%)
Parent's qualification	none	314 (42.1%)
	matric	154 (20.7%)
	diploma	96 (12.9%)
	degree	66 (8.9%)
	honours	115 (15.4 %)
Race	African	729 (97.9%)
	other	16 (2.1 %)
Childhood residence	rural	392 (52.6 %)
	urban	341 (45.8 %)
	farm land	12 (1.6 %)

Table 5.3: Independent variables considered for the analysis of time to alcohol intake.

Variable (abbreviation)	Category	Frequency (sample proportion)
Sexual abuse	Yes	12 (1.64%)
	No	733 (98.4 %)
Stress	Yes	663 (89.0 %)
	No	82 (11.0%)

Table 5.2 presents the summary statistics of the explanatory variables considered in this study. It is seen from Table 5.2 above that males comprised more than half of the respondents in the study. A great number (91.94%) of the students reported to have never experienced any physical abuse and almost 30% confessed to have ever experienced a negative life event in their lives. Almost half of the students recorded that their parents either abuse alcohol or drugs. Approximately 34% of the respondents are from a dysfunctional home and a little over 50% of the students reported to having a family income of between R1000 and R4000. More than half of the students have a relationship with an adult.

According to Table 5.3 47% have between 0 and 3 siblings and 59.3% recorded that their parents discipline them. A little over half (52.69%) grew up in a rural place and almost 2% grew up on farm.

Almost all (97.85%) in the study are African and a little less than 10% of them reported to have ever been married. Most students were in their second year and 95.30% of them were registered at the university of Venda. More than half of the students depend on a social welfare and 42% had parents who did not have any formal education. Almost 90% have a history of stress.

5.1.2 Results obtained from LASSO variable selection algorithm

The discrete survival models have a special feature that they have a baseline hazard that is specified by parameters in λ_0. Even though a moderate number of intervals can be used, the estimation of the baseline might be very unstable. Consequently one has to incorporate an additional penalty in the model objective. The GlmmLasso algorithm uses an L_1 penalty term of the form $\lambda \sum |\beta_i|$ to enforce variable selection, where the tuning parameter λ determines the strength of the selection. When $\lambda = 0$ implies that there is no selection while for $\lambda \to \infty$ means all the explanatory variables are not included in the model. The size of λ can be chosen using

information criteria such as AIC or BIC. To determine the optimal tuning parameter we used BIC. The variables with zero coefficients in Table 5.4 are considered not to have any effect on the outcome of interest and were thus not included in the final estimated model.

Table 5.4: Estimated covariate effects obtained from the GLMMlasso variable selection algorithm.

Variable	Estimate	Standard error	P - value
Gender	-3.8107×10^{-1}	1.0911×10^{-1}	0.0005
Stress	0.0000	0.0000	0.0000
Family income	0.000	0.000	0.000
Siblings	0.0000	0.0000	0.0000
Parent's qualification	0.000	0.0000	0.0000
Taking other drugs	-5.9221×10^{-1}	1.4807×10^{-1}	<0.0001
Parent's discipline	0.0000	0.0000	0.0000
Negative life event	-2.2370×10^{-1}	1.1874×10^{-1}	0.0060
Dependant on welfare	0.0000	0.0000	0.0000
Physical abuse	-4.7066×10^{-1}	1.7948×10^{-1}	0.0087
Sexual abuse	0.0000	0.0000	0.0000
Relation with adult	0.0000	0.0000	0.0000
Family dysfunction	0.0000	0.0000	0.0000
Parents' abuse of drugs or alcohol	-3.7622×10^{-1}	1.0728×10^{-1}	0.0004
Institution	0.0000	0.0000	0.0000
Faculty	0.0000	0.0000	0.0000
Race	0.0000	0.0000	0.0000
Marital status	0.0000	0.0000	0.0000
Drinking peers	-5.680×10^{-1}	1.4360×10^{-1}	<0.0001

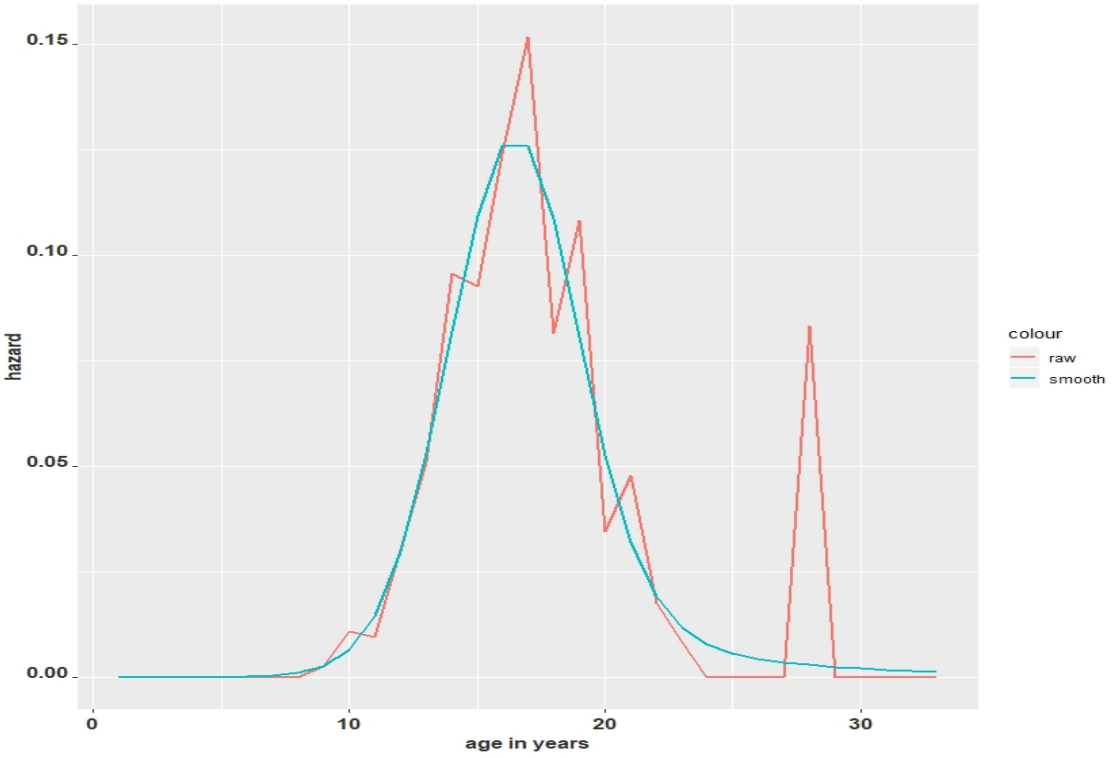


Figure 5.1: Estimated hazard rates obtained from the analysis of time to alcohol intake.

Figure 5.1 depicts the estimated hazard rates of duration to the debut of alcohol intake.

Examination of the raw estimates reveal that the time to alcohol initiation peaks at 17 years and another peak is observed at 27 years. From this plot it is also observed that the raw estimates are highly jagged, possibly as a result of measurement errors that are inherent in survey data. The smoothed version of the hazard rates is clearly seen to alleviate the jaggedness. The smoothed hazard rates were obtained by using a p-spline estimator.

5.1.3 Bivariate analysis

Within group survivor functions, employed here, were obtained using life table estimates and they analyse the compounded effect of covariates on the event occurrence and also allow for a comparison of median lifetime across all predictor levels. In Sithuba (2017), the compounded effect of the covariates on the event occurrence was analysed using Kaplan Meier curves. These type of curves are primarily used to analyse continuous time to event data. One of the drawbacks of the Kaplan-Meier curves is that they assume that individuals are at risk to experience the event outcome on specific times whereas in reality an individual is at risk at any time.

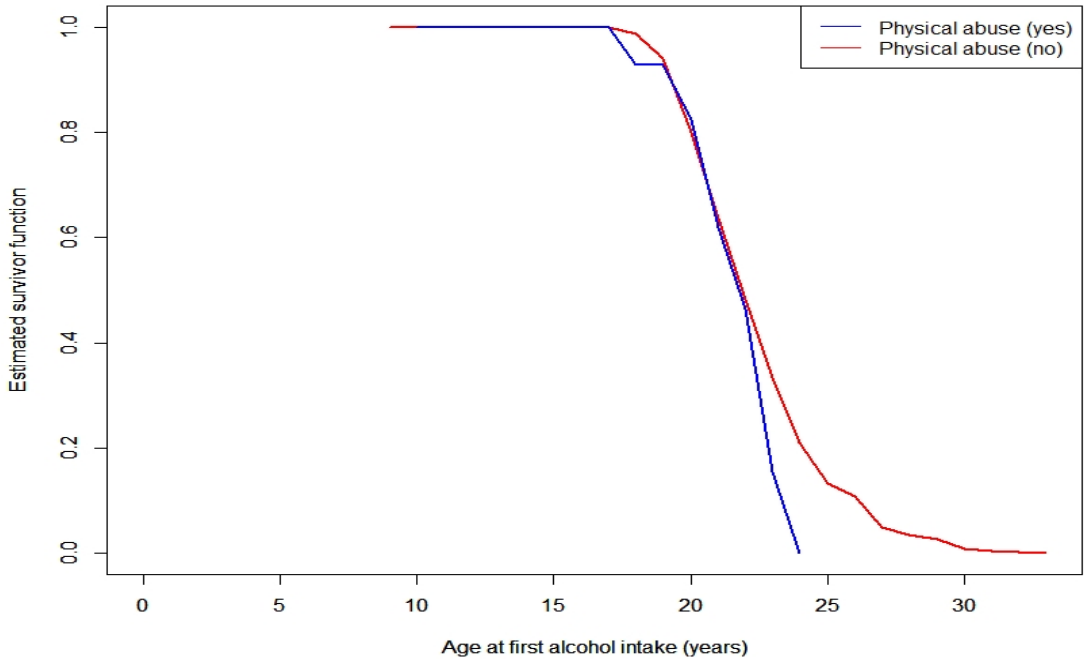


Figure 5.2: Estimated survivor function for ever experiencing physical abuse.

It appears from figure 5.2 that being physically abused increases the risk of early onset of alcohol. At age 24 years, all the respondents that reported to have ever experienced physical abuse had already started drinking.

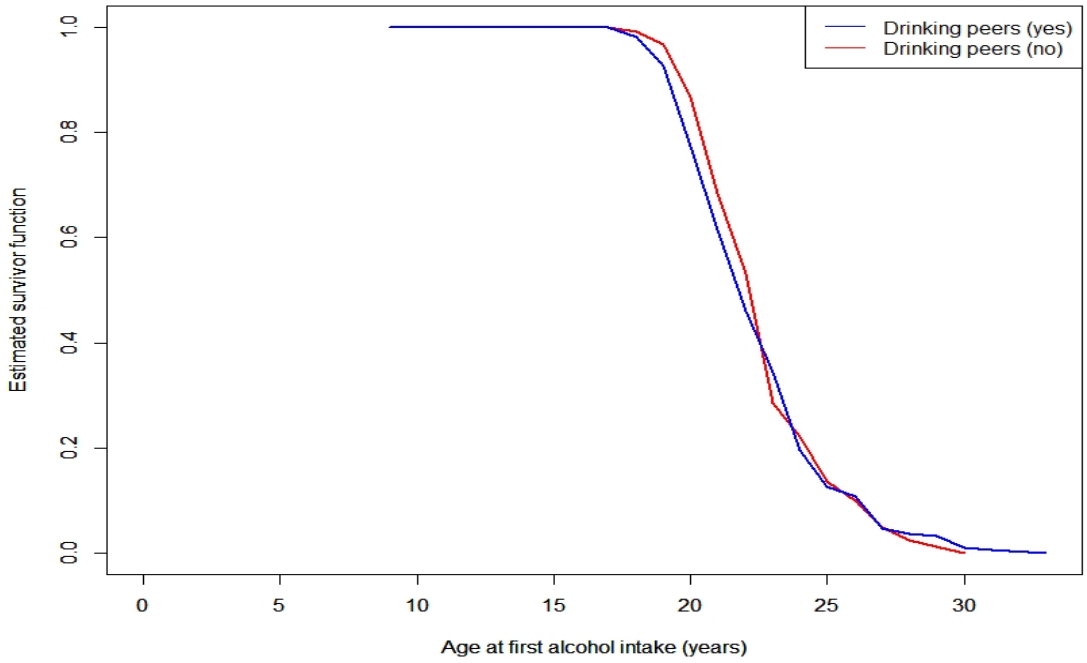


Figure 5.3: Estimated survivor function for drinking peers.

The estimated survival probability of drinking alcohol for the first time is lower for students who had drinking peers as revealed in figure 5.3. At age 21, about 60% of the students that had drinking peers survived early onset of alcohol intake while it was 65% survival for those who did not have any drinking peers.

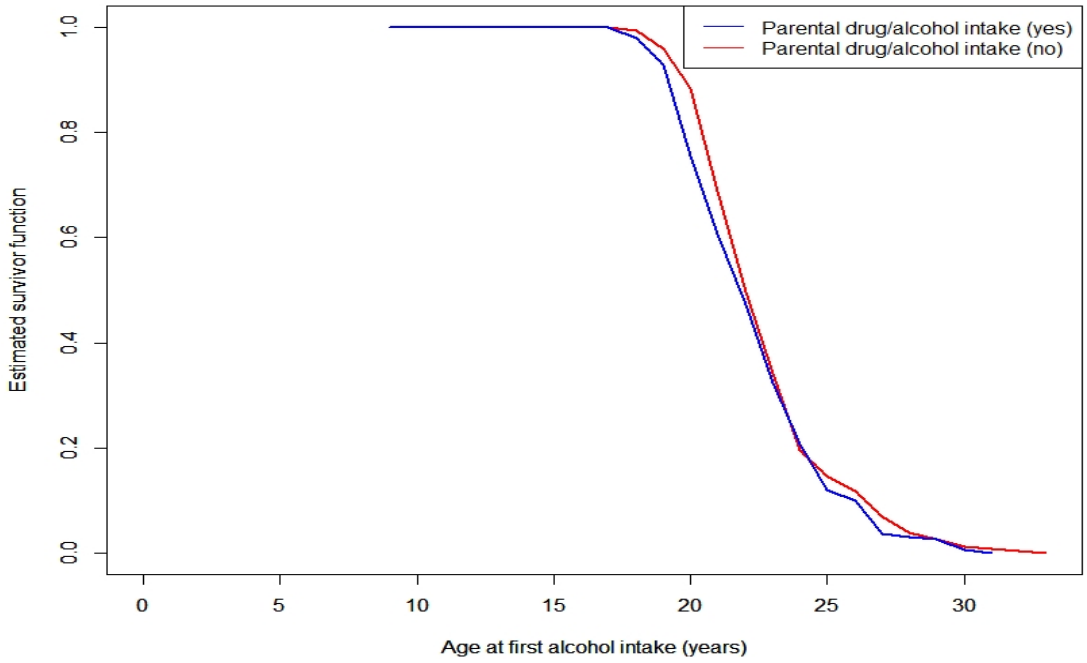


Figure 5.4: Estimated survivor function for parental alcohol intake.

Examination of figure 5.4 reveals that the probability of surviving is lower for students with parents who drink than those students whose parents do not drink. At 22 years, about 60% of students with parents that drink alcohol survived while about 65% of students who did not have parents that drink survived early onset of alcohol intake.

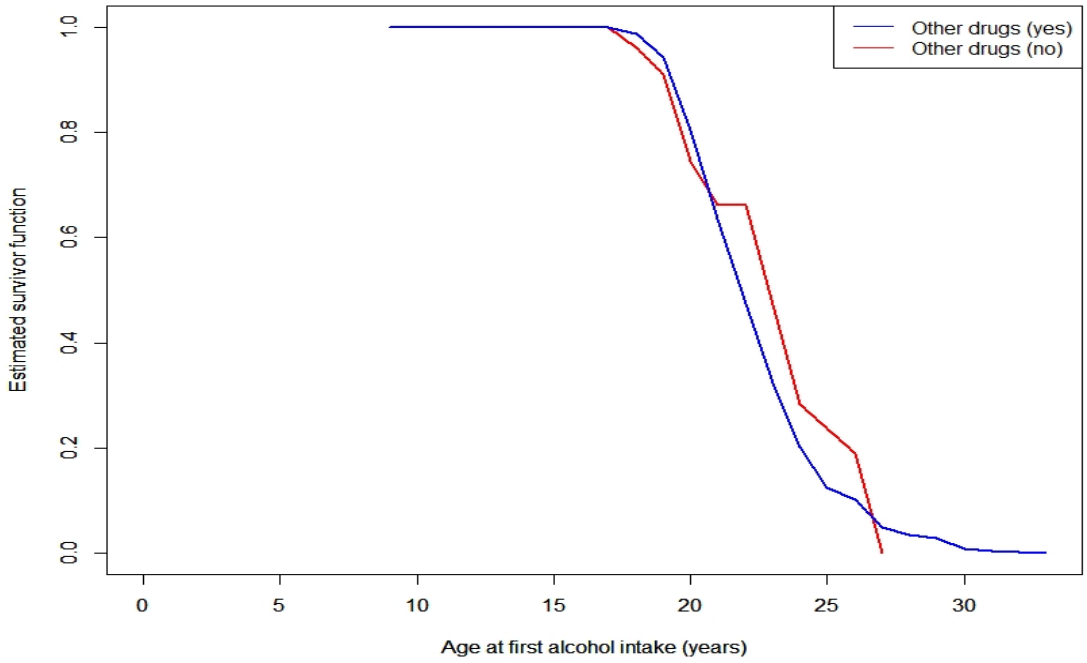


Figure 5.5: Estimated survivor function for other drug usage.

Figure 5.5 depicts the comparison of survival probability for taking other drugs and not taking other drugs. From this diagram it is observed that consumption of other drugs seem to have a catalytic role in early onset of alcohol intake. Fifty percent of students who consumed other drugs started drinking at age 22 years, which is at least a year earlier than students who did not take other drugs.

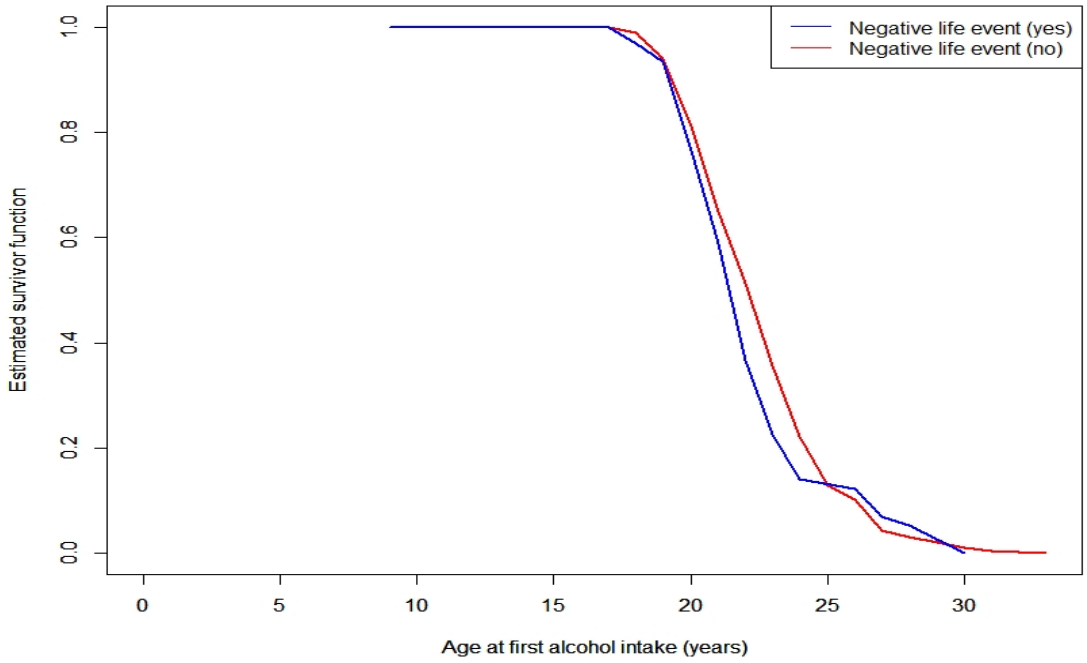


Figure 5.6: Estimated survivor function for ever experiencing a negative life event.

Figure 5.6 shows the survival probability of ever experiencing a negative life event of students. At age 21 years, sixty percent of students who experienced a negative event in their lives survived early onset of alcohol intake whereas for those who did not experience any negative life event seventy percent survived.

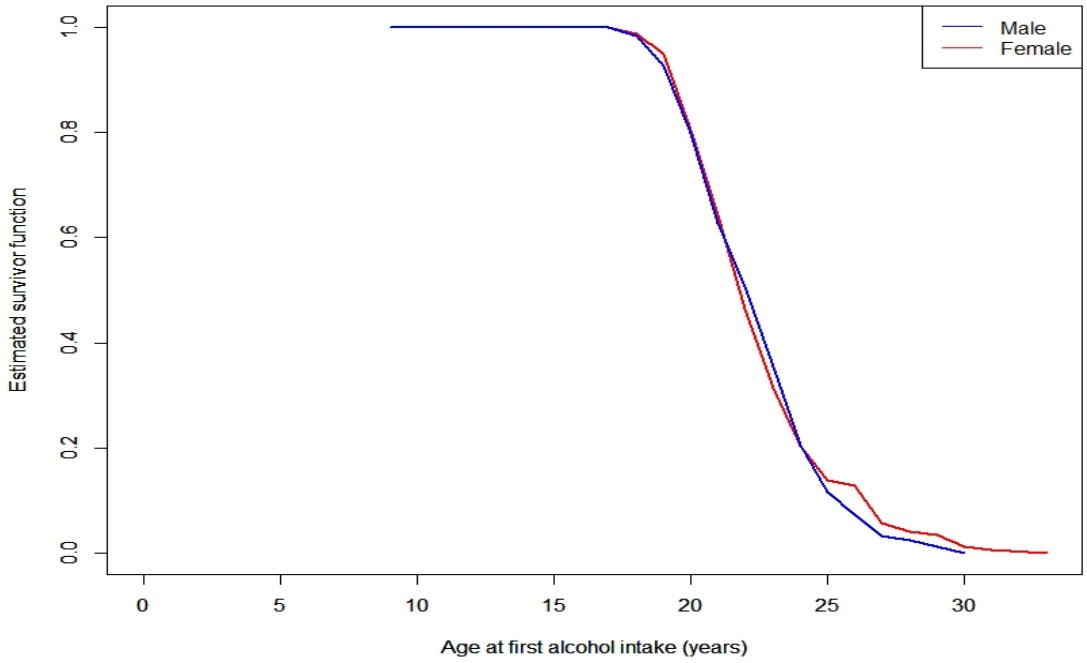


Figure 5.7: Estimated survivor function for gender.

Inspection of figure 5.7 reveals for the first 24 years the probability of surviving early initiation of alcohol intake is the same for both males and females. Beyond 24 years male students seem to have a lower survival rate than their female counterparts.

5.2 Regression modelling

This section discusses the results that were obtained from the parametric model

$$\lambda(t|\mathbf{x}) = h(\lambda_0(t) + \beta_1 gender + \beta_2 \text{negative } life\ event + \beta_3 \text{physically abused} + \beta_4 \text{Other drugs} + \beta_5 \text{drinking peers} + \beta_6 \text{parent alcohol intake}) + \varepsilon_t \tag{5.2.1}$$

where $h(\cdot)$ is the link function chosen as the logistic function and $\lambda_0(t)$ is assumed to be fixed and represents the hazard function with any given set of covariates and ε_t is the error term. When $\lambda_0(t)$ is modeled as a dummy variable equation (5.2.1) above becomes:

$$\lambda(t|\mathbf{x}) = h[(\alpha_1 d_1 + ... + \alpha_{k-1} d_{k-1}) + \beta_1 gender + \beta_2 \text{negative } life\ event + \beta_3 \text{physically abused} + \beta_4 \text{Other drugs} + \beta_5 \text{drinking peers} + \beta_6 \text{parent alcohol intake}] + \varepsilon_t; \quad k = 1, 2, ..., 24 \tag{5.2.2}$$

where

$$d_k = \begin{cases} 1 & \text{if } t = k; \\ otherwise & 0. \end{cases}$$

,

and for the smoothing case equation (5.2.1):

$$\lambda(t|\mathbf{x}) = h\left[\sum_{s=1}^{m} \beta_{0s}\phi_s(t) + \beta_1 gender + \beta_2 \text{negative } life\ event + \beta_3 \text{physically abused} + \beta_4 \text{Other drugs} + \beta_5 \text{drinking peers} + \beta_6 \text{parental alcohol intake}\right], \tag{5.2.3}$$

where $\phi_s(t)$'s are the nonlinear functions of t specified as the penalised cubic regression spline, p-spline and thin plate regression spline.

5.2.1 Baseline life table estimates

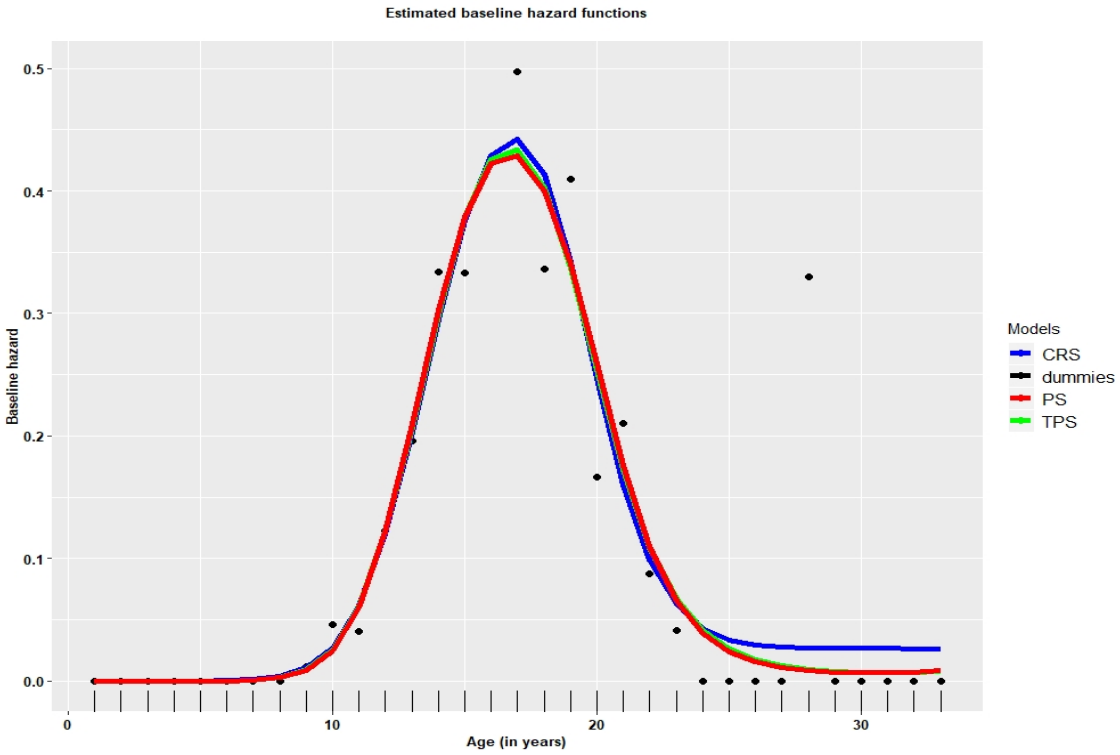


Figure 5.8: This figure shows the estimated baseline hazard functions of the dummies, penalised cubic regression, thin plate regression spline and p-spline.

Figure 5.8 depicts the discrete baseline hazard (dots) and the smoothed (curve) estimates of the baseline $\exp(\lambda_0(t))/(1+\exp(\lambda_0(t)))$ obtained from a logistic discrete hazard model. Inspection of the plot reveals that the conditional probability of alcohol intake increases steadily until the 17 years, decreases rapidly until 25 years and then becomes constant. It is a bit difficult to deduce the exact shape of the baseline hazard from the dummies.

5.2.2 Model comparison and evaluation

Table 5.5: Estimated covariates (est.), standard errors (se) and p-values of the explanatory variables obtained from the model with a discrete baseline hazard (dummies), penalised cubic regression spline (CRS), thin plate regression spline (TPS) and p-splines (PS).

Covariate	Dummies			CRS			TPS			PS		
	est.	se	p-value	est.	se	p-value	est.	se	p-value	est.	se	p-value
Gender (female)	-0.3811	0.1091	0.0005	-0.3793	0.1088	0.0005	-0.3792	0.1088	0.0005	-0.3778	0.1088	0.0005
Other drugs (No)	-0.5922	0.1481	0.0000	-0.5823	0.1475	0.0000	-0.5811	0.1474	0.0000	-0.5811	0.1474	0.0000
Drinking peers (No)	-0.3922	0.1146	0.0069	-0.3126	0.1143	0.0062	-0.3118	0.1142	0.0063	-0.3117	0.1142	0.0064
Neg. life ev.(No)	-0.2237	0.1187	0.0596	-0.2258	0.1184	0.0566	-0.2258	0.1184	0.0564	-0.2257	0.1184	0.0565
Phy. ab. (No)	-0.4707	0.1795	0.0087	-0.4703	0.1789	0.0086	-0.4683	0.1788	0.0088	-0.4680	0.1788	0.0088
Par. alc. ab. (No)	-0.3762	0.1073	0.0005	-0.3819	0.1070	0.0004	-0.3814	0.1069	0.0004	-0.3814	0.1069	0.0004
AIC		2941.99			2932.61			2932.74			2932.48	
BIC		3151.22			3019.47			3013.10			3011.76	

Table 5.5 contains estimated covariate effects, standard errors (se) and the corresponding p-values of the estimated covariate effects obtained from a model with a discrete baseline hazard (dummies) and the models with a smooth baseline hazard. Duration was modeled using dummy variables and also by three different smoothing methods, namely, penalised cubic regression spline, thin plate regression spline and p-spline. There appears to be very small differences amongst the parameter estimates. All the estimated effects are significantly (p-value ≤ 0.05) related to time to first alcohol intake except ever experiencing negative life events. The estimated covariate effects do not change signs across the models with different time specifications. Both the AIC and BIC do not give us enough evidence that the smoothing methods are significantly different from one another. The estimated coefficients and the standard errors obtained in this study are different from Sithuba (2017, unpublished)'s study. The effect of negative life event was statistically significant in (Sithuba, 2017)'s study which is not the case in our study. Most of the estimated coefficients (though the signs are similar) and standard errors are bigger in magnitude when compared to what was obtained in this study.

5.2.3 Model diagnostics

Calibration plot

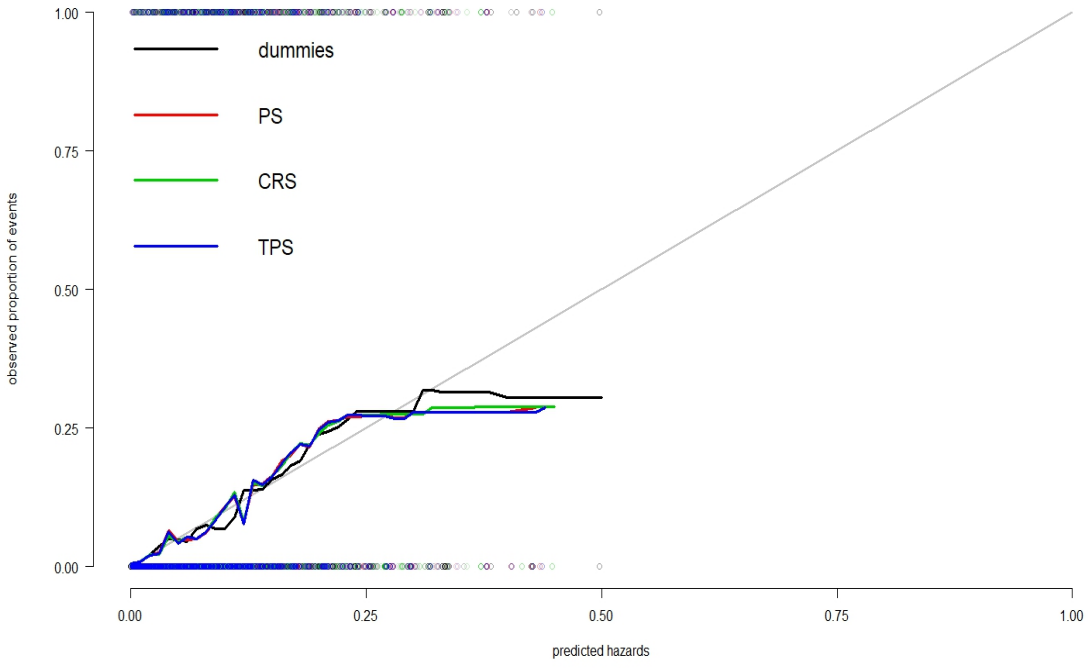


Figure 5.9: Calibration plot for the models obtained from dummies, PS, CRS and TPS.

Figure 5.9 is a calibration plot that is used to assess the performance of the fitted models by describing the differences between the observed hazard probabilities and the predicted hazard rates. If a model fits well, it should lie perfectly on the 45 degree line. On the plot, all the fitted models do not fall perfectly on the 45 degree line but have moderate deviations from the 45 degree line, which still indicate a moderate fit.

Residual analysis

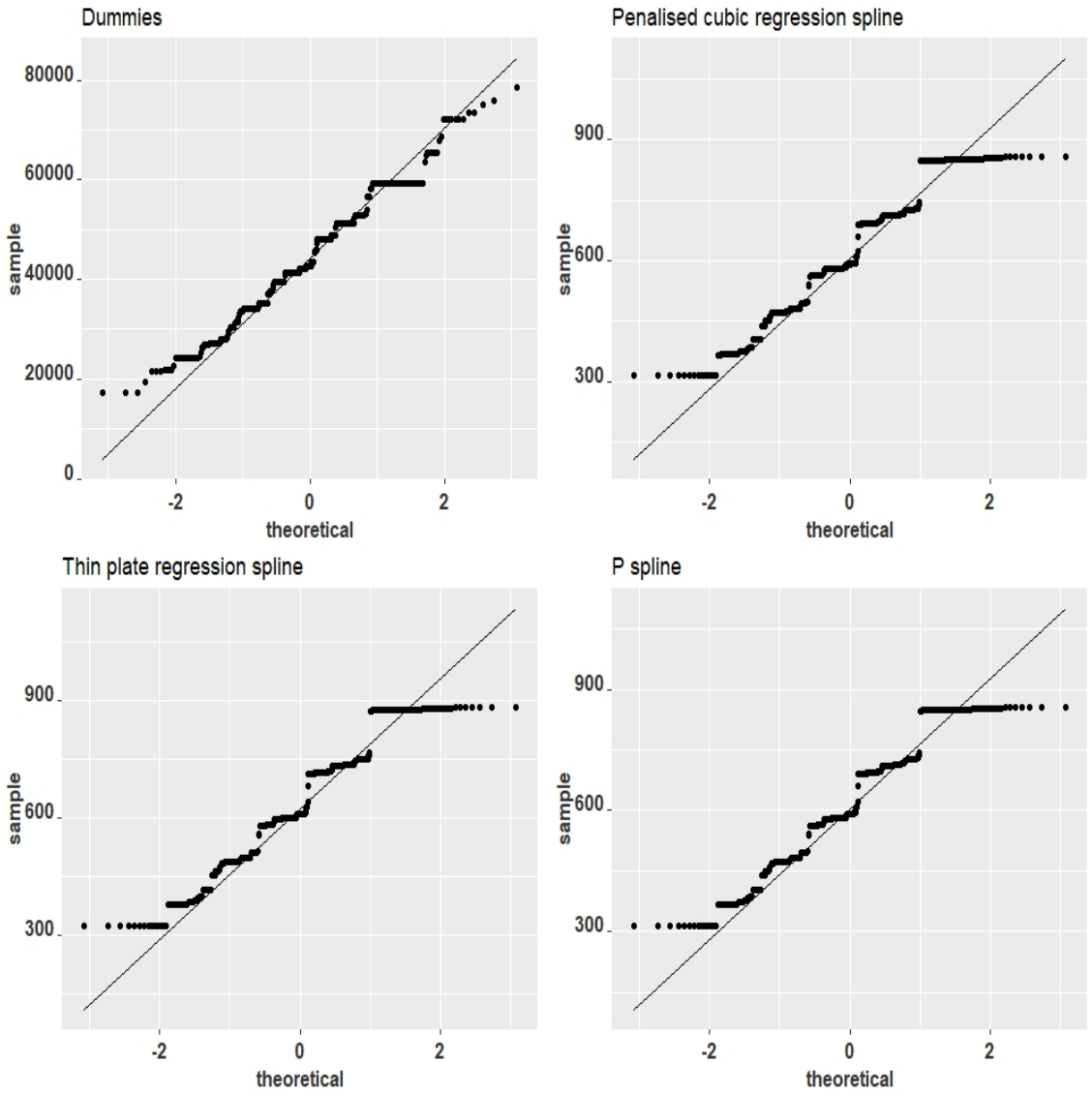


Figure 5.10: Normal quantile-quantile plots of the adjusted residuals obtained from the analysis of alcohol intake.

Figure 5.10 shows the normal quantile-quantile plots of the adjusted deviance residuals obtained from the model with the baseline modeled as dummy variable as well as the models with the smooth baseline. From this plot, it can be deduced that all the models performed moderately

well.

Receiver operating curves

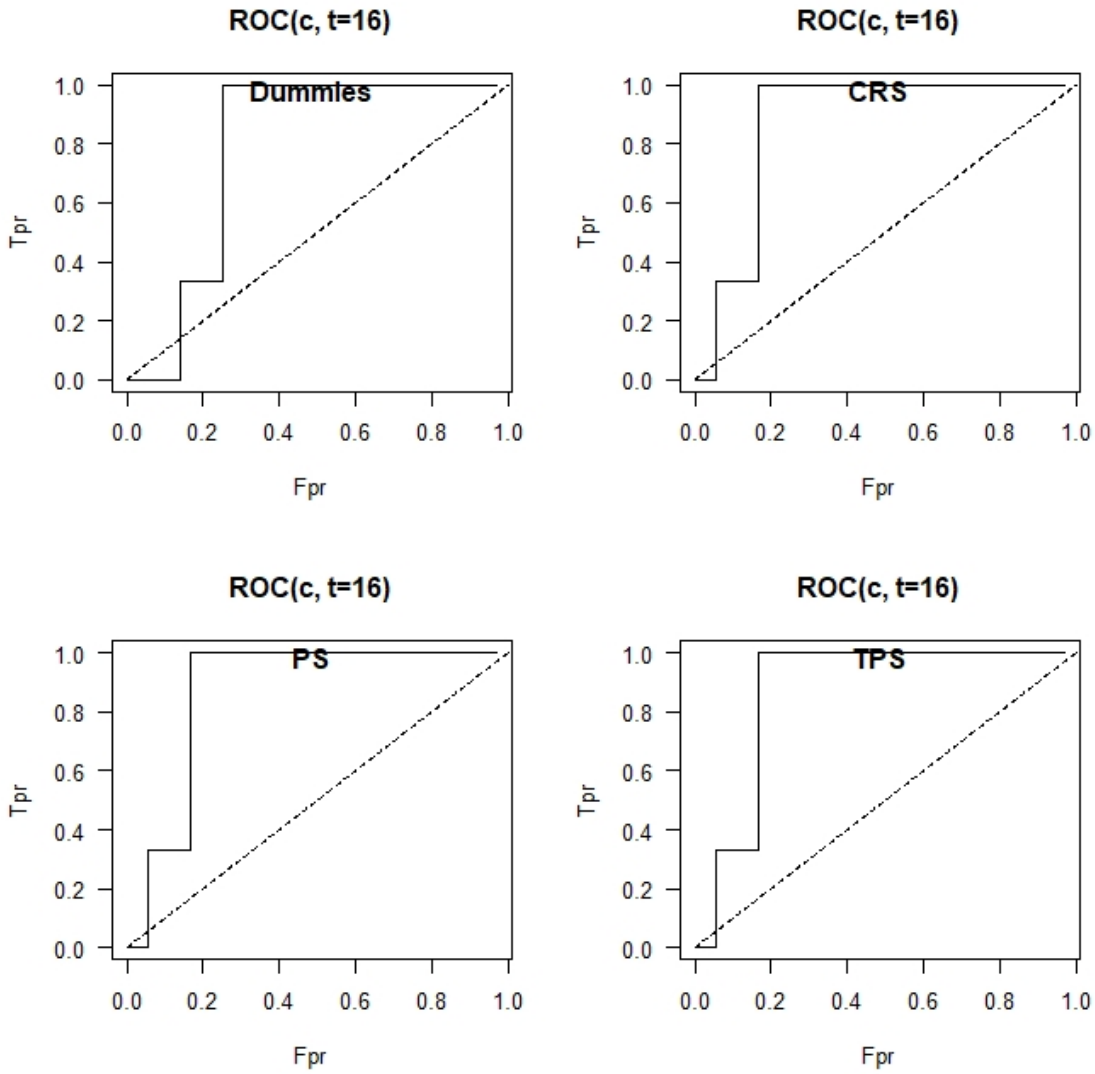


Figure 5.11: ROC curves (calculated from the test data) that were obtained from the all the fitted models at time $t = 16$.

Figure 5.11 display ROC curves obtained from the models with a discrete baseline hazard and the models with a smooth baseline hazard at time $t = 16$. A model that predicts better than chance should be above the diagonal line. The diagonal line represents a discrete survival model without covariates (AUC $= 0.5$). From Figure 5.11, it is seen that all the ROC curves for all the fitted models are above the diagonal line indicating that all the models show a good predictive performance. Furthermore, the AUC values at $t = 16$ calculated using the test data, were 0.787 for the model with the discrete baseline hazard and 0.8704 for all the models with a smooth

baseline hazard. This confirms the results obtained from the ROC curves. Finally the results from the C^* indices for all models indicate that all the models showed to be performing well. The C^* indices are 0.7492, 0.7862, 0.7862 and 0.814 for the CRS, PS, TPS, and the dummies respectively.

5.3 Interpretation of the results based on the model with a p-spline baseline hazard

5.3.1 Interpretation of the parameter estimates

Table 5.6: : Parameter estimates (baseline category) for the PS model: est (estimate), se (standard error).

Predictor	est.	se	p-value	Hazard ratios (95% CI)
Gender (female)	-0.3778	0.1088	0.0005	0.6854 (0.5537-0.8483)
Other drugs (No)	-0.5811	0.1474	0.0000	0.5593 (0.4190-0.7466)
Drinking peers (No)	-0.3117	0.1142	0.0064	0.7322 (0.5854-0.9159)
Negative life event (No)	-0.2257	0.1184	0.0565	0.7980 (0.6327-1.0063)
Physical abuse (No)	-0.4680	0.1788	0.0088	0.6263 (0.4411-0.8891)
Parental alcohol intake (No)	-0.3814	0.1069	0.0004	0.6829 (0.5538-0.8420)

Based on the estimates obtained from the p-spline model specification, the conditional probability of drinking alcohol for the first time is significantly lower for female students (p-value = 0.005) at 5% level of significance. The hazard ratio for this group was 0.69, indicating that this group is 31% less likely to drink alcohol very early in life. At 95% confidence level, the hazard for females ranges from 15% to 45% lower than that for males. Figure 5.12 is a graphical depiction of this finding.

The negative estimate (-0.5811) and the p-value <0.05 for not consuming other drugs indicate that the students in this category have a significantly lower chance of early intake of alcohol. We are 95% confident that for this group, the hazard for initiation of alcohol intake is between 25% and 58% lower than for those students who are consuming other drugs. The same pattern of results is seen in Figure 5.12.

It is also observed from Table 5.6 that having peers that drink is associated with higher chances of early drinking. The probability of drinking early in life at any given time for students who do not have drinking peer is between 8% and 41% lower than for students who have drinking

peers at 95% confidence level. Figure 5.12 is in agreement with this finding.

The effect of ever experiencing a negative life was found not be significant (p-value = 0.06) in describing the hazard for early onset of alcohol intake at 5% level of significance. Moreover, at 95% confidence level, the hazard ratio interval contains 1 and this shows that this explanatory variable does not have a significant impact on the hazard of alcohol intake. The relatively small differences between the predicted hazard rates as seen in Figure 5.13 are in concordance with this result.

From Table 5.6 it seen that the hazard of drinking alcohol early in life is 19% and 56% lower for students who did not experience any physical abuse in their lives than for those who experienced any form of physical abuse at 95% level of confidence. A p-value of 0.008 for this predictor indicates that this effect had a significant impact on the hazard and Figure 5.13 is in support of this finding.

Finally, having parents that drink alcohol is observed to be associated with a high estimated risk of early onset of drinking alcohol. Controlling for the other variables in the model, we are 95% confident that the hazard of alcohol is between 16% and 45% lower for the baseline category. Figure 5.13 shows the same pattern.

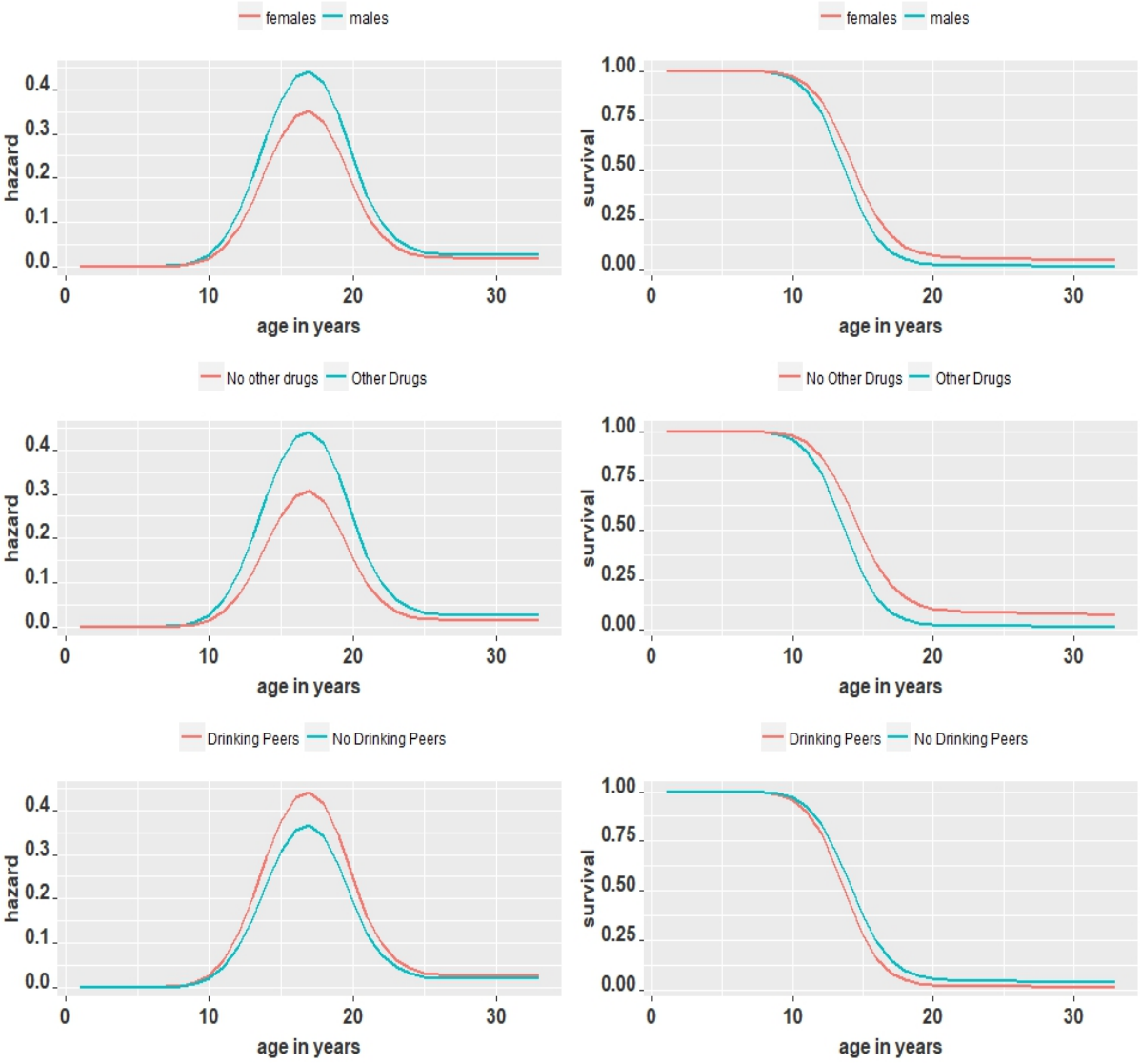


Figure 5.12: The discrete hazard function and the corresponding survival function probability that describe the risk for early alcohol consumption by gender, consumption of other drugs and having drinking peers

Figure 5.12 depicts a collage of fitted hazard probability and the corresponding survival probability by gender, consumption of other drugs and having drinking peers. The evidence from Figure 5.12 suggests that female students are less likely to start drinking very early in life. The probability of taking alcohol for the first time is considerably lower for females than for males across all the intervals. Furthermore, the big gap between the estimated hazards reveals that gender is a significant predictor of drinking alcohol early in life and this is supported by the small p-value (p-value$<$ 0.05) of this estimate as seen in Table 5.6. The estimated median lifetime for age at first alcohol intake is lower for males (13 years) than their female (14 years) counterparts as seen in the fitted survivor function that describes the chance of early intake of alcohol intake by gender.

The fitted hazard function that describes the risk of early onset of alcohol intake by drinking peers seems to show that having drinking peers is related to higher risk of drinking very early in life compared to not having any drinking peers. This result is confirmed by the negative estimated effect of this variable seen in Table 5.6. According to the survivor probability by consumption of other drugs, at 15 years 25% of the students who consumed other drugs survived versus about 49% of those students who did not drink alcohol.

A comparison of the estimated hazard rates for students who consumed any form of drugs and those who have never used drugs reveals a higher probability of early alcohol consumption for those who consumed drugs than those who did not consume any form of drugs. The considerably huge difference in the hazard rates suggest that consumption of drugs is strongly associated with early onset of alcohol intake (p-value =0.0000 as seen in Table 5.6). It is seen from the survivor function that explains the risk of drinking early by drinking peers that the students who had drinking peers have a lower likelihood to survive that the students who did not have any drinking peers.

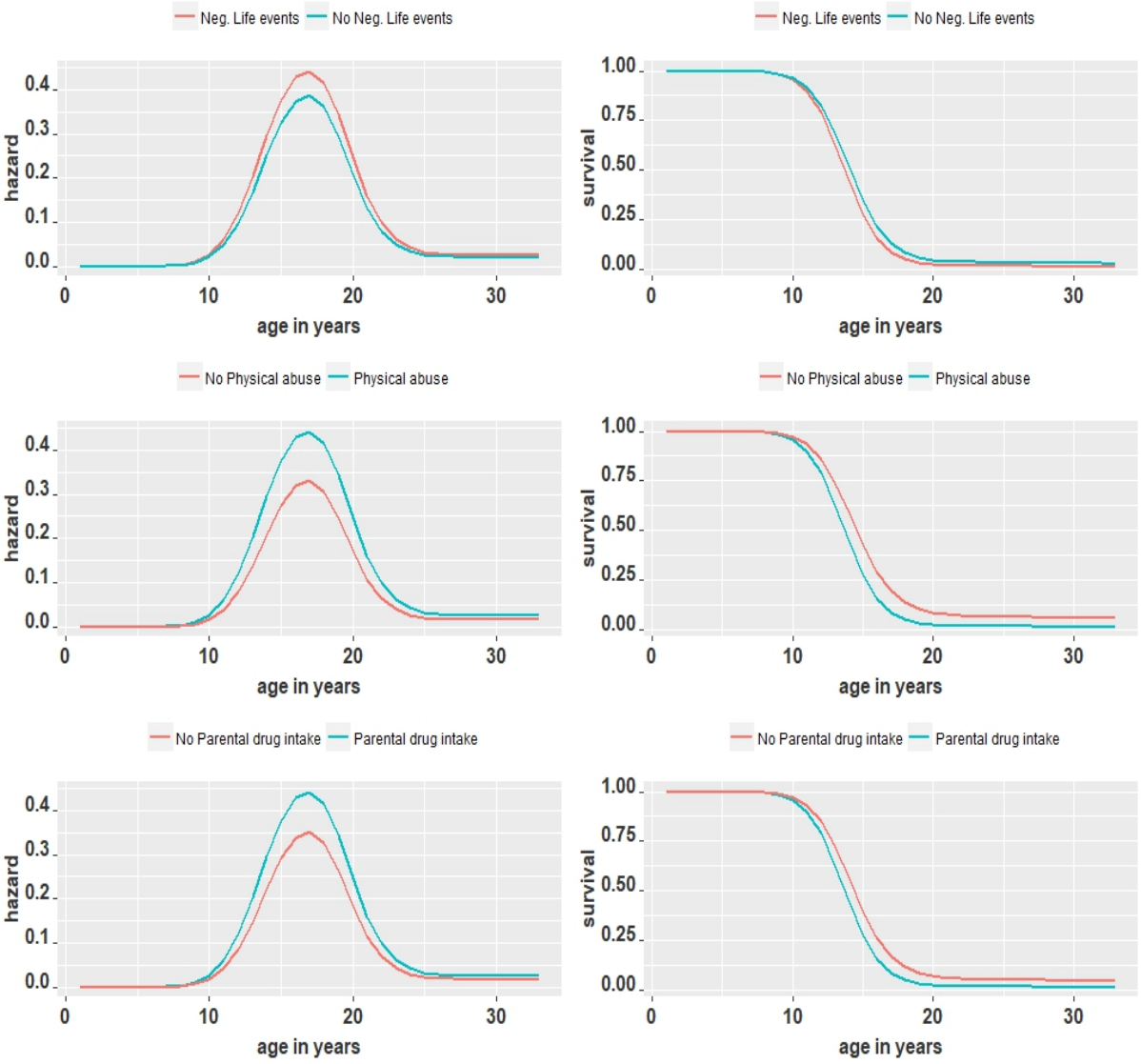


Figure 5.13: Predicted discrete hazard function and the corresponding survival function that describe the risk for early alcohol consumption by negative life event, ever experiencing physical abuse and parents who take any form of drugs.

Figure 5.13 shows estimated hazard and survivor probability for negative life event, eve experiencing physical abuse and parental drug intake. The small gap between the predicted hazard rates for ever experiencing a negative life is an indication that the effect of this covariate on the hazard of alcohol initiation is not significant as depicted in 5.13. The 95% confidence interval of the hazard ratio of this covariate as observed in Table 5.6 is in agreement with this result because the value 1 was included in the estimated interval, which is an indication that effect of this predictor on the hazard of alcohol initiation is not significant. The survivor curves for negative life experience suggest that the chances of survival for students who have experienced a negative life event is almost similar to that of students who did not experience any negative life event.

Inspection of the hazard rate for parental drug intake indicate that the students whose parent(s) do not drink alcohol have a lower likelihood of early onset of alcohol intake. This is in concordance with the negative coefficient observed for this category as shown in Table 5.6. The fitted survivor function for this covariate suggests that students whose parent(s) do not take any form of drug have a higher likelihood to survive than their counterparts whose parent(s) take drugs. At 15 years the survival probability for the latter group is 25% while it is 35% for the former group.

The predicted hazard rates for experiencing physical abuse as seen in Figure 5.12 confirms the finding indicated in Table 5.6 that a student who has suffered physical abuse in his or her own life is more likely to drink alcohol very early in life. The survivor probability for students who have ever experienced physical abuse is consistently lower across all time points than those students who have not experienced any form of physical abuse.

5.4 Concluding remarks

In this chapter the results obtained from the analysis of duration to first alcohol intake of students in Thohoyandou were presented. For the analysis, a proportional continuation ratio model where the baseline hazard was modeled as a dummy variable and also by using three variants of penalised regression splines was employed. The purpose was to check if all the models will produce similar results.

The magnitude of the estimated effects obtained from the four models are almost equal and the signs are similar and therefore there seems to be no major differences in the model fits with different specifications of the baseline hazard. The results obtained in this study are different to the results obtained in (Sithuba, 2017)'s study. Diagnostic results attained from the calibration plot, adjusted residuals, ROC curves and AUC revealed that no model in particular seemed to be performing better than the other models.

The results obtained from the analysis suggest that being a female, not having parents and peers who drink and not consuming drugs is associated with lower chances of alcohol initiation. Ever experiencing a negative event in life before did not have a significant impact on the hazard of drinking alcohol early in life. This is contrary to the findings in (Sithuba, 2017), as this predictor was found to have a significant impact on the hazard of alcohol initiation. This could

be as a result of adding a penalty term in the model objective.

Chapter 6

Conclusions

6.1 Conclusions

This study attempted to illustrate and compare the accuracy of penalised regression splines and dummy variables in modeling the baseline hazard function. Three variants of the penalised regression splines: penalised cubic regression spline, p-spline and thin plate regression spline. To illustrate the comparison of these splines in smoothing the baseline hazard, two simulation experiments with different transformations of the baseline hazard function were employed. In the first experiment we considered a logarithmic baseline hazard function and a gamma baseline was used in the second simulation experiment.

From both experiments, the RMSE analysis suggested that the penalised regression splines performed better than the model with the discrete baseline hazard (modeled using a dummy variable). Inspection of the plots obtained from the simulation experiments are in concord with the RMSE results, albeit no specific smoothing method showed to outperform the other models. It was also observed that the splines performed relatively better than the dummies in settings where the ratio of the sample size to the time intervals becomes smaller. When the ratio of the sample size to the number of intervals increases, all the models appeared to be improving. Therefore in cases where the ratio of the sample size to the number of time points is small, one should consider smoothing the baseline hazard to avoid numerical problems associated with the estimation of the baseline hazard function.

The afore mentioned methods were also applied using data on duration to first alcohol intake that were obtained from students from University of Venda and Vhembe FET college. Duration was discretized into one year intervals that resulted into 24 time points. The baseline life table

estimates showed that the discrete baseline hazard had an irregular pattern and thus, it was hard to interpret but a reasonable interpretation could be derived from the smooth baseline hazards. From the plot, a sharp increase of the hazard of drinking alcohol for the first time was seen, peaking at 17 years and a steady increase was observed until 25 years and subsequently plateaus until the end.

The findings from the regression analysis indicated no significant differences in the estimated coefficients, standard errors and p-values of all the fitted models. This observation is further confirmed by the calibration plot that revealed that all the models performed moderately well. The signs of the coefficients were similar. Both the AIC and the BIC showed not much difference among the estimated methods.

All the estimated effects, except having experienced a negative life event, obtained from all the models had a significant impact on the probability of drinking alcohol for the first time. According to the signs of the estimates, being male, ever experience physical abuse, having drinking parent(s) and drinking peers accelerate the onset of alcohol debut. When a student does not use any drugs, he or she is significantly less likely to start drinking very early in life. Parents have to be informed about the impact of their alcohol behaviour on their children. Students who have suffered any form of physical abuse should be encouraged to seek therapy from professionals who are trained to offer such kind of help. And finally students should also be educated about the dangers of consuming alcohol.

6.2 Limitations of the study and Future research

This study only focussed on the spline based approaches in the modeling of the baseline hazard function. Perhaps as future research a comparison between spline based approach and other approaches such as the radial basis, kernel functions or a dynamic model with a random walk prior can be investigated.

References

[1] Adebayo S.B and Farhmeir L., "Analysing child mortality in Nigeria with geoadditive discrete-time survival models", *Statistics in Medicine*, 24, pp. 709-728, 2005.

[2] Akaike H., "A new look at the statistical model identification", *IEEE Transactions on Automatic Control*, 19(6), pp. 716-723, 1974.

[3] Allison, P.D., "Discrete-time methods for the analysis of event histories", *American Sociologica Association*, 13,pp 61-98 , 1982.

[4] Allison, P.D., "Survival analysis using SAS: Apractical:A Practical guide", *American Sociologica Association*, Vol 13,pp 61-98 , 1982.

[5] Berger M. and Schmid M., "Semiparametric regression for discree time-to-end data", *Statistical Modelling*, 18(3-4), pp. 322-345, 2018.

[6] Biggeri L. ; Bini M. and Grill L., "The transition from University to work: A Multilevel Approach to the Analysis of the Time to obtain the first job", *Statistics in Society*, 164(2), pp. 293-305, 2001.

[7] Campolieti M., "Bayesian estimation and smoothing of the baseline hazard in discrete time duration models", *Review of Economics*, 85(4), pp. 685-694, 2002.

[8] Chauke T.M.; Van der Heever H. and Hoque M.E., "Alcohol use amongst learners in rural high school in South Africa, *African Journal of Primary Health Care and Family Medicine* , 7(1), pp. 1-6, 2015.

[9] Cox D.R., "Regression models and life-tables", *Royal Statistical Society*, 34(2), pp. 187-220, 1972.

[10] DeWit D.J.; Adlaf E.M; Offord D.R and Ogborne A.C, "Age at first alcohol use: a risk factor for development of alcohol disorders", *American Journal of Psychiatry*, 157(5), pp. 745-750, 2000.

[11] Durrleman S. and Simon R., "Flexible regression models with cubic spline", *Statistics in Medicine*, 8(5), pp.551-561, 1989.

[12] Efron B., "Logistic, Survival analysis and the Kaplan Meier Curve", *American Statistical Association*, 83(402), pp. 414-425, 1988.

[13] Eilers P.H.C. and Marx B.D.,"Flexible smoothing B-splines and penalties", *Statistics Science*, 11(2), pp. 89-121, 1996.

[14] Eisen E.A.; Aggaliu I.; Thurston S.W.; Coull B.A. and Checkoway H., "Smoothing in cohort studies: An illustration based on penalised splines", *Occupation and Environmental Medicine*, 61(10), pp. 854-860, 2004.

[15] Farhmeir L. and Knorr-held L, "Dynamic discrete-time duration model (Revised)", *Sonderforschungsbereich*, Paper 4, 1997.

[16] Fahrmeir L. and Wagenpfeil S., "Smoothing Hazard functions and time-varying effects in discrete duration and competing risks model", *American Statistical Society*, 91(436), pp. 1584-1594, 1996.

[17] Gray R.T., "Flexible methods for analysing survival data using splines", *American Statistics Society*, 87(420), pp. 942-951, 2016.

[18] Govindarajulu U.S; Spielgeman D.; Thurston S.W.; Ganguli B. and Eisen E.A., "Comparing smoothing techniques in Cox model for exposure-response relationships", *Statistics in Medicine*, 26, pp. 3735-3752, 2007.

[19] Groll A. and Tutz G., "Variable selection for Generalized Linear Mixed Models by L_1 Penalised Estimation", *Stats Comp.*, 24, pp. 137-154, 2014.

[20] Harmele A., "Regression analysis for discrete-time event history or failure data", *Statistical Papers*, 27, pp. 77-86, 2013.

[21] Heargerty P.J. and Zheng Y., "Survival model predictive accuracy and ROC curves", *Biometrics*, 61(1), pp. 92-105, 2005.

[22] Hess W. and Persson M.,"Duration of Trade Revisited: Continuous time versus Discrete-time", *Empirical Economics*, 43(3), pp. 1083-1107, 2010.

[23] Kalbfleisch, J. and L. Prentice, R., "The Statistical Analysis of Failure Data", *Reliability, IEEE Transactions on Realibity*, 34, pp. 11, 1980.

[24] Kim Y. and Gu C., "Smoothing spline Gaussian regression: more scalable computation via efficient approximation", *Statistical Methodology*, 66(2), 2004.

[25] Kyei K.A. and Ramagoma M., "Alcohol consumption in South African universities: Prevalence and Factors at the University of Venda", *Journal of Social Science*, 36(1), pp. 77-86, 2013.

[26] Langova L., "Survival analysis for clinical studies",*PubMed*, 152(2), pp. 303-307, 2008

[27] Liang W. and Chikritzhs T., "Age at first use of alcohol and risk of heavy alcohol use: A population-based study", *BioMed Research International*, 2013, pp. 1-5, 2000.

[28] Manda S. and Meyer R., "Age at first marriage in Malawi: A Bayesian Multilevel analysis using a discrete time to event model", *Statistical Social Association*, 150, pp. 439-455, 2005.

[29] Mantel N. and Hankey B., "A logistic regression model for use with response data where the hazard function is time-dependent", *Communication Statistics: Theory and Methods*, 7(4), pp. 333-347, 1978.

[30] Nkhoma P. and Maforah F., "Drinking patterns among students at the self-catering residence in the University of Cape Town", *Urbanisation Health Newsletter*, 21(2), pp. 54, 1994.

[31] Pitkanen T.; Lyrra A. and Pulkkinen L., "Age of onset of drinking and the use of alcohol in adulthood: a follow-up study from age 8-42", *American Journal of Psychiatry*, 100, pp. 652-661, 2005.

[32] Ramsoomer L. and Morojele N.K., "Trends in alcohol prevalence, age of initiation and association with alcohol-related harm among South Africa", *SAMJ*, 102(7), pp. 609-612, 2012.

[33] Reiss P.T and Ogden R.T., "Smoothing parameter selection for a class of semiparametric linear models", *Statistical Methodology*, 71(2), pp. 505-523, 2009.

[34] Rizopoulos D., *Joint Models for Longitudinal and Time-to-event data: With Applications in R*, Chapman and Hall/CRC Biostatistics Series, 1st edition, 2012.

[35] Roshani D. and Ghaderi, "Comparing smoothing techniques for fitting non-linear effect of caovariate in Cox model", *Acta Informatica Medica*, 24(1), pp. 38-41, 2016.

[36] R Core Team, "R: A Language and Environment for Statistical Computing",*R Foundation for Statistical Computing*, Vienna, Austria, 2017, https://www.R-project.org/.

[37] SchelldorferJ.; Meier L. and Buhlmann P., "GLMMLasso: An algorithm for High-Dimensional Generalized Linear Mixed Models Using L_1 Penalization ", *Computational and Graphical Statistics*, 23(2), pp. 460-477, 2014.

[38] Schmid M. ; Tutz G. and Welchowski T., "Discrimination measures for discrete time-to-event predictions", *Economics and Statistics*, 7, pp. 153-164, 2018.

[39] Sithuba H.G, "The determinants of age at first alcohol use by students at tertiary institutions around Thohoyandou, *Unpublished honours dissertation*, 2017.

[40] Tayob S.M and Van der Heever H., "Alcohol use among students at the University of Limpopo, South Africa", *African Journal for Physical, Health Education, Recreation and Dance*, September (Suppl. 1), pp. 364-373, 2014.

[41] Tshitangano T.G. and Tosin O.H., "Substance use amongst secondary school students in a rural setting in South Africa: Prevalence and possible contributing factors", *African Journal of Primary Health Care and Family Medicine* , 8(2), pp. 1-6, 2016.

[42] Tutz, G. and Scmid, Daly P.W., *Modeling discrete time to event data*, Springer, Switzerland, 2016.

[43] Welchowski T. and M. Schmid, "discSurv: Discrete Time Survival Analysis", *R package version 1.3.1*, 2018, https://CRAN.R-project.org/package=discSurv.

[44] Wood S.N. and Augustin N.H., "GAMs with integrated model using penalised regression splines and applications to environmental modelling", *Ecological Modelling*, 157, pp. 157-177, 2002.

[45] Wood S.N., "Thin plate regression splines", *Journal of Royal Statistical Society*, 69, pp. 95-114, 2003.

[46] Wood S.N., "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalised linear models", *Statistical Methodology*, 73(1), 2011.

[47] Wood S.N., "General Additive Model: an introduction with *R*", *Chapman and Hall/CRC Press*, 1st edition, 2006.

[48] Wood S.N., "General Additive Model: an introduction with R", *Chapman and Hall/CRC Press*, 2nd edition, 2017.

[49] Wood S.N. ; Pya N. and Säfken B., "Smoothing parameter and model selection for general smooth models", *American Statistical Association: Theory and Methods*, 111(516), 1548-1575, 2016.

[50] Wood S., " mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation", *R package version 1.3.1*, 2018, https://CRAN.R-project.org/package=mgcv.

Appendices

Appendix A

```
#rm(list=ls())
library(foreign)
library(survival)
library(mgcv)
library(Ecdat)
library(dplyr)
library(caret)
library(discSurv)
library(ggplot2)
library(gamlss.mx)
library(mgcv)
library(MASS)
library(nlme)
library(glmmLasso)
library(Hmisc)
dataset = read.spss("C:/Users/mashamr/Desktop/afa.sav", to.data.frame=TRUE,
use.value.labels = TRUE)
#(dataset,2,table)
#dataset<dataset[(507),]
attach(dataset)
Tp<data.frame(cbind(Sex,Ethnicity, AcadL, MaritalS, Stress, IncomeF, ChildPR
Siblings, PQuali, ACS,OtherDr, Discipn, DrPeers, WelfDep, NegetiveLE, PhysicA
SexAb, RelationAdult, DisfunctF, PDrIntake, Institutions,Faculties))

apply(Tp,2,table)
NmaritalS<NULL
NEthn<NULL
AFI<NULL
```

```
CNS<NULL
Ctime<NULL
Pr<NULL
for (i in 1:length(AFA)){

  #print(MaritalS[i]==1)
  Pr[i]<ifelse(ChildPR[i]=="Urban",1,0)
  AFI[i]<ifelse(is.na(AFA[i])==TRUE,Age[i],AFA[i])
  CNS[i]<ifelse(is.na(AFA[i])==TRUE,0,1)
  NmaritalS[i]<ifelse(MaritalS[i]=="Single","Single","Ever_married")
  NEthn[i]<ifelse(Ethnicity[i]=="African","African","Other_race")

  if (AFI[i]==9) {
    Ctime[i]<1

  }else if(AFI[i]==10) {
    Ctime[i]<2

  }else if(AFI[i]==11) {
    Ctime[i]<3

  }else if(AFI[i]==12) {
    Ctime[i]<4

  }else if(AFI[i]==13) {
    Ctime[i]<5

  }else if(AFI[i]==14){
    Ctime[i]<6

  }else if(AFI[i]==15){
    Ctime[i]<7

  }else if(AFI[i]==16){
    Ctime[i]<8

  }else if(AFI[i]==17){
```

```
    Ctime[i]<9

}else if(AFI[i]==18){
    Ctime[i]<10

}else if(AFI[i]==19){
    Ctime[i]<11

} else if(AFI[i]==20){
    Ctime[i]<12
} else if (AFI[i]==21){
    Ctime[i]<13
} else if (AFI[i]==22){
    Ctime[i]<14
}else if (AFI[i]==23){
    Ctime[i]<15
}else if (AFI[i]==24){
    Ctime[i]<16
}else if (AFI[i]==25){
    Ctime[i]<17
}else if (AFI[i]==26){
    Ctime[i]<18
}else if (AFI[i]==27){
    Ctime[i]<19
}else if (AFI[i]==28){
    Ctime[i]<20
}else if (AFI[i]==29){
    Ctime[i]<21
}else if (AFI[i]==30) {
    Ctime[i]<22
}else if (AFI[i]==31){
    Ctime[i]<23
}else if (AFI[i]==32){
    Ctime[i]<23
}else{
    Ctime[i] <24
```

```
  }

}

#Ctime
#NmaritalS<-(factor(NmaritalS,
#levels = c(0,1),
#labels = c("Married", "Single")) )
table(NmaritalS,AFA)
#NEthn<-factor( NEthn,
#levels = c(0,1),
#labels = c("Other", "African"))
summary(Age)
table(NEthn)
table(Sex)
dataset<-data.frame(cbind(dataset,AFI,CNS, NEthn,NmaritalS,Ctime,Pr))
dataset$AFI<-as.integer(dataset$AFI)
summary(dataset$AFI)

LFTB <- lifeTable (dataSet=dataset, timeColumn="AFI", censColumn="CNS")

head(LFTB$Output, 25)
plot(x=1:dim(LFTB$Output)[1], y=LFTB$Output$hazard,
type="l", xlab="Time interval",
ylab="Hazard", las=1,main="Life table
estimated marginal hazard rates")
length(LFTB$Output[,4])
x<-seq(1:33)

LFTB$output<-cbind(LFTB$Output,x)



p1<-ggplot(aes(x=1:dim(LFTB$Output)[1],y=LFTB$Output$hazar),
data=LFTB$Output)
+ geom_line(size=1) +
```

```
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"))+
  xlab("age in years") + ylab("hazard")



lr.fit < gam(hazard~s(x,bs="ps"),data=LFTB$Output)
Testdata<data.frame(x=c(1:33))


Fits<predict(lr.fit,newdata=Testdata, type='response',se.fit=TRUE)


predicts<data.frame(LFTB$output,Fits)


p2<ggplot(aes(x=x,y=fit),data=predicts) +
geom_smooth(aes(ymin=fit 1.96*se.fit
,ymax=fit+1.96*se.fit),fill="gray80",size=1,stat="identity")
+geom_rug(sides="b")+
  theme(axis.text.x = element_text(colour="grey20",size=12,
  angle=0,hjust=.5,vjust=.5,face="bold"),


  xlab("age in years") + ylab("hazard")
```

```
library(gridExtra)

grid.arrange(p1, p2, ncol=2)

#plot(lr.fit)

dtLg <as.data.frame(dataLong (dataSet=dataset,
timeColumn="AFI", censColumn="CNS"))



family = binomial(link="logit")

#################################################################


################### More Elegant Method ######################
## Idea: start with big lambda and use the estimates of the
 previous fit (BUT: before
## the final re estimation Fisher scoring is performed!)
as starting values for the next fit;
## make sure, that your lambda sequence starts at a
value big enough such that all covariates are
## shrinked to zero;

## Using BIC (or AIC, respectively) to determine the
 optimal tuning parameter lambda
lambda < seq(200,0,by=5)


BIC_vec<rep(Inf,length(lambda))


# specify starting values for the very first fit;
pay attention that Delta.start has suitable length!
Delta.start<as.matrix(t(rep(0,18)))
```

```
Q.start<0.5

for(j in 1:length(lambda))
{
  print(paste("Iteration ", j,sep=""))
  glm3 < glmmLasso(y~1 +as.factor(NmaritalS)+as.factor(Sex)+as.factor(DrPeer
+as.factor(DisfunctF)
                        +as.factor(PDrIntake)+
  as.factor(NEthn)+as.factor(NegetiveLE)+as.factor(PhysicAb)
   +as.factor(SexAb)+as.factor(RelationAdult)
   +as.factor(WelfDep)  +
   as.factor(Discipn)+as.factor(OtherDr)
   +as.factor(Stress)+as.factor(IncomeF)
   +as.factor(Institutions),rnd = NULL,
   family = binomial(link = logit), data = dtLg,
   lambda=lambda[j], switch.NR=F,final.re=T,
   control = list(start=Delta.start[j,],smooth=
    list(formula=~ 1 + as.numeric(timeInt), nbasis=7,
            spline.degree=3, diff.ord=2, penal=10)
                  ,print.iter=TRUE,center=TRUE,steps=4000,
                  epsilon=1e 6,eps.final=1e 6,maxIter=2000))
  print(colnames(glm3$Deltamatrix)[2:18]
  [glm3$Deltamatrix[glm3$conv.step,2:18]!=0])
  BIC_vec[j]<glm3$aic
  Delta.start<rbind(Delta.start,glm3$Deltamatrix[glm3$conv.step,])
  #Q.start<c(Q.start,glm3$Q_long[[glm3$conv.step+1]])
}

opt3<which.min(BIC_vec)
print(lambda[opt3])



glm3_final<glmmLasso(y~1 +as.factor(NmaritalS)+as.factor(Sex)+
as.factor(DrPeers)
```

```
+as.factor(DisfunctF) +as.factor(PDrIntake)+
as.factor(NEthn)+as.factor(NegetiveLE)+as.factor(PhysicAb)
+as.factor(SexAb)+as.factor(RelationAdult)
+as.factor(WelfDep)  + as.factor(Discipn)+as.factor(OtherDr)+
as.factor(Stress)+as.factor(IncomeF)
+as.factor(Institutions),rnd = NULL,
family = binomial(link = logit), data = dtLg,
lambda=lambda[opt3], switch.NR=F,final.re=T,
control = list(start=Delta.start[opt3,],q_start=Q.start[opt3],smooth=
 list(formula=~ 1 + as.numeric(timeInt), nbasis=7,
 spline.degree=3, diff.ord=2, penal=15),print.iter=TRUE,
 center=TRUE,steps=4000,epsilon=1e 6,eps.final=1e 6,maxIter=2000))


print(summary(glm3_final))

## plot coefficient paths
par(mar=c(6,6,4,4))
plot(lambda,Delta.start[2:(length(lambda)+1),2],
type="l",ylim=c( 1e 1,1e 1),ylab=expression(hat(beta[j])))
lines(c( 1000,1000),c(0,0),lty=2)
for(i in 3:7){
  lines(lambda[1:length(lambda)],Delta.start[2:(length(lambda)+1),i])
}
abline(v=lambda[opt3],lty=2)


#psmod<gam(y~s(as.numeric(timeInt), bs="ps",
 m=c(2,3))+Sex+OtherDr+ DrPeers+ NegetiveLE+ PhysicAb +PDrIntake,
 method="REML", scale=1, family = binomial(link = logit),  data =dtLg)
#summary(psmod)

Crmod<gam(y~s(as.numeric(timeInt), bs="cr", k=12)
+NmaritalS+DisfunctF+Sex+OtherDr+ DrPeers+ NegetiveLE+ PhysicAb
+PDrIntake+Discipn,method="REML", scale=1,
family = binomial(link = logit),  data =dtLg)
summary(Crmod)
```

```
##shortadjusted residuals
set.seed(0)
Indizes < sample(unique(dataset$Id), 500)
randSample < dataset[unlist(sapply(1:length(Indizes),
function(x) which(dataset$Id==Indizes[x]))),]


datLong < dataLong(dataSet=randSample,
                        timeColumn="AFI", censColumn="CNS",
                        timeAsFactor=FALSE)
gamFit<gam(y~s(as.numeric(timeInt), bs="cr",
k=12)+NmaritalS+DisfunctF+Sex+OtherDr+
 DrPeers+ NegetiveLE+ PhysicAb +PDrIntake+Discipn,
 method="REML", scale=1, family = binomial(link = logit),
   data =datLong)

#gamFit < gam(y ~ s(timeInt, bs="ps", m=c(2,3))+
Sex + OtherDr+DrPeers+NegetiveLE+PDrIntake    ,
data=datLong, family="binomial")
hazPreds < predict(gamFit, type="response")
adj  < adjDevResidShort (dataSet=datLong,
 hazards=hazPreds)$Output$AdjDevResid
l<quantile(adj,0.01)
u<quantile(adj,0.99)
 tadj<adj[adj>l & adj<u]

plot(density(tadj),ylim=c(0,0.005),
      main="Density comparison: Normal vs nonparametric estimate",
       lwd=2, las=1, col="red")
dnorm1 < function (x) {dnorm (x, mean=mean(tadj), sd=sd(tadj))}
curve (dnorm1, n=500, add=TRUE, col="black", lwd=2)
legend("topright", legend=c("Normal", "Nonpar"), lty=1,
 lwd=2, col=c("black", "red"))


##Gender
```

```
#psp<gam(y~s(as.numeric(timeInt), bs="ps", m=c(2,3))+Sex + OtherDr+DrPeers+
NegetiveLE+ PhysicAb +PDrIntake, data=dtLong, method =
 "REML", scale =1, family=binomial(link=logit))
Testdata<data.frame(timeInt=factor(1:33),
            Sex=factor(rep("Male", 33), levels=c("Male", "Female")),
            OtherDr=factor(rep("Yes", 33), levels=c("Yes", "No"))
            ,DrPeers=factor(rep("Yes", 33), levels=c("Yes", "No"))
            ,NegetiveLE=factor(rep("Yes", 33), levels=c("Yes", "No")),
            PhysicAb=factor(rep("Yes", 33), levels=c("Yes", "No"))
            ,PDrIntake=factor(rep("Yes", 33), levels=c("Yes", "No")) )

Testdata1<data.frame(timeInt=factor(1:33),
 Sex=factor(rep("Female", 33), levels=c("Male", "Female")),
OtherDr=factor(rep("Yes", 33), levels=c("Yes", "No"))
 ,DrPeers=factor(rep("Yes", 33), levels=c("Yes", "No")),
NegetiveLE=factor(rep("Yes", 33), levels=c("Yes", "No")),
PhysicAb=factor(rep("Yes", 33), levels=c("Yes", "No"))
,PDrIntake=factor(rep("Yes", 33), levels=c("Yes", "No")) )

Fits < predict(Crmod, newdata=Testdata, type="response")
Fits1 < predict(Crmod, newdata=Testdata1, type="response")



SurvM < cumprod(1 Fits)
SurvF< cumprod(1 Fits1)

predicts<data.frame(Testdata,Fits,SurvM)
predicts1<data.frame(Testdata1,Fits1,SurvF)



p3<ggplot(aes(x=as.numeric(timeInt),y=Fits,color="males"),
data=predicts) + geom_line(size=1) +
```

```
  geom_line(aes(x=as.numeric(timeInt),y=Fits1,color="females"),
  data=predicts1,size=1)+
  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"))+
  xlab("age_in_years") + ylab("hazard")


ps3<ggplot(aes(x=as.numeric(timeInt),y=SurvM,color="males")
, data=predicts) + geom_line(size=1) +
  geom_line(aes(x=as.numeric(timeInt),y=SurvF,color="females")
  , data=predicts1,size=1)+  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"))+
  xlab("age_in_years") + ylab("survival")


grid.arrange(p3,ps3,p4,ps4,p5,ps5,  ncol=2)


#grid.arrange(p6,ps6,p7,ps7,p8,ps8,  ncol=2)
```