University of Venda

DISCRETE SURVIVAL MODELS WITH FLEXIBLE LINK FUNCTIONS FOR AGE AT FIRST MARRIAGE AMONG WOMEN IN SWAZILAND

By

Nevhungoni Thambeleni Portia (Student No: 16011737)

SUBMITTED IN FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN STATISTICS
AT
UNIVERSITY OF VENDA
THOHOYANDOU, SOUTH AFRICA
NOVEMBER 2018

UNIVERSITY OF VENDA
DEPARTMENT OF
STATISTICS

The undersigned hereby certify that they have read and recommend to the Faculty of Science for acceptance a dissertation entitled "**Discrete survival models with flexible link functions for age at first marriage among women in Swaziland**" by **Nevhungoni Thambeleni Portia (Student No:** 16011737) in fulfillment of the requirements for the degree of **Master of Science in Statistics**.

Dated: November 2018

Supervisor: _____
Dr A. Bere

Co-supervisor: _____
Dr C Sigauke

Readers: _____

_____

UNIVERSITY OF VENDA

Date: **November 2018**

Author: **Nevhungoni Thambeleni Portia (Student No:** 16011737**)**

Title: **Discrete survival models with flexible link functions for age at first marriage among women in Swaziland**

Department: **Statistics**

Degree: **M.Sc.** Convocation: **November** Year: **2018**

Permission is herewith granted to University of Venda to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.

Signature of Author

I dedicate this work to God, my father Late MJ Nevhungoni, my loving husband Vele KR (who always believed in me since I started this project), my son Murendeni and my daughter Muvhuya.

Table of Contents

List of Tables

List of Figures

Abstract

This study explores the use of flexible link functions in discrete survival models through a simulation study and an application to the Swaziland Demographic and Health Survey (SDHS) data. The objective of the research study is to perform simulation exercises in order to compare the effectiveness of different families of link functions and to construct a discrete multilevel survival model for age at first marriage among women in Swaziland using a flexible link function. The Pareto hazard model, Pregibon and Gosset families of link functions were considered in models with and without unobserved heterogeneity. The Pareto model where the family parameter is estimated from the data was found to outperform the other models, followed by the Pregibon and the Gosset family of link functions. The results from both simulation study and real data analysis of the SDHS data illustrated that, misspecication of the link function causes bias on the estimation of results. This demonstrates the importance of choosing the right link. The findings of this study reveal that women who are highly educated, stay in the Manzini and Shiselweni region, those who reside in urban areas were more likely to marry later compared to their counterparts in Swaziland. The results also reveal that the proportion of early first marriages is declining since the difference among birth cohorts is found to be very high, with women of younger cohorts getting married later compared to older women.

Keywords: *Flexible families of link functions, discrete survival models, heterogeneity, simulation study, misspecication of the link function.*

Abbreviations

AFM	Age at first marriage.
DHS	Demographic and Health Survey.
SDHS	Swaziland Demographic and Health Survey.
BDHS	Bangladesh Demographic and Health Survey.
PH	Proportional Hazard.
GLM	Generalized Linear Model.
cdf	Cumulative distributive function.
LR	Likelihood Ratio.
AIC	Akaike information criterion.

Acknowledgements

I would like to acknowledge the continuous support given to me by my supervisor Dr A Bere during this research, without him this work would never have come into existence. His guidance, insightful criticisms and patient encouragement helped me in all the time of research and writing of this thesis.

I would also like to thank my core supervisor Dr C Sigauke for his suggestions, support and motivation. Special thanks goes to CSIR DST-Interbusary Support (IBS) bursary office for the financial support.

I am grateful to my loving mom for her patience and love. I wish to thank my in laws Mrs Nyambeni Matevhutevhu and Mrs Nyambeni Ester for being patient with me throughout this research.

Finally, I also wish to thank the following people: Mr Thakhani Ravele, Ms Mpfumali Phathutshedzo, Mr P Nematswerani, Ms V Ramarumo; my younger brother Elia Nevhungoni, my sister in law Langanani Nyambeni and also those who directly and indirectly helped me during this research.

Chapter 1

Introduction

1.1 Survival analysis

The term survival analysis refers to statistical methods that can be used to analyse data, where the focus or interest is on the duration of time until the occurrence of an event. For example, the event of interest can be occurrence of disease, marriage, divorce or school dropout. The main focus is on the time at which the subject experiences the event and that can be measured in years, months, days or other smaller units of time measurement (Cox and Oakes 1984, Clayton 1978, Mills 2011, Singer and Willett 2003). In the current study we are interested in modelling the time from birth until a woman gets married.

Survival data has two unique features which make it impossible to use the orthodox statistical techniques such as multiple linear regression. These are censoring and time varying covariates.

When the subject is not experiencing an event during an observational period, the particular subject is said to be censored. In that case we are not sure whether the

event will happen after the observational period or not. For example: suppose a group of individuals are followed for a period of one year to observe their age at first marriage. An individual who will not get married during that period is said to be censored, because we don't know what will happen to that individual after the study. This may be due to a number of reasons. For example: loss to follow up or non-occurrence of the event.

Generally there are three types of censoring which are used. Firstly we have the right censoring, which occurs when the subject leaves the study before the event of interest takes place. This type of censoring is commonly used because the true unobserved event is to the right of censoring time. It is divided into two types namely: Type l censoring (the study ends when a fixed time is reached with no event having occurred), Type ll censoring (the study ends when a fixed number of events has taken place among the subjects). Secondly we have the left censoring which occurs when the subject has already experienced the event before the study begins. Obviously if the event of interest is death it means that in that regard data cannot be left censored. Finally we have the interval censoring which occurs when the event of interest occurred to the subject but the time at which it occurred is not exactly known because it is between two values (Klein and Moeschberger (2006), Jenkins (2005), Collett (2015)).

In some situations a covariate may not necessarily be stable throughout the whole study. For example: if one wishes to examine the link between place of residence and age at first marriage, this can be complicated due to the fact that the subjects in the

study may move from one place of residence to another. Therefore place of residence can then be the time varying covariate in a statistical model. Other examples of time varying covariates are age and marital status.

Sometimes one might find that only those individuals who lie within a certain interval can be observed and that may result in sampling bias, we call it truncation. There are two types of truncation which are right truncation and left truncation. Right truncation occurs when the event of interest has already happened on the entire study population. Left truncation occurs when the subjects have been at risk of experiencing the event before entering the study (Guo (1993), Cox (2018), Pollock et al. (1989)). An example of left truncation is a life insurance policy where a study can begin at a fixed time but the policy holders will have been at risk of dying before the study.

In survival modelling, the time to event can be considered to be discrete or continuous. Here we start by discussing continuous survival models. When using continuous time models we assume that the time to event is measured as a continuous random variable.

Assuming that T is a continuous random variable representing the waiting time until the occurrence of an event. T is non-negative (*i.e.* $T \geq 0$) and it has the probability density function (pdf) denoted by $f(t)$ and cumulative density function (cdf) denoted by $F(t)$. The cdf is mathematically represented as

$$F(t) = \int_0^t f(u)du = P_r(T \leq t), \tag{1.1.1}$$

where t is the specific value of T and $P_r(T \leq t)$ is the probability that the waiting time until the event happens is less than or equal to the specific value of T. $F(t)$ gives the probability that an event of interest has occurred by duration t. For example: if we try to relate this with the current study $F(t)$ will give the probability that an individual got married by duration t. To get $f(t)$ we differentiate $F(t)$ with respect to t, then

$$\Rightarrow f(t) = \lim_{\Delta t \to 0} \frac{(F(t + \Delta t) - F(t))}{\Delta t}, \tag{1.1.2}$$

where Δt represents a small time interval. Relating this to the current study $f(t)$ will give the unconditional instantaneous probability that an individual got married in the interval $(t, \Delta t)$ and it is formally written as follows

$$f(t) = \lim_{\Delta t \to 0} \frac{P_r(t \leq T < t + \Delta t)}{\Delta t}. \tag{1.1.3}$$

Now the survival function denoted by $S(t)$ and the hazard function denoted by $\lambda(t)$ are defined as the two quantitative terms which are important in survival analysis. $S(t)$ focuses on surviving and $\lambda(t)$ focuses on failing of the subject. Relating to the current study, the survival function will give us the probability that an individual survives longer than time t without getting married (or being single) and the hazard function will give the instantaneous rate of getting married after time t given that the individual did not get married up to time t. The survival function is given by

$$S(t) = 1 - F(t) = P(T \geq t). \tag{1.1.4}$$

The relationship between the hazard, density and survival function is derived in the following way

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P_r(T \in [t, t + \Delta t | T \geq t))}{\Delta t} \tag{1.1.5}$$

$$= \lim_{\Delta t \to 0} \frac{P_r(T \in [t, t+\Delta t))}{P(T \geq t)} \div \Delta t \tag{1.1.6}$$

$$= \lim_{\Delta t \to 0} \frac{P_r(T \in [t, t+\Delta t))}{P(T \geq t)} \times \frac{1}{\Delta t} \tag{1.1.7}$$

$$= \lim_{\Delta t \to 0} \frac{P_r(T \in [t, t+\Delta t))}{\Delta t.S(t)} \tag{1.1.8}$$

$$= \frac{1}{S(t)} \lim_{\Delta t \to 0} \frac{P_r(T \in [t, t+\Delta t))}{\Delta t} \tag{1.1.9}$$

$$= \frac{f(t)}{S(t)}. \tag{1.1.10}$$

The relationship between the survivor function and the cumulative distribution function is that the survivor function is equal to 1 minus the probability of failure by time t (i.e. $S(t) = 1 - F(t)$), because total probability must sum to 1. According to Mills (2010), the continuous survival models have the following advantages. First, they are flexible and one is not required to choose a particular probability distribution in advance when using them. Second, the parameter estimates can still be produced for the evaluation of multiple explanatory variables regardless of the fact that the baseline hazard in the model remain unspecified. Finally, they have the ability to perform well when event time is recorded with a precise and fine metric such as an hour, a day or a week.

According to Hess and Person, (2010), continuous-time models such as the Cox model have three major problems which are as follows: They face problems when there are

many ties, leading to biased coefficient estimates and standard errors. It is also difficult to properly control for the unobserved heterogeneity and if such heterogeneity is not accounted for, it can cause spurious negative duration dependence of the estimated hazard function as well as parameter bias. Finally, the Cox model brings the restrictive and empirically questionable proportional hazard assumptions. With regards to this particular study, it also seems appropriate to consider discrete models since age at first marriage is reported in complete years.

According to Hess and Person (2010), the advantages of discrete survival models are as follows. First, there are no problems on handling ties and the unobserved heterogeneity can be easily controlled. Second, they have the ability to accommodate time varying covariates and the fact that they do not require proportional hazards (PH) assumptions like continuous time models. The disadvantage of discrete survival models is the fact that one needs to restructure data into smaller time-intervals which can be done by creating a subject period file. Such a file can become large and unruly, especially when the time intervals are short and the observation period is very long.

In the discrete-time framework, we focus on the probability that an event happens in a given time interval $[t_k, t_{k+1})$, where $k = 1, 2, ..., K$, and $t_1 = 0$, given that the event did not happen up to the beginning of the interval. The discrete time hazard function is defined as

$$\lambda_{ik} = P(T_i = t_k \mid T_i \geq t_k, \mathbf{x}_{ik}). \tag{1.1.11}$$

One may wish to model and investigate the influence of some covariates on duration times. In this regard the discrete duration model must include covariates and it is mathematically written as

$$\lambda(t|\mathbf{x}_{it}) = F(\gamma_{0t} + \mathbf{x}_{it}^T\mathbf{\Gamma}), \tag{1.1.12}$$

where $F(.)$ is a fixed link function, γ_{0t} is the base line hazard, $\mathbf{x}_{it}^T = (\mathbf{x}_{1,it}, ..., \mathbf{x}_{p,it})^T$ is a p-dimensional vector of covariates for observation $i\epsilon\{1, ..., N\}$. The baseline hazard represents the hazard when all the covariates take the value 0. It measures the effect of time on the hazard. The contribution of the covariates is absorbed in the term $\mathbf{x}_{it}^T\mathbf{\Gamma}$ and the $\mathbf{\Gamma} = (\mathbf{\Gamma}_1, ..., \mathbf{\Gamma}_p)^T$ are the regression coefficients (Tutz and Schmid, 2016).

1.2 Statement of the problem

The link function relates the response probability $P(T_i = t_k \mid T_i \geq t_k, \mathbf{x}_{ik})$ to the linear predictor. According to Hess et al. (2014), the link function has to be chosen in such a way that it is non-negative *i.e.* $0 \leq \lambda(t|\mathbf{x}_{it}) \leq 1$ for all t, because the discrete-time hazard is a conditional probability. Parametric cumulative distribution functions are therefore used as link functions because they provide substantial benefits in bias reduction (see, for example, Baldi et al., 2009), and they directly link the hazard rate (also called the failure rate) to an additive combination of multiple covariates (see, for example,Tutz and Schmid, 2016).

The commonly used cumulative distribution functions are the logit, probit and the clog-log (see, for example, Manda and Meyer (2005), Muggeo and Ferrara (2007), Nicolettia and Rondinelli (2010), Nam et al. (2008), Hess et al. (2014), Prentice

(1976), Laaksonen (2005), Buckley and Westerland (2004)).

The misspecification of the link function in discrete survival models should be avoided, because it results in estimated effects of covariates being biased. The estimated hazard rates may also become biased (Hess et al., (2014); Czado and Santner, 1992); Tutz and Petry, 2010). Misspecification of the link function may also affect the variable selection procedures (Tutz and Petry 2010).

However this can be avoided if the link function is fixed into a family of link functions or if it is estimated nonparametrically. Using the nonparametric approach may however bring some difficulties, because when the number of predictors is large the nonparametric methods fail since the full model can not be estimated (Tutz and Petry, 2010). In this study we will use the flexible families of link functions to model age at first marriage for women in Swaziland. The 2007 Swaziland Demographic and Health survey (SDHS) data will be used (Fisher and Way 1988).

Families of link functions are not widely used and they have never been used in studies on age at first marriage to the best of our knowledge. No study has been done to compare the families of link functions. Recent work has been done by Hess et al. (2014), where they proposed flexible specification of the link function. The authors only used one family of link functions. The current study will use the same family used by Hess et al. (2014) and compare it with other families of link functions.

A common feature of Demographic and Health survey (DHS) data which will be

used in this study is where by individuals are clustered in families and communities. Women from the same family tend to be more alike in their mental characteristics (for example, they may believe in the same religion) than individuals chosen at random from the population at large. They therefore may get married at almost the same age influenced by the religion that they believe in. In such cases, we can not assume independence between the responses of the individuals. As a result we can use multilevel models which are models that take into account the existence of data hierarchies because they allow for residual components at each level of hierarchy, therefore acknowledging the possible existence of unobserved variables which can influence the response at each level of the hierarchy the residual components account for variation due to unobserved source of variation applicable to that level of the hierarchy (Raudenbush and Bryk (2002), Gromping et al. (2015)).

The ignorance of unobserved heterogeneity can lead to inconsistent parameter estimates, wrong standard errors, and misleading estimates of duration dependency (Nicolettia and Rondinelli (2010), Groll and Tutz (2016)). In this study, discrete multilevel survival models (also known as random effects models) which allow for unobserved heterogeneity at the household level are required to ensure that the existence of groups is acknowledged in this study.

1.3 Aim and Objectives

1.3.1 Aim

The main aim of this study is to explore the use of flexible link functions in discrete survival models with unobserved heterogeneity, through a simulation study and an

application to real-life data.

1.3.2 Objectives

The objectives of this study are:

(1.) to perform a simulation study in order to evaluate the effectiveness of families of link functions and compare their performance in predicting the hazard probabilities,

(2.) to build a discrete survival model for age at first marriage among women in Swaziland using a flexible link function.

The rest of the dissertation is arranged in the following manner. Chapter 2 provides an overview of the previous research that has been conducted survival modelling on age at first marriage and families of link functions. Chapter 3 discusses the methodology, Chapter 4 provides results of the simulation study, chapter 5 provides analysis of real data and chapter 6 provide the conclusion and discussion.

Chapter 2

Literature review

2.1 Introduction

This chapter provides a literature review of some studies on age at first marriage and families of link functions.

2.2 Studies on age at first marriage

In the unpublished work, Ermans (2010) employed the multilevel survival model when modeling age at first marriage in Nigeria. Ermans used the cloglog link function. The author fitted a Cox proportional hazard model for continuous outcomes, which is not the focus of the present study since the current study will use discrete survival models. As indicated previously, discrete survival models have the ability to accommodate time varying covariates and they do not require proportional hazards (PH) assumptions like continuous time models.

Ikamari (2005), Aryal (2007), Agaba et al. (2010) and Kumchulesi et al. (2011) also used the Cox proportional hazard model when modeling age at first marriage among women. The multilevel structure of the data was ignored in their studies.

A study was conducted by Manda and Meyer (2005) on age at first marriage in Malawi. Data were obtained from the 2000 Malawi Demographic and Health Survey (DHS). The authors used a multilevel discrete survival model to assess the determinants of age at first marriage among women in Malawi. The baseline hazard was modeled using nonparametric techniques and the computation of the model parameters was done using the Markov Chain Monte Carlo algorithm. Manda and Meyer (2005) also used the complimentary log-log (cloglog) link function in their study, but they did not compare the cloglog link with other link functions. The current study will attempt to address this issue by comparing diferent link functions (probit, cloglog and logit) and apply a family of link functions with the aim of reducing the chances of misspecification of the link function.

In all studies cited above, the flexible families of link functions were not used to model age at first marriage. Only few studies used multilevel models. Most studies employed the Cox proportional hazard models.

2.3 Studies on families of link functions

This section presents a summary of some studies that have been done on link functions.

Hess et al. (2014) considered the flexible link functions to avoid the problem of misspecifying the link function. They considered only one family of link functions which is the generalized logistic distribution family, using log-Burr distribution as

the response function. The current study will also use the same family used by Hess et al. (2014), but it will compare it with other families. Through a simulation study Hess et al. (2014) concluded that the choice of the link function has important implications for the estimated covariate effects and the predicted hazards.

Muggeo and Ferrara (2008) fitted the generalized linear model (GLM) with unspecified link function. They considered data concerning $n = 683$ workers of a single manufacturing facility. The authors estimated a generalized linear model (GLM) with the logit link function and the unspecified link function. The unspecified link function was modelled using a penalized B-spline approach (Dierck, 1997). The logit, probit and cloglog links did not perform well.

Mallick and Gelfand (1994) permitted the data to estimate the link function by including it as unspecified in the model. The unspecified link function was modelled as a mixture of Beta distributions, using a Bayesian approach for model fitting. The authors recommended their approach for other statistical studies that involve modeling monotone functions and integrated hazard in survival analysis.

Jiang et al. (2013) proposed a new family of flexible link functions (which is the symmetric power link family) for modeling binomial data. The authors introduced a power parameter into the cumulative distribution function (cdf) corresponding to a symmetric link function and its mirror reflection. They achieved a great flexibility in skewness. The authors argued that the proposed family of link functions has an advantage of accommodating flexible skewness in both positive as well as negative

directions, while retaining the baseline standard link as a special case. They also performed a simulation study where the proposed link function was seen to be more flexible and it performed better than the standard link functions against the misspecificatiion of the link.

Nicolettia and Rondinelli (2010) carried out a study on the misspesification of discrete survival models with unobserved heterogeneity. The authors considered three sequential models with random effects which are: the sequential logit, sequential probit and sequential complementary loglog models. They conducted a Monte Carlo simulation to evaluate the consequences of ignoring the unobserved heterogeneity in a discrete survival model. In their study they found out that omitting the unobserved heterogeneity and misspecifying the link function can both cause bias in the duration dependence estimation. The present study will use discrete multilevel survival models with unobserved heterogeneity.

Some researchers who have done much work of introducing the flexibility into the link functions include Cordeiro and de Castro (2011) who proposed a new family of generalized distributions in order to extend the normal, gamma, Wiebull and Gumbel inverse distributions among other distributions. Jones (2004) generated a family of distribution F, by employing two parameters, where those parameters were used to introduce skewness and to vary tail weight. Czado and Raftery (2006) proposed the use of approximate Bayes factors to assess the improvement in fit over a GLM with canonical link using a parametric link family. This study will use the flexible families of link functions for discrete survival models with unobserved heterogeneity.

Based on the previous studies cited above, there is no study that has been done which compare the families of link functions in order to see their performance.

2.4 Conclusion from literature

To the best of our knowledge, there is no study which has been done on age at first marriage using flexible families of link functions and only few studies used multilevel survival models. Ignoring multilevel structure may result in the coefficients of the covariates being biased (see, e.g., Baker and Melino, 2000). In all studies cited in Sections 2.2 and 2.3 no effort was made to assess the adequacy of the link functions and in studies which used the flexible families of link functions, they were not compared with other families to assess their performance. The present study will employ a discrete multilevel survival model to model age at first marriage among women in Swaziland using flexible link functions.

Chapter 3

Methodology

3.1 Introduction

This chapter will introduce some of the methods which will be used in this study.

3.2 Discrete hazard and survival function

Let k intervals be given by $[a_0, a_1), [a_1, a_2), ..., [a_{q-1}, a_q), [a_q, \infty)$, where $q = k - 1$. Discrete time $T\epsilon\{1, ..., k\}$ means that $T = t$ is observed if the event happens within the interval $[a_{t-1}, a_t)$. Therefore the stochastic behaviour of T given $\mathbf{X}$ can be described by a probability density function of the form

$$P(T = t|\mathbf{X}), t = 1, ..., k, \tag{3.2.1}$$

or as cumulative density function,

$$F(t|\mathbf{x}) = P(T \leq t|\mathbf{X}). \tag{3.2.2}$$

We now define the discrete survival function. The discrete survival function is the

probability that the event happens later than time t, and it is given by

$$S(t|X) = P(T > t|\mathbf{X}) = \prod_{i=1}^{t}(1 - \lambda(i|\mathbf{X})). \tag{3.2.3}$$

A model for time to event data can also be specified by choosing a parametrization of the hazard function. The hazard function at time $t, \lambda(t|X)$ is given by

$$\lambda(t|\mathbf{x}) = P(T = t|T \geq t, \mathbf{x}). \tag{3.2.4}$$

Model 3.2.4 models a binary response that differentiate between the event occurring in category t or not (given that $T \geq t$).

An important class of models for binary response $Z\epsilon\{0, 1\}$ with given covariate x is given by

$$\lambda(Z = 1|\mathbf{x}) = F(\gamma_0 + \mathbf{x}_{it}^T\mathbf{\Gamma}), \tag{3.2.5}$$

where $F(\cdot)$ is a fixed link function which relates the response probability and the linear predictor $(\gamma_{0t} + \mathbf{x}_{it}^T\mathbf{\Gamma}), \gamma_{0t}$ is a base line hazard, $\mathbf{x}_{it}^T = (\mathbf{x}_{1,it}, ..., \mathbf{x}_{p,it})^T$ is a p-dimensional vector of covariates for observation $i\epsilon\{1, ..., N\}$. The linear predictor $(\gamma_{0t} + \mathbf{x}_{it}^T\mathbf{\Gamma})$ contains distribution functions since the probability can only fall in the interval $[0, 1]$. The binary model for decision between t and categories $(t + 1, ..., k)$ given that $T \geq t$ is used to obtain discrete hazard model give by

$$\lambda(t|\mathbf{x}_{it}) = F(\gamma_{0t} + \mathbf{x}_{it}^T\mathbf{\Gamma}), \tag{3.2.6}$$

The difference between equation 3.2.5 and 3.2.6 is the fact that the base line hazard inn equation 3.2.6 (γ_0) now depends on time (Tutz and Schmid, 2016).

3.3 Popularly used link functions

There are three popularly used link functions which are: the logit, probit and cloglog link functions.

The logit model is the popularly used binary regression model, which uses the logistic distribution function which is characterized by the CDF

$$F(n) = \frac{\exp(n)}{1 + \exp(n)}. \tag{3.3.1}$$

The corresponding logistic discrete hazard model which models the failure time t is given by

$$\lambda(t|\mathbf{x}) = \frac{\exp(\gamma_{0t} + \mathbf{x}_{it}^T \mathbf{\Gamma})}{1 + \exp(\gamma_{0t} + \mathbf{x}_{it}^T \mathbf{\Gamma})}. \tag{3.3.2}$$

Equation (3.3.2) can alternatively be written as

$$\log\left(\frac{\lambda(t|\mathbf{x})}{1 - \lambda(t|\mathbf{x})}\right) = \gamma_{0t} + \mathbf{x}_{it}^T \mathbf{\Gamma}, \tag{3.3.3}$$

or

$$\log\left(\frac{P(T = t|\mathbf{x})}{P(T > t|\mathbf{x})}\right) = \gamma_{0t} + \mathbf{x}_{it}^T \mathbf{\Gamma}, \tag{3.3.4}$$

where $\dfrac{P(T = t|\mathbf{x})}{P(T > t|\mathbf{x})}$ is the continuation ratio and it compares the probability of an event at t to probability of an event later than t, given that $T \geq t$ (Tutz and Schmid, 2016).

The probit link function uses an inverse normal link function to relate the binary response to the set of predictors through the equation

$$\lambda(t|\mathbf{x}) = \phi^{-1}(\gamma_{0t} + \mathbf{x}_{it}^T\mathbf{\Gamma}), \tag{3.3.5}$$

where ϕ is the inverse of the normal CDF. The CDF corresponding to the clog-log link function is given by

$$F(n) = 1 - \exp(-\exp(n)), \tag{3.3.6}$$

and it yields the model given by

$$\lambda(t|\mathbf{x}) = 1 - \exp(-\exp(\gamma_{0t} + \mathbf{x}_{it}^T\mathbf{\Gamma})). \tag{3.3.7}$$

All the above mentioned popularly used link functions (logit, probit and cloglog) will be used in the current study and they will be compared with families of link functions.

3.4 Flexible families of link functions

Although any link function may yield satisfactory results in many applications, it is still possible to improve the model fit by making use of more flexible choices of link functions. The corresponding flexible choices can either be obtained if the link function is fixed into a family of link functions or if it is estimated in a nonparametric way (Tutz and Schmid, 2016). There are different families of link functions. Firstly we have the generalized logistic distribution family: This family was used by Hess et al. (2016). The corresponding CDF is given by

$$F_\xi(u) = 1 - (1 + \xi \exp(u))^{\frac{-1}{\xi}}, \tag{3.4.1}$$

where ξ is the shape (family) parameter and it is non-negative ($i.e\ \xi > 0$). When $\xi = 1$ one obtains the logistic distribution. For $\lim_{\xi \to 0}$, one obtains the complementary log-log (cloglog) model. Thus the generalized logistic distribution family consists of two widely used discrete survival models which are the logistic and the grouped proportional hazard models. The density function for this family is given by

$$f_{\xi}(u) = 1 - (1 + \xi \exp(u))^{\frac{-1}{\xi} - 1} \exp(u). \tag{3.4.2}$$

Secondly we have the family which was proposed by Aranda -Ordaz (1983). The CDF for the Aranda family is given by

$$F_{\alpha}(u) = \begin{cases} 1 - \exp(-(1 + \alpha u)^{\frac{1}{\alpha}}), & \text{for} \quad u\ \epsilon\ [-\frac{1}{\alpha}, \infty) \\ \\ 0, & \text{otherwise}, \end{cases}$$

where $F_{\alpha}(u)$ depends on the parameter α. When α approaches zero one obtains the cloglog or grouped proportional hazard model and when $\alpha = 1$ one obtains the model given by

$$-\log(1 - \lambda(t|\mathbf{x}_{it})) = \gamma_{0t} + \mathbf{x}_{it}^{T}\mathbf{\Gamma}, \tag{3.4.3}$$

where the constant 1 is being absorbed by the baseline hazard (γ_{0t}) parameters. When $\alpha = 1$ we obtain a discretized version of the additive continuous time model which is given by

$$\lambda(t|\mathbf{x}_{it}) = \lambda_{0t} + \mathbf{x}_{it}^{T}\mathbf{\Gamma}. \tag{3.4.4}$$

Thus, this family contains special cases which are the grouped proportional hazards model and a discretized version of an additive model. Its disadvantage is the restriction of the predictor's range; the linear predictor has to fulfill the condition $\gamma_{0t} + \mathbf{x}_{it}^T \mathbf{\Gamma} > -\frac{1}{\alpha}$. The third family is a family of link functions for binary responses having two parameters. This family was proposed by Pregibon (1980) and the corresponding inverse CDF is written as

$$G_{a,b}(\lambda) = \frac{\lambda^{a-b} - 1}{a - b} - \frac{(-a)^{a-b} - 1}{a + b}, \tag{3.4.5}$$

which consists of the symmetric logit link as a, b approach zero, but it is not symmetric for $a, b > 0$. The family can be implemented using the r-package gld.

The fourth family is the Gosset family of links which enable one to account for symmetric distributed heavy tails in the latent variable model for binary response. The family is given by the inverse of the cumulative distribution function of the $t-$distribution written as

$$F(t) = \int_{-\infty}^{t} f(u)du = 1 - \frac{1}{2} I_{x(t)}(\frac{\nu}{2}, \frac{1}{2}), \tag{3.4.6}$$

where ν is the degrees of freedom parameter and $x(t) = \frac{\nu}{t^2 + \nu}$. When $\nu > 6$, the Gosset link cannot be easily distinguished from the Probit link. When $\nu = 7$ or 8 one can approximate the logit model well using the Gosset model (Koenker and Yoon 2009). The likelihood becomes difficult to evaluate when ν is closer to zero or less than 0.2, due to the fact that the tails of inverse $t-$distribution are dramatically heavy.

The above mentioned families of link functions are the most recent families and they

are not widely used. The focus of this study will be on the generalized logistic distribution family used by Hess et al. (2016), but comparing it with other families of link functions.

In the current study we will use the generalized logistic distribution family, Pregibon family and the Gosset family.

3.5 Model building

We now assume that we have k individuals clustered in family j and community i with a vector of covariates $\mathbf{x}_{ijk(t)}$ at time $t(i = 1, ..., N; j = 1, ..., j_i; k = 1, ..., K_{ij})$ Manda and Meyer (2005). The hazard function for a particular individual is written as

$$F(t|\mathbf{x}_{ijk(t)}, b_{ij}, b_i) = P_r(T = t|T \geq t; \mathbf{x}_{ijk(t)}, b_{ij}, b_i), \tag{3.5.1}$$

and the discrete survival function is now given by

$$S(t|\mathbf{x}_{ijk(t)}, b_{ij}, b_i) = P_r(T > t|\mathbf{x}_{ijk(t)}, b_{ij}, b_i) = \prod_{i=1}^{t}(1 - h(t|\mathbf{x}_{ijk(t)}, b_{ij}, b_i)). \tag{3.5.2}$$

The multilevel model with two level of hierarchy is then written as follows

$$\lambda(t|\mathbf{x}_{ijk}, b_{ij}, b_i) = F(\gamma_{0t} + \mathbf{x}^T_{ijk(t)}\mathbf{\Gamma} + b_{ij} + b_i), \tag{3.5.3}$$

where b_i represents the unobserved covariates at a community level and and b_{ij} represents the unobserved covariates at family level. The examples of unobserved

covariates at community level are the amount of bride price and approved age at marriage, and the examples of unobserved covariates at family level are poverty and social status of the family.

The multilevel model with one level of hierarchy has the form

$$\lambda(t|\mathbf{x}_{it}, b_i) = F(b_i + \gamma_{0t} + \mathbf{x}_{it}^T\mathbf{\Gamma}), \tag{3.5.4}$$

where b_i is a random effect that is assumed to follow a normal distribution.

The discrete duration model without unobserved heterogeneity is already presented in chapter one (equation 1.1.13).

In the current study we will consider multilevel models with one level of hierarchy and models without unobserved heterogeneity.

3.6 Parameter estimation

3.6.1 Introduction

In this section we illustrate how parameters from the models in this study are estimated. The objective is to find the estimates of the covariates coefficients and the baseline hazard.

3.6.2 Models without unobserved heterogeneity

Now we let the basic discrete model be given in the form $\lambda(t|\mathbf{x}_{it}) = F(\gamma_{0t} + \mathbf{x}_{it}^T\mathbf{\Gamma})$

$$\lambda(t|\mathbf{x}_{it}) = F(\mathbf{x}_{it}^T\beta), \tag{3.6.1}$$

where $\mathbf{x}_{it}^T = 0, ..., 0, 1, 0, ..., 0, \mathbf{x}_i^T$ is a vector of explanatory variables with 1 being the t^{th} digit and all parameters are collected in $\beta^T = (\gamma_{01}, ..., \gamma_{01q}, \gamma^T)$. We then assume that observations are subjected to censoring, meaning that only a portion of the observed times can be considered as exact survival times. One may only know that the survival time exceeds a certain time point for the rest of the observation.

Now we let T_i be the response for the i^{th} subject, C_i be the censoring time for the i^{th} subject and δ_i be the event indicator.

The observed time is written as follows

$$t_i = \min(T_i, C_i), \tag{3.6.2}$$

which is the minimum of the response T_i and the censoring time C_i.

Now the event indicator is given by

$$\delta_\alpha(u) = \begin{cases} 1, & \text{if the event was observed } (T_i \leq C_i), \\ 0, & \text{if the respond was cencored } (T_i > C_i), \end{cases}$$

,

We now assume that censoring happens at the end of the interval and we let total data with potential censoring be given by $(t_i, \delta_i, \mathbf{x}_i),\ i = 1, ...,$n, where n is the sample

size.

The probability of observing the exact survival time $(t_i), \delta_i = 1$ is given by

$$P(T_i = t_i, \delta_i = 1) = P(T_i = t_i)P(C_i \geq t_i), \tag{3.6.3}$$

since time and censoring are independent by assumption. The probability of observing censoring time is given by

$$P(c_i = t_i, \delta_i = 0) = P(T_i > t_i)P(C_i = t_i). \tag{3.6.4}$$

A failure in interval $[a_{t_i-1}, a_{t_i}]$ implies that $C_i \geq t_i$, and censoring interval $[a_{t_i-1}, a_i)(i.e. C_i = t_i)$ implies survival beyond a_{t_i}. The contribution of the ith observation to the likelihood function is then given by

$$L_i = P(T_i = t_i)_i^{\delta} P(T_i > t_i)^{1-\delta_i} P(C_i \geq t_i)^{\delta_i} P(C_i = t_i)^{1-\delta_i}. \tag{3.6.5}$$

We then assume that the censoring contributions that involve C_i do not depend on the parameters that determine the survival time. Now we separate the factor $C_i := P(C_i \geq t_i)^{\delta_i} P(C_i = t_i)^{1-\delta_i}$ to obtain a simpler form which is given by

$$L_i = c_i P(T_i = t_i)^{\delta_i} P(T_i > t_i)^{1-\delta_i}. \tag{3.6.6}$$

By using the discrete hazard function and including the covariates, one obtains a likelihood given by

$$L_i = c_i \lambda(t_i, \mathbf{x}_i)^{\delta_i} (1 - \lambda(t_i, \mathbf{x}_i))^{1-\delta_i} \prod_{j=1}^{t_{i-1}} (1 - \lambda(t_i, \mathbf{x}_i)). \tag{3.6.7}$$

The full likelihood for the discrete survival model is given by

$$l(\alpha) = \Sigma_{i=1}^{n}\Sigma_{s=1}^{t_i} y_{is} \log \lambda(s|\mathbf{x_i}) + (1 - y_{is}) \log(1 - \lambda(s|\mathbf{x_i})), \tag{3.6.8}$$

where $s = 1, ..., t_i$. Equation 3.6.8 represents the log likelihood of a binary regression model. This means that the usual methods for estimating parameters in binary regression models are applicable here. The first step is to differentiate the log-likelihood equation (3.6.8) with respect to each of the unknown parameters and equating each of the derivatives to zero.

In most cases, it is not possible to obtain closed form expressions for the parameter estimates. Numerical methods such as the $Newton\_Raphson$ method are then used.

Before we can derive the matrices, we first consider a discrete hazard model with linear predictor $\varsigma_{ir} = \gamma_{0r} + \mathbf{x}_i^T\gamma$. The linear predictor is given by $\mathbf{x}_{ir}^T\beta$ with parameter vector $\beta^T = (\gamma_{0,1}, ..., \gamma_{0q}, \gamma^T)$. One needs to adapt the covariate vector to the observations under consideration for the linear predictors of the binary model. one also have to distinguish between censored and non censored observations for the generation of designed matrix. If $T = t_i$ and $\delta_i = 1$ therefore the t_i binary observations and design variables are given by the argument data matrix given by

$$\begin{pmatrix} Binary\ observations & D_1 & D_2 & D_3 & . & . & . & D_{t-i} & Covariates \\ 0 & 1 & 0 & 0 & . & . & . & 0 & x_i^T \\ 0 & 0 & 1 & 0 & . & . & . & 0 & x_i^T \\ 0 & 0 & 0 & 1 & . & . & . & 0 & x_i^T \\ . & . & . & . & . & . & . & . & x_i^T \\ . & . & . & . & . & . & . & . & x_i^T \\ . & . & . & . & . & . & . & . & x_i^T \\ 1 & 0 & 0 & 0 & . & . & . & 1 & x_i^T \end{pmatrix},$$

where D represent the Dummy variables. Similarly, If $T = t_i$ and $\delta_i = 0$ the t_i binary observations and design variables are given by the argument data matrix given by

$$\begin{pmatrix} Binary\ observations & D_1 & D_2 & D_3 & . & . & . & D_{t-i} & Covariates \\ 0 & 1 & 0 & 0 & . & . & . & 0 & x_i^T \\ 0 & 0 & 1 & 0 & . & . & . & 0 & x_i^T \\ 0 & 0 & 0 & 1 & . & . & . & 0 & x_i^T \\ . & . & . & . & . & . & . & . & x_i^T \\ . & . & . & . & . & . & . & . & x_i^T \\ . & . & . & . & . & . & . & . & x_i^T \\ 0 & 0 & 0 & 0 & . & . & . & 1 & x_i^T \end{pmatrix}$$

An alternative way can be by means of constructing the observations with running time t, which has to be considered as a factor. Then if $T = t_i$ and $\delta_i = 1$ one uses the coding given by

$$\begin{pmatrix} \textit{Binary observations} & \textit{time} & \textit{Covariates} \\ 0 & 1 & x_i^T \\ 0 & 2 & x_i^T \\ 0 & 3 & x_i^T \\ . & & \\ . & & \\ . & & \\ 1 & t_i & x_i^T \end{pmatrix}$$

In a similar way, if $T = t_i$ and $\delta_i = 0$ one uses the coding given by

$$\begin{pmatrix} \textit{Binary observations} & \textit{time} & \textit{Covariates} \\ 0 & 1 & x_i^T \\ 0 & 2 & x_i^T \\ 0 & 3 & x_i^T \\ . & & \\ . & & \\ . & & \\ 0 & t_i & x_i^T \end{pmatrix}$$

3.6.3 Models with unobserved heterogeneity

We consider general frailty model which has the form

$$\lambda(t|\mathbf{x}_{it}, \mathbf{b}_i) = F(\mathbf{w}_{it}^T\gamma + \mathbf{z}_{it}^T\mathbf{b}_i), \tag{3.6.9}$$

where $\mathbf{w}_{it}$ and $\mathbf{z}_{it}$ are the explanatory variables and $\mathbf{b}_i$ is the random effect. Different sets of predictor are collected in vectors $\mathbf{w}_{it}$ and $\mathbf{z}_{it}$, this is referred to the fixed

and random effects γ and $\mathbf{b}_i$ respectively. A good example can be a simple frailty model with random intercepts which uses $\mathbf{w}_{it}^T = (0, ..., 1, ..., 0, \mathbf{x}_i^T)$, with parameters $\beta^T = (\gamma_{01}, ..., \gamma_{01q}, \gamma^T)$ and $\mathbf{z}_{it} = 1$.

An assumption which is commonly made for random effects is a normal distribution, $\mathbf{b}_i \sim N(\mu, \sigma^2)$, where β and δ can be estimated simultaneously by applying numerical integration, which can be done by maximizing the marginal log-likelihood given by

$$l(\beta, \delta) = \sum_{i=1}^{n} \log(\int \prod_{s=1}^{ti} \lambda(s|x_i)^{y_{is}} (1 - \lambda(s|x_i)^{1-y_{is}}) db_i. \tag{3.6.10}$$

There are two popular approaches for numerical integration which are: the Gauss-Hermite quadrature and the Monte Carlo approximations. The Gauss Hermite procedure approximates the integral in equation (3.6.8) by using a pre-specified number of quadrature points and the Monte Carlo approximation approximates the integral in equation (3.6.8) by transforming the integration problem into a procedure of sampling values from a tractable probability distribution and calculating the average of those samples.

Another way of estimating random effects models is by using penalized quasilikelihood estimation, which uses the Laplace approximation (Breslow and Clayton 1993) and it is given by

$$l(\delta) = \sum_{i=1}^{n} \log f(y_i|b_i, \beta) - \frac{1}{2} \sum_{i=1}^{n} b_i^T \delta^{-1} b_{ij}, \tag{3.6.11}$$

where $(f(y_i|b_i, \beta)) = \prod_{s=1}^{t_i} \lambda(s|x_i)^{y_{is}} (1 - \lambda(s|x_i))^{1-y_{is}}$. The disadvantage with this method is that it makes the variance of the random effect to be underestimated.

However, this can be avoided by using the modifications proposed by Breslow and Lin (1999). With the penalized quasilikelihood estimation, the parameters β and $b_1; ...; b_n$ and the mixture parameters are estimated separately.

3.7 Model evaluation

Here we consider the diagnostic and model selection tools for the discrete survival model that we will apply to the real life data.

3.7.1 Relevance of predictors Tests

In survival modelling the main interest is in the impact of explanatory variables. One strategy to evaluate the relevance of variable j within the model is to test for the null hypothesis as follows

$$H_0 : \gamma_j = 0, \tag{3.7.1}$$

against

$$H_1 : \gamma_j \neq 0, \tag{3.7.2}$$

which only works if a variable is represented by only one parameter. For example if one has a factorial explanatory variable or if a quadratic term of continuous variables are included in the model, one has to test simultaneously if all corresponding parameters are zero. A linear hypotheses that covers these cases are given by

$$H_0 : C\beta = \xi, \tag{3.7.3}$$

against

$$H_1 : C\beta \neq \xi, \tag{3.7.4}$$

where C is a fixed matrix of full rank $s \leq p$ and ξ is a fixed vector. β is a vector that collects all the parameters of the model $(i.e, \beta^T = \gamma_{01}, ..., \gamma_{0q}, \Gamma^T)$.

Tests can be derived from the binomial representation of the log-likelihood given by

$$l(\alpha) = \sum_{i=1}^{n}\sum_{s=1}^{n} Y_{is} \log \lambda(S|x_i) + (1 - Y_{is}) \log(1 - \lambda(S|x_i)). \tag{3.7.5}$$

This approach allows for using tools from binary regression. For example, a common test statistic for linear hypothesis is the likelihood ratio test, which is based on the comparison between two models, the models without constrains and the models fitted under linear constrains. We let $\hat{\beta}$ denote the ML estimate for the full model (model 3.2.5), and $\bar{\beta}$ denote the estimate under constraint $C\beta = \xi$. Then the likelihood ratio statistic is given by

$$LR = -2\{l(\bar{\beta}) - l(\hat{\beta})\}, \tag{3.7.6}$$

which quantifies the change of log likelihood l given in equation (3.8.5) when evaluated at $\hat{\beta}$ and $\bar{\beta}$. The likelihood ratio statistic follows asymptotically a chi-square distribution under regularity conditions, with degrees of freedom given by

$$S = rk(C), \tag{3.7.7}$$

where $rk(C)$ denotes the rank of the matrix C.

Other test statistics that can be derived as approximations of LR are Wald and the Score statistic. The Wald statistic is given by

$$w = (C\hat{\beta} - \xi)^T [CF^{-1}(\hat{\beta})C^T]^{-1}(C\hat{\beta} - \xi), \tag{3.7.8}$$

where $F(\beta) = E(-\partial^2 l(\beta)/\partial\beta\partial\beta^T)$ denotes the Fisher information matrix. The advantage of using the Wald test is that only the full model has to be fitted to obtain $\hat{\beta}$.

The Score statistic is given by

$$u = s^T(\bar{\beta})F^{-1}(\bar{\beta})s(\bar{\beta}), \tag{3.7.9}$$

where $s(\beta) = \partial l(\beta)/\partial\beta = (\partial l(\beta)/\partial\beta_1, ..., \partial l(\beta)/\partial\beta_p)^T$ is the score function evaluated at the fit of the constrained model.

The likelihood ratio, Wald and Score statistic have the same distribution where $LR, w, u \sim \chi^2(rank\ C)$, see; for example, Fahrmeir and Tutz (2001), Tutz and Schmid (2016).

3.7.2 Residuals and goodness-of-fit

When analysing residuals and goodness-of-fit, one should always know that the original data is consists of n independent observations (t_i, δ_i, x_i). In a model without censoring, the discrete time $T_i \epsilon (1, ..., q\}$ follows a multinomial distribution and if data is censored, analysis of the fitted model becomes more complicated. In such cases one models the observed time periods which is given by

$$t_i = \min(T_i, C_i), \tag{3.7.10}$$

which is the minimum of the response T_i and the censoring time C_i. The Deviance for grouped observations is written as:

$$D = 2\sum_{i=1}^{N} n_i \left(\sum_{r=1}^{q} P_{ir}^{(1)} \log(\frac{P_{ir}^{(1)}}{\hat{\pi^{(1)}}}) + \sum_{r=1}^{q} P_{ir}^{(0)} \log(\frac{P_{ir}^{(0)}}{\hat{\pi}^{(0)}}) \right), \tag{3.7.11}$$

where

$$\log\left(\frac{P_{ir}^{(1)}}{\hat{\pi}^{(1)}}\right) = \log\left(\frac{P(T_i = r)}{\hat{P}(T_i = r)}\right) + \log\left(\frac{P(C_i \geq r)}{\hat{P}(C_i \geq r)}\right), \tag{3.7.12}$$

and

$$\log\left(\frac{P_{ir}^{(0)}}{\hat{\pi^{(0)}}}\right) = \log\left(\frac{P(T_i > r)}{\hat{P}(T_i > r)}\right) + \log\left(\frac{P(C_i = r)}{\hat{P}(C_i = r)}\right). \tag{3.7.13}$$

The deviance asymptotically follow chi-square distribution with $N(k-1)-p$ degrees of freedom, were p is the number of estimated parameters. If one is using a significance level α the model will be considered inappropriate if both or one of the test statistics are larger than $1-\alpha$ (see; Tutz and Schmid 2016 for more details).

3.7.3 Martingale Residuals

A Martingale residuals is an alternative type of residual that takes censoring into account and is particularly suited for assessing the functional forms of predictor effects. It is given by

$$m_i = \delta_i - \sum_{s=1}^{ti} \hat{\lambda}_{is}, \qquad i = 1, ..., n, \tag{3.7.14}$$

where $\hat{\lambda}_{is} = \hat{\lambda}(s|\mathbf{X}_i)$ and $\wedge(t_i) = \sum_{s=1}^{ti} \hat{\lambda}_{is}$ measures the cumulative risk of observation i up to time t_i.

The idea of the martingale residual is to compare for each individual the observed number of events up to t_i (measured by δ_i) with the expected number of events up to t_i (measured by $\wedge(t_i)$).

If one wishes to use the binary variables representation with $(Y_{i1}, ..., Y_{it_i}) = (0, ..., 0, \delta_i)$, then the residual can be defined as

$$m_i = \sum_{s=1}^{ti} (Y_{is} - \hat{\lambda}_{is}), \qquad i = 1, ..., n. \tag{3.7.15}$$

Thus the martingale residuals use the differences between the transition indicators and the estimated probabilities of failure, in contrast to the deviance residuals which use the log-transformed proportion of observation (see; Tutz and Schmid 2016 for more details and examples)

3.7.4 Model selection tools

Once an appropriate transformations have been conducted and the model fits the data set, one can run each model and compute the AIC and BIC. The same data set must be used for each model *i.e* the same observations must be used for each analysis. Missing values for a certain variable in the data set can cause variation in number of observation. The same response function must be used in all models.

AIC is an estimate of a constant plus the relative distance between the unknown true likelihood function of the data and the fitted likelihood function of the model and is given by

$$AIC = 2k - 2\ln(\hat{L}), \tag{3.7.16}$$

where k is the number of estimated parameters in the model and $(\hat{L})$ is the maximum value of the likelihood function for the model. A model with the minimum AIC value is said to be the best model.

BIC (also known as SBC) is an estimate of a function of the posterior probability of a model being true, under a certain Bayesian setup and is given by

$$BIC = \ln(n)k - 2\ln(\hat{L}), \tag{3.7.17}$$

where $\hat{L}$ is the maximize value of the likelihood function of the model M and it is given by

$$\hat{L} = P(\mathbf{X}|\hat{\theta}), \tag{3.7.18}$$

where $\hat{\theta}$'s are the parameter values, $\mathbf{X}$ is the observed data set, n is the number of data points in x and k is the number of parameters estimated by the model. The model with a lower value of BIC is considered to be more likely the true model.

3.8 Predictions from the model

The prediction accuracy of survival models can be measured using three different approaches which are the predictive Deviance and R^2 coefficients, the prediction error curves and the discrimination measures. These three approaches which are designed to measure the prediction accuracy of survival models will be considered in this section.

3.8.1 The predictive Deviance and R^2 coefficients

The predictive deviance is given by

$$D = -2\Sigma_{j=1}^{n^\tau}\{\sigma_j^\tau \log(\hat{P}(T^\tau = t_j^\tau)) + (1-\sigma_j^\tau)\log(\hat{P}(T_j^\tau) > t_j^\tau)\} \tag{3.8.1}$$

$$= -2\Sigma_{j=1}^{n^\tau}\Sigma_{s=1}^{t_j^\tau}\{Y_{js}^\tau \log(\hat{\lambda_{js}}) + (1-Y_{js}^\tau)\log(1-\lambda_{js})\}, \tag{3.8.2}$$

where $\lambda_{js} = \hat{P}(T_j^\tau) = S|T_j^\tau \geq S_{xj}^\tau$ and where T_j^τ are the unobserved survival time. The predictive deviance is an unbounded sure, thus in order to facilitate interpretation one needs to consider R^2 coefficients which is given by

$$R^2 = \frac{1-\exp(c\Sigma_{j=1}^{n^\tau}t_j^\tau)^{-1}(D-D_0)}{1-\exp(c\Sigma_{j=1}^{n^\tau}t_j^\tau)^{-1}D_0}, \tag{3.8.3}$$

where D_0 is the predictive deviance which can be obtained from a null model without covariate information.

3.8.2 Prediction Error Curves

The predictive error curves (PE curves) are time dependent measure of prediction error that is based on the squared distance between the predicted survival functions. The PE values for fixed values of t is given by

$$\hat{PE}(t) = \frac{1}{n^\tau}\Sigma_{j=1}^{n^\tau}\left[\frac{\sigma_j^\tau(1-\hat{S}_j(t))}{\hat{G}_j(t_j^\tau - 1)} + \frac{\hat{S}_j(t)}{\hat{G}_j(t)}\right](\hat{S}_j(t) - \hat{S}_j(t))^2 \tag{3.8.4}$$

$$= \frac{1}{n^\tau}\Sigma_{j=1}^{n^\tau}W_j^\tau(t)(\hat{S}(t) - \hat{S}_j(t))^2 \tag{3.8.5}$$

3.8.3 Discrimination Measures

Prediction accuracy can also be evaluated by computing the discrimination measures for the predictor $\varsigma := x^T\Gamma$. The main reason of computing the discrimination measures is to consider survival outcomes as time-dependent binary variables with levels event at t and event after t thus, prediction accuracy can be measured by using established concepts for the evaluation of binary classification rules.

This is done as follows: For each fixed time point t, one considers observations having an event at t as cases and observations having an event after t as controls. Using this definition, ς has a high prediction accuracy if it has a high discriminative power. Discriminative power can be measured using time-dependent sensitivities and specificities, which are defined as follows

$$sens(c,t) := P(\varsigma > c|T = t), \tag{3.8.6}$$

and

$$spec(c,t) := P(\varsigma > c|T > t), \tag{3.8.7}$$

where c is a threshold of the predictor ς. Therefore a large sensitivity represent a large hazard.

In the current study we will use the AIC to select the best model since there is no software package which implements the measures described in this section using the generalized logistic family.

3.9 Simulation study

The main purpose of the simulation study is to compare the accuracy of the link functions. We cannot compare the link functions using real life data because we don't know the true characteristics of the real data. The true response function is assumed to be the logit link function through out the study. The focus will be based on the effect of misspecifying the hazard functional form and assessing the performance of the flexible Pareto hazard models, the Gosset and the Pregibon hazard model. The estimation of the models will be done using 'gamlss.mx' (Stasinopoulos and Rigby (2012)), 'glmx' (Zeileis et al. (2013)) and 'gld' packages found in R software programme (Wuertz, 2010).

We generate survival data in long format with known covariate coefficient values and specified baseline hazard functions using logistic link and fit other link functions and families of link functions, and compare the accuracy with which the different link functions recover the known parameter values and hazard ratios.

The data that will be used in this study was generated from the logistic distribution using 'gamlss.mx' in R software program (Stasinopoulos and Rigby (2012)). The hazard rates are given by

$$\lambda(t|x) = \frac{\exp(\gamma_{0t} + x_{1,i}\gamma_1 + x_{2,i}\gamma_2)}{1 + \exp(\gamma_{0t} + x_{1,i}\gamma_1 + x_{2,i}\gamma_2)}, \tag{3.9.1}$$

where variable X_1 is generated from a normal distribution with mean of zero and standard deviation of 1.5 . Variable X_2 is generated from a demeaned gamma distribution with unit variance and $\gamma_1 = \gamma_2 = 0.5$. The values of parameters X_1 and X_2

were chosen arbitrarily. Each generated hazard was compared with a uniform random number. Whenever the uniform random number was exceeded by the computed hazard, the binary response y would take the value 1 otherwise it would take the value 0 (no event).

The base line hazard function γ_{0t} is specified as follows

$$\gamma_0 = -\ln(t) \tag{3.9.2}$$

The hazard rate for models with the unobserved heterogeneity is given by

$$\lambda(t|x_{it}, b_i) = \frac{\exp(b_i + \gamma_{0t} + x_{1,i}\gamma_1 + x_{2,i}\gamma_2)}{1 + \exp(b_i + \gamma_{0t} + x_{1,i}\gamma_1 + x_{2,i}\gamma_2)}, \tag{3.9.3}$$

where the individual effects b_i are drawn independently from a normal distribution with mean of zero and standard deviation of 0.1. Parameters γ_1 and γ_2 are fixed at 0.5 and 0 respectively. Again the values of parameters γ_1 and γ_2 were chosen arbitrarily. All observations which survived beyond period 12 were considered to be censored. Models based on each of the 5 link functions were fitted to the data in order to determine the accuracy with which each model recovers the parameter values used in generating the data.

The simulation exercise for Pareto, probit, logit, cloglog, Gosset and the Pregibon without unobserved heterogeneity will be repeated 100 times. This is consistent with the work done by Hess et al. (2016). The simulation exercise for the Pregibon with unobserved heterogeneity will be repeated 20 times only since it takes time to produce the results. The reported parameter and hazard estimates are therefore averages of

the total number of replicates of the simulation. The main reason of repeating the simulation exercise several times is to account for sampling variability of the estimates.

Chapter 4

Results of the simulation study

4.1 Introduction

Here a brief exemplification showing the importance of the link function on results estimation will be given. The results that will be presented in this chapter are found from a simulated data set consisting of 1500 and 3000 individuals. Two simulation experiments will be performed. In the first simulation experiment, the hazard at each time point will be generated using equation (3.9.1), i.e. the logit link was used to generate data. The Pareto hazard model with family parameter ξ, Pareto hazard model with $\xi = 1$, logit, probit, cloglog, Pregibon and the Gosset models are then considered. In the second simulation experiment, the hazard at time t is generated using equation (3.9.3). The difference between the two simulation experiments is that in the first simulation experiment the model does not have a frailty component while the model in the second has. In the second simulation experiment, we fit the Pareto hazard model with family parameter ξ, Pareto hazard model with $\xi = 1$, logit, probit, cloglog and the Pregibon model with frailty.

4.2 Results for models without Frailty

This section provides the results and interpretation of the models without unobserved heterogeneity.

The value of ξ for Pareto model with family parameter ξ is estimated at regular of fixed values ξ starting from 0.01 to 5 moving in steps of 0.1 allowing R software programme to select the position where the smallest value of AIC is from the model using 'gamlss.mx' package. When R is done selecting the position it was then allowed to select the value in that particular position and the ξ value was taken as the value of the family parameter ξ (Hess et al., 2014; Stasinopoulos and Rigby, 2012).

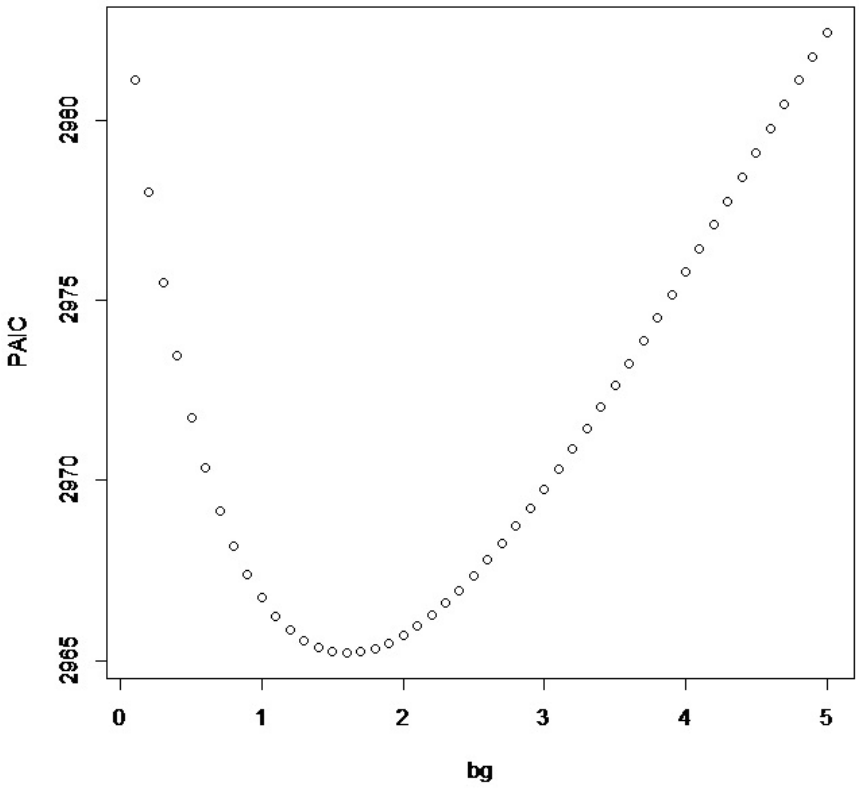

Figure 4.1: Akaike information criterion (AIC).

.

Figure 4.1 indicates the AIC graph, where bg represent the values of ξ and PAIC represents the AIC values. The AIC graph reaches the minimum value at 1.7. This suggests that the value of ξ for Pareto ($\xi = \xi$) is equal to 1.7 for models without frailty.

The values for γ_1 and γ_2 in Table 4.1 for the misspecified link functions were standardised using Amemiya conversion factors proposed by Amemiya (1981). The main aim of standardising the parameter estimates was to enable meaningful comparison of the estimated coefficients. The coefficients, estimates are weighted by factor $F'(\bar{u})$ using

the conversion mentioned above, where $F'(\cdot)$ is the first derivative of the response function used and $\bar{u}$ is the mean of index function $u = \gamma_{0t} + x^T_{ijk(t)}$ at estimated values of γ. The results are given in Table 4.1. Standard errors of the estimated coefficients are given in brackets.

Table 4.1: Comparing Covariate Effects Across Model Specifications.

	Logit	Cloglog	Probit	Pareto ($\xi = 1$)	Pareto($\xi = \xi$)	Gosset	Pregibon
γ_1	0.491 (0.031)	0.433 (0.024)	0.437 (0.017)	0.498 (0.031)	0.504 (0.067)	0.475 (0.019)	0.494(0.042)
γ_2	0.504(0.046)	0.404 (0.031)	0.447 (0.026)	0.501 (0.044)	0.504 (0.087)	0.483 (0.031)	0.506(0.046)
γ_1/γ_2	0.974	1.072	0.978	0.994	1.000	0.983	1.136
Hazard Ratio at $t = 1(x_1)$	1.225 (0.013)	1.279 (0.021)	0.959 (0.067)	1.244 (0.017)	1.251 (0.035)	0.954 (0.013)	0.946(0.023)
Hazard Ratio at $t = 10(x_1)$	1.500 (0.053)	1.430 (0.032)	1.374 (0.019)	1.553 (0.042)	1.569 (0.096)	1.551 (0.091)	1.475 (0.107)
N	1500	1500	1500	1500	1500	1500	1500
γ_1	0.497 (0.022)	0.433 (0.014)	0.441 (0.011)	0.498 (0.019)	0.504 (0.061)	0.446 (0.014)	0.501 (0.026)
γ_2	0.498 (0.031)	0.391 (0.025)	0.444 (0.021)	0.493 (0.037)	0.491 (0.076)	0.421 (0.020)	0.494 (0.029)
γ_1/γ_2	0.998	1.110	0.993	1.10	0.999	1.059	1.184
Hazard Ratio at $t = 1(x_1)$	1.234 (0.012)	1.279 (0.021)	0.962 (0.007)	1.244 (0.11)	1.242 (0.029)	0.946 (0.004)	0.939 (0.015)
Hazard Ratio at $t = 10(x_1)$	1.520 (0.019)	1.430 (0.032)	1.361 (0.012)	1.551 (0.031)	1.574(0.089)	1.596 (0.076)	1.472 (0.069)
N	3000	3000	3000	3000	3000	3000	3000

The pattern of results in Table 4.1 in terms of how the generalized logistic distribution family compares with the logit, cloglog and probit is comparable to those of Hess et al. (2016), who used Burr distribution to generate their data.

All p-values were found to be less than 0.005, which means that all variables used are significant at 5 % level of significance. The results indicate that the covariate effects are biased when the misspecified link function is used and accurately estimated when the correct link function (logit) is used. The results for Pareto model ($\xi = 1$) are almost the same as the ones obtained from the correct model (logit). This is due to the fact that with the Pareto models, for $\xi = 1$ one obtains the logistic distribution function. The results for the Pareto ($\xi = \xi$) are very close to those of the correct link function. The results also reveal that the covariates effect for the Pregibon link ($\gamma_1 = 0.494$ and $\gamma_2 = 0.506$) are close to those of logit link compared to those of the Gosset family of link functions.

The third row and the ninth row of Table 4.1 indicate the covariates effects ratios. The results show that the relative effects of covariates become biased when the link function is misspecified.

Row four, five, ten and eleven of Table 4.1 indicate the hazard ratios and they are estimated at the shortest ($t = 1$) and the long ($t = 10$) observed duration. The hazard ratios for all models are calculated for an increase in x_1 from zero to one, while x_2 is fixed at its mean of zero. The results reveal that the hazard ratios become over

estimated for the cloglog ($H = 1.279$) and Pareto model with $\xi = \xi$ ($H = 1.251$) compared to the correct model, and under estimated for the Pareto model with ($\xi = 1$) ($H = 1.149$). Moreover, the hazard ratio for the probit model becomes substantially small at shortest duration ($H = 0.959$) and at long duration ($H = 1.374$).The hazard ratios of the Pregibon and Gosset models are under estimated at the shortest and long duration (H=0.939 and 0.946 respectively) compared to those of the correct link (H=1.225). This confirms the results in Table (4.1) above that the choice of a link function affects the strength of the estimated covariates effects. The hazard ratios differ substantially between models at the shortest duration ($t = 1$) and long duration ($t = 10$).

The sample size on the upper part of the table is 1500 with approximately 5264 binary observations. It is then increased to 3000 with approximately 11057 binary observations on the lower part of the table. The standard errors of the parameter estimates becomes smaller when the sample size is increased. Increasing the sample size does not change the biasness of the covariates. However when the sample size is increased the results for the flexible Pareto model are almost as acurate as the ones obtained from the logit link function.

4.3 Results for models with Frailty

This section provides the results and interpretation of models with frailty. All models used object id as a random effect.

Again, the pattern of results in Table 4.2 in terms of how the generalized logistic

distribution family compares with the logit, cloglog and probit is comparable to those of Hess et al. (2016), even if they generated their data from the Burr distribution.

The results displayed in Table 4.2 below are the averages and standard deviations found after performing the simulation exercise 100 times. It is then found that there is a difference in the parameter estimates when the random intercept is included in the models.

Table 4.2: Comparing Covariate Effects Across Model Specifications.

Model estimates						
	Logit	Cloglog	Probit	Pareto ($\xi = 1$)	Pareto($\xi = \xi$)	Pregibon
γ_1	0.511 (0.039)	0.442 (0.023)	0.453 (0.028)	0.527 (0.037)	0.289 (0.093)	0.458(0.014)
σ	0.212 (0.177)	0.069 (0.077)	0.300 (0.125)	0.292 (0.195)	0.295 (0.376)	0.023
Hazard Ratio at $t = 1(x_1)$	1.264 (0.016)	1.285 (0.019)	0.943 (0.016)	1.261 (0.0111)	1.217 (0.094)	0.935 (0.017)
Hazard Ratio at $t = 10(x_1)$	1.591 (0.053)	1.443 (0.031)	1.317 (0.026)	1.1601 (0.054)	1.712 (0.204)	1.404 (0.020)
N	1500	1500	1500	1500	1500	1500
γ_1	0.512 (0.026)	0.439 (0.017)	0.473 (0.023)	0.508 (0.024)	0.313 (0.077)	0.484(0.012)
σ	0.231 (0.151)	0.046 (0.064)	0.325 (0.081)	0.212 (0.144)	0.371 (0.245)	0.034(0.006)
Hazard Ratio at $t = 1(x_1)$	1.241 (0.001)	1.281 (0.015)	0.949 (0.0128)	1.249 (0.013)	1.237 (0.038)	0.941 (0.008)
Hazard Ratio at $t = 10(x_1)$	1.561 (0.034)	1.431 (0.024)	1.358 (0.017)	1.578 (0.036)	1.589 (0.106)	1.485(0.011)
N	3000	3000	3000	3000	3000	300

Rows 2 and 7 show the variance of the random effect which gives us the evidence that there is significant within group correlation in all models since the values of σ are both not equal to zero.

When the random effect is included in the models covariates effects of the flexible Pareto model i.e. Pareto ($\xi = \xi$) differ markedly ($\gamma_1 = 0.289$) from those obtained from the correct model i.e. logit model ($\gamma_1 = 0.511$). This is probably due to the poor choice of the optimal value of $\xi = 1$. The covariates effects of the Pareto ($\xi = 1$) ($\gamma_1 = 0.527$) are almost the same as those of the correct model ($\gamma_1 = 0.511$), while the covariates effects of cloglog and probit are under estimated ($\gamma_1 = 0.442$ and $\gamma_1 = 0.453$ respectively). The Pregibon model ($\gamma_1 = 0.458$) outperforms the flexible Pareto model with family parameter (ξ).

All $p-$values are found to be less than 0.001, thus all variables used are significant at 5% level of significance.

The results show that there is a huge difference in the hazard ratios at $t = 1$ and at $t = 10$. The hazard ratio of the true model is found to be 1.264 at the shortest duration ($t = 1$) and 1.591 at the long duration. The hazard estimates obtained from cloglog model exhibit a larger bias ($H = 1.285$) than the probit ($H = 0.943$), Pareto ($\xi = \xi$) ($H = 1.201$) and Pareto($\xi = 1$) ($H = 1.261$) at shortest duration. This might be due to the fact that the cloglog link function has a fixed negative skewness resulting in its lack of flexibility to let the data tell how much skewness should be incorporated and the ability to allow for positive skewness.

Again the sample size is increased from 1500 with approximately 48384 binary observations to 3000 with approximately 98487 binary observations. The parameters of the misspecified link functions remain biased even after increasing the sample size. However, the estimates exhibit smaller standard deviation when the sample size is increased.

4.4 Conclusion

The flexible families of link functions seem to perform quite well on the model estimation than the probit and the cloglog model, folowed by the Pregibon model.

Including unobserved heterogeneity did not help reducing the biasness either, it made the covariates effect to be highly under estimated for all models. However when the unobserved heterogeneity is included, the Pregibon model out performs the Pareto model with ($\xi = \xi$).

Increasing the sample size did not help reducing the biassness, however it worked in making the standard deviation smaller.

There is evidence of frailty among all models with frailty since the variance of random effect is not equal to zero.

Chapter 5

Real data analysis

5.1 Introduction

In this chapter the performance of the Pareto hazard model in real data analysis of the 2006-2007 Swaziland Demographic and Health survey data is analysed. The Pareto hazard model is then compared with other discrete models (logit, probit, cloglog and Pregibon), to investigate the importance of the choice of link function in discrete-duration models in practice and to determine the factors that explain the variation in women's age at first marriage. All the analysis was done using R software program (R×643.3.1).

5.2 The data and variables in the study

The real data analysis is based on data obtained from the Swaziland Demographic and Health Survey (SDHS) carried out between July 2006 and March 2007. The survey is the first of its kind to be conducted in Swaziland. The SDHS involved three basic questionnaires which are: The questionnaire on household, individual women and men questionnaire. Age at first marriage is the dependent variable measured as

the length of time from birth until the age at first marriage in years. During the survey women were asked questions regarding their marital status and whether they had ever lived with a man or not. Those who indicated that they have lived with a man or they were married before were asked how old they were when they started to live with a man. The answer to this question was considered as the woman's age at first marriage. Those women who had not experienced the event by the time of the survey were considered as censored at their age on the occasion of the survey. A total of 4987 women of reproductive age (15-49) were interviewed and information on socio-economic variables and age at first marriage were collected. A household was defined as one or more people related or unrelated living together and sharing food resources, and we take this sampling unit as a family unit for the current study.

Five explanatory variables were included for analysis in this study to examine the determinant factors for the timing of age at first marriage and to compare the performance of the link functions. All variables were selected for specific reasons: The region is used to study ethnic differentials, place of residence is used to capture the effects of urbanization and modernization, the year of birth and the age group are used to capture the effect of changing time on early marriages and, finally, the woman's educational level is also used to capture the effects of modernization (Manda and Meyer, 2005). Data was first converted to long format in R software programme before analysis, to meet the discrete standard.

5.3 Univariate analysis

5.3.1 Introduction

In this section we will explain the covariates that will be included in this study and present them in tables and graphical form.

Figure 5.1 shows that majority of women are between the ages of 15 and 20. About 30% of women in the sample are at age of 21 years.

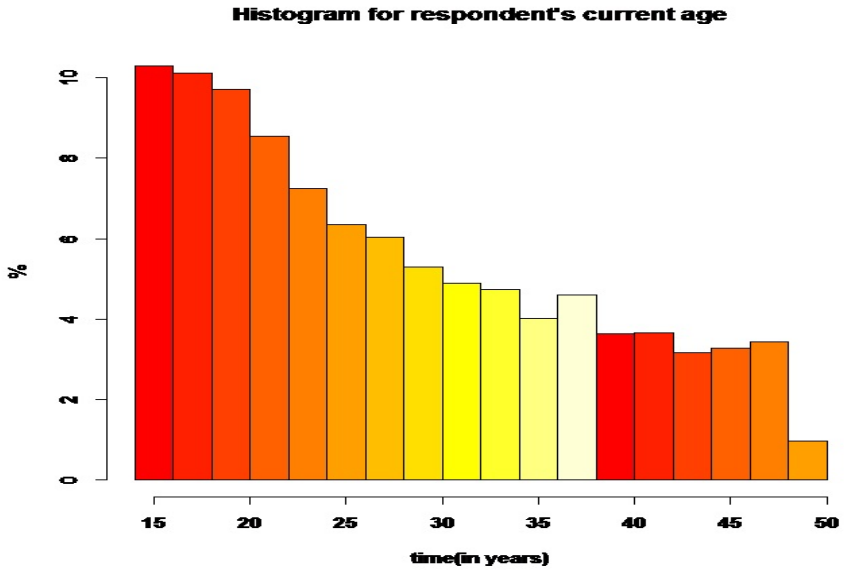

Figure 5.1: Histogram for respondent's current age.

Place of residence is a binary variable having two levels coded 0 for rural and 1 for urban. More than 50 % of women in Swaziland stay in rural areas while only 3 in 10 stay in urban areas. The results are displayed on Figure 5.2 and Table 5.1.

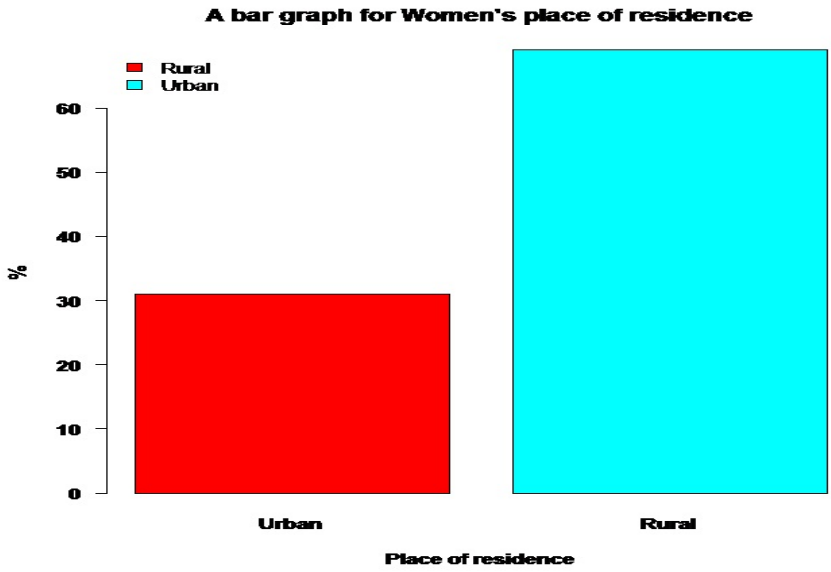

Figure 5.2: A bar graph for place of residence.

Table 5.1: Total number of women and percentages for residence.

Place of Residence	Unweighted number	Unweighted percentage
Urban	1544	69.0
Rural	3443	31.0
Total	4987	100

Women in the Manzini region accounted the highest part of the sample which is 3 in 10 while the remaining part was for women in the Hhohho, Lubombo and Shiselweni regions who accounted less than 30 % of the sample each. The results are displayed on Figure 5.3 and Table 5.2.

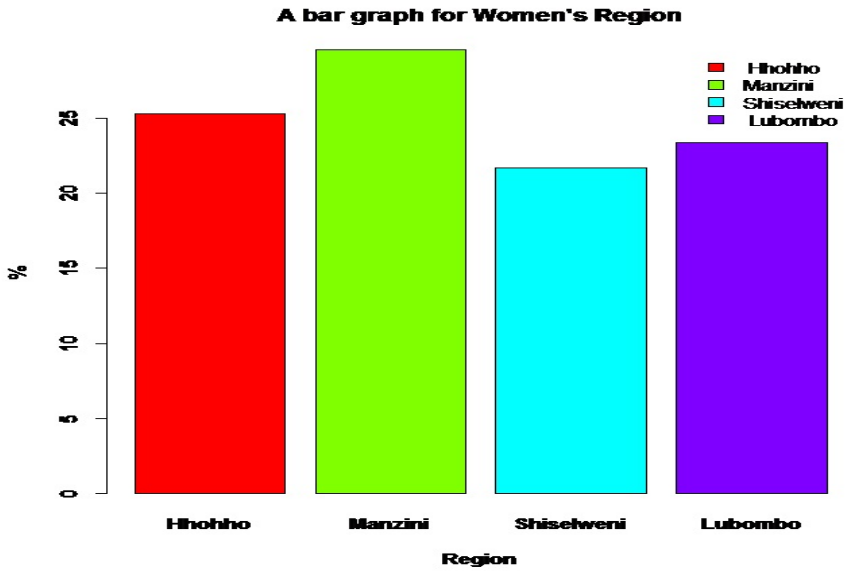

Figure 5.3: A bar graph for region.

Table 5.2: Total number of women and region percentages.

Region	Unweighted number	Unweighted percentage
Hhohho	1263	25.0
Lubombo	1166	23.0
Manzini	1475	30.0
Shiselweni	1083	22.0
Total	4987	100

Education is categorized as : No education, primary, secondary and higher. Categories are coded 0 to 3 corresponding respectively to the levels of education mentioned above. Secondary education is the modal category which accounted 51 % of the sample. Women with no education and those with higher education accounted the smallest part of the sample which is less than 10 % each as shown in Figure 5.4 and Table 5.3.

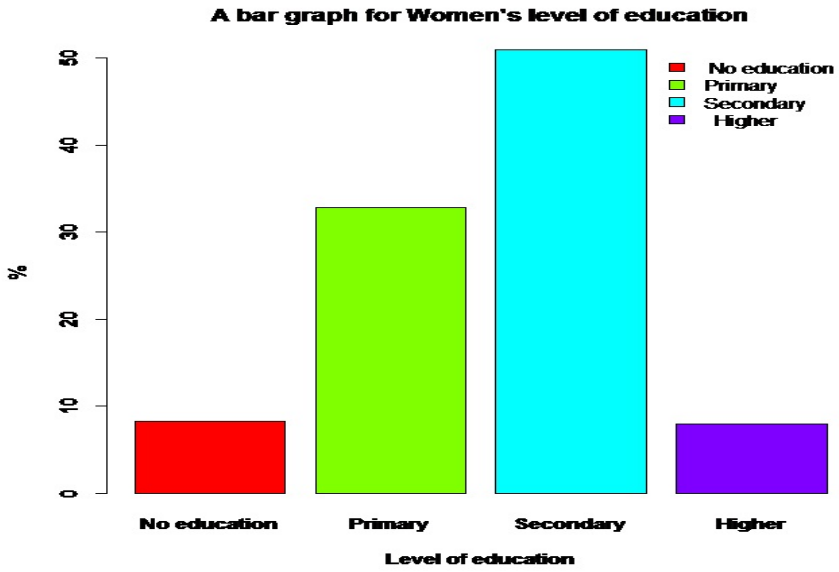

Figure 5.4: A bar graph for level of education.

Table 5.3: Total number of women and level of education percentages.

Education Level	Unweighted number	Unweighted percentage
No Education	413	8.0
Primary	1636	33.0
Secondary	2541	51.0
Higher	397	8.0
Total	4987	100

Table 5.4 and Figure 5.5 shows that majority of women were born in the year 1980to 1988. About 22% of women in the sample were born in the year 1972 to 1980.

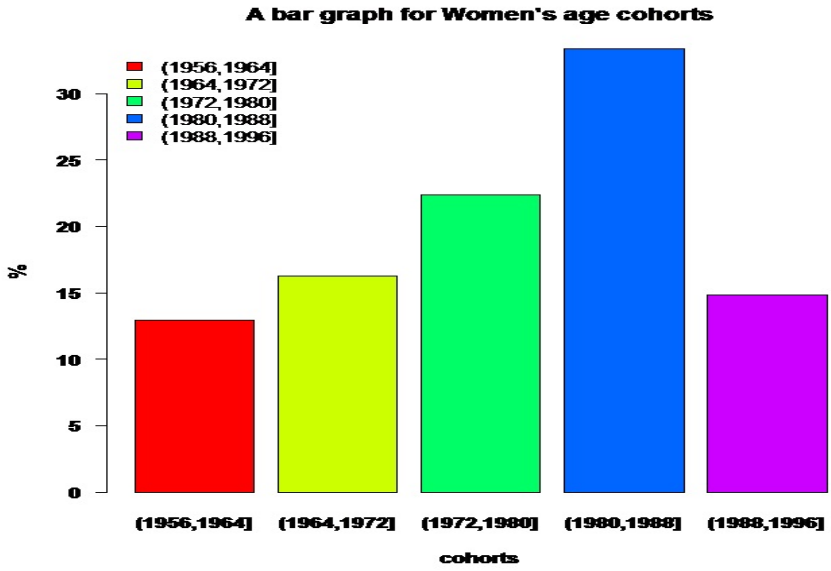

Figure 5.5: Bar graph for year of birth.

Table 5.4: Total number of women and percentages for Year of birth.

Year of birth	Unweighted number	Unweighted percentage
1956-1964	654	13.114
1964-1972	812	16.282
1972-1980	1117	22.398
1980-1988	1663	33.347
1988-1996	741	14.859
Total	4987	100

Figure 5.6 displays the corresponding hazard rates estimates of transition from birth to age at first marriage where there is an increase of first marriages from the age of 14 to 19 years and a pick at 27 years old.

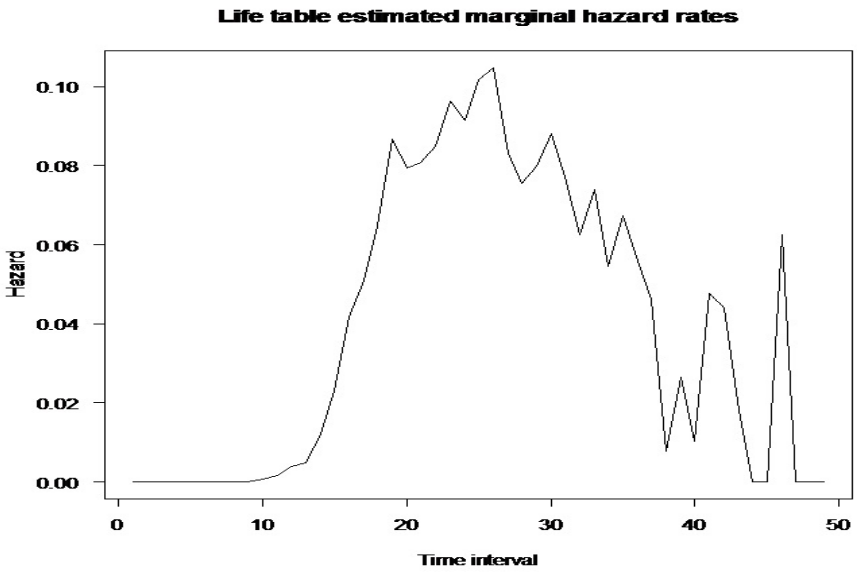

Figure 5.6: Raw hazard rate estimate of AFM.

In Figure 5.7 the wider the confidence interval, the late first marriages. Figure 5.7 which is a smoothed baseline hazard also shows a peak at the age 27 years.

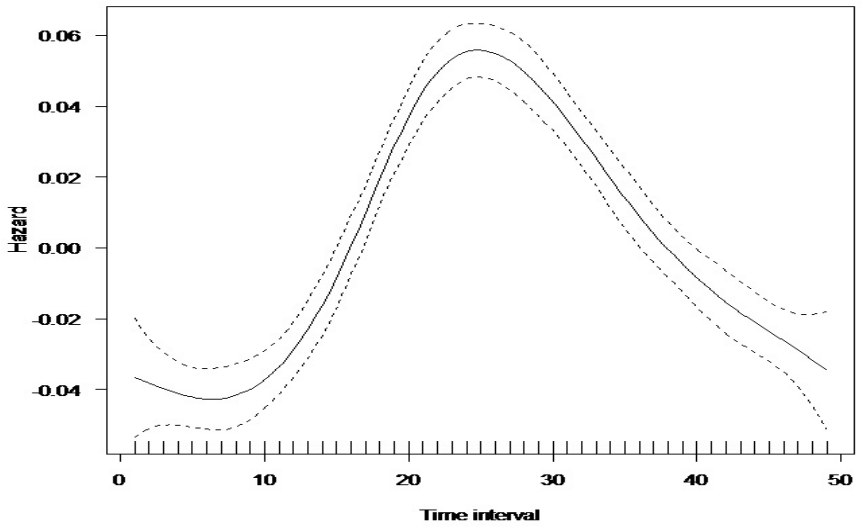

Figure 5.7: Smoothed hazard rate estimate of AFM.

5.4 Bivariate analysis

Here the possible effects that can be expected when the categorical variables are used in the model are examined by exploring the corresponding within groups sample survival functions. The log rank test is employed to assess if there is a significant difference in the survival patterns groups and all the $p-$values were found to be zero.

In Figure 5.8 women who reside in rural areas seem to have greater probability of experiencing early first marriages compared to those who reside in urban areas. At the age of 28 years only 34% of women in rural areas were not married in comparison to about 47% for women in urban areas.

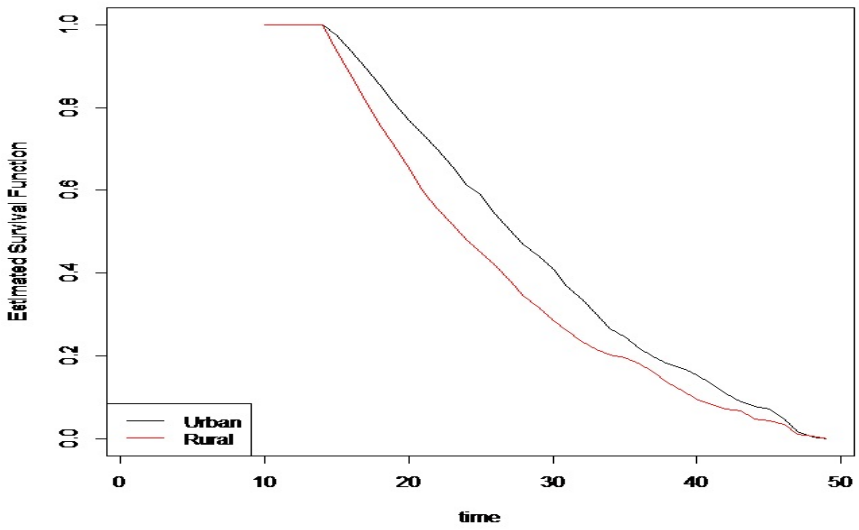

Figure 5.8: Estimated survival function for place of residence.

Figure 5.9 suggests that women who stay in the Hhohho and Lubombo region have lower probability of surviving early first marriages in comparison to those who stay in the Shiselweni and Manzini regions. Women in the Hhohho and Lubombo regions have only 30% and 33% probability of surviving early first marriages respectively at the age of 30 years, in comparison to about 36% women in the Manzini and Shiselweni regions.

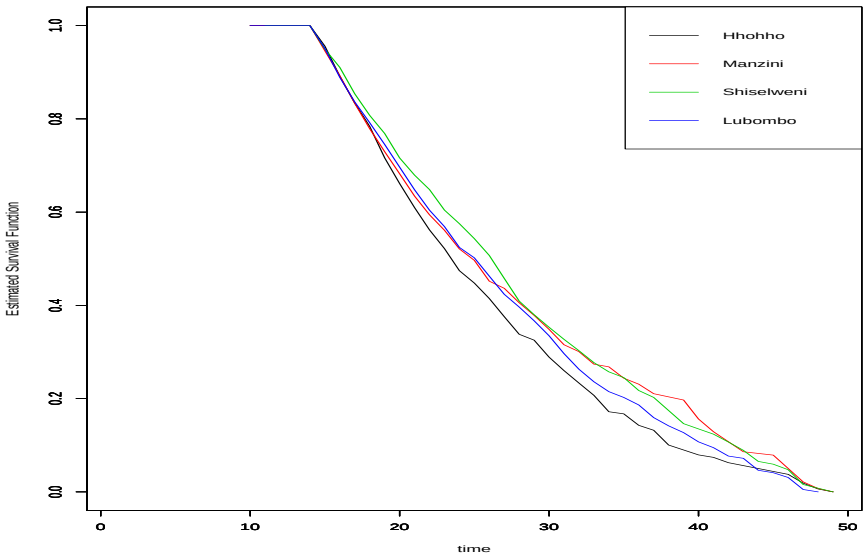

Figure 5.9: Estimated survival function for region.

Figure 5.10 women with no formal education and primary education seem to have lower probability of surviving early first marriages compared to those with higher and secondary education. At the age of 20 years, women who survived early first marriages for both categories no education and primary is only 64% in comparison to 90% and 97% for women with secondary education and higher respectively.

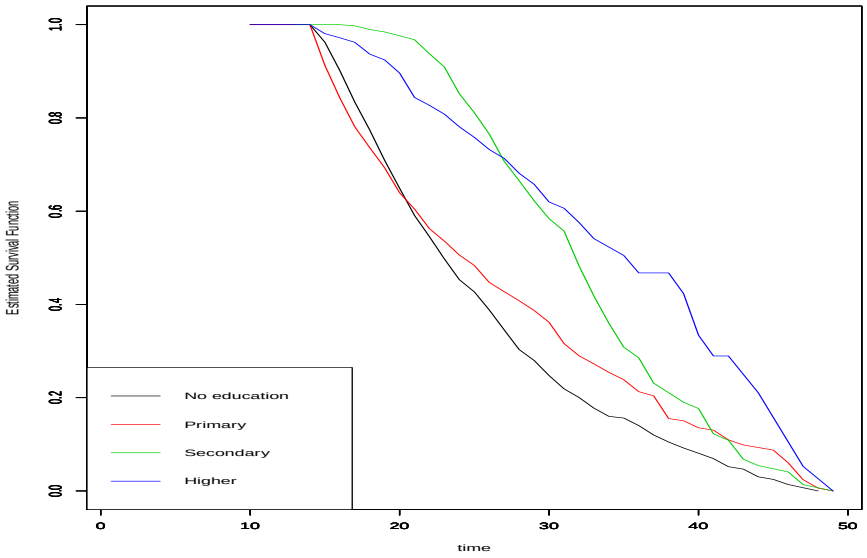

Figure 5.10: Estimated survival function for level of education.

Figure 5.11 shows that there is a pattern of early first marriages with older women getting married late compared to younger women.

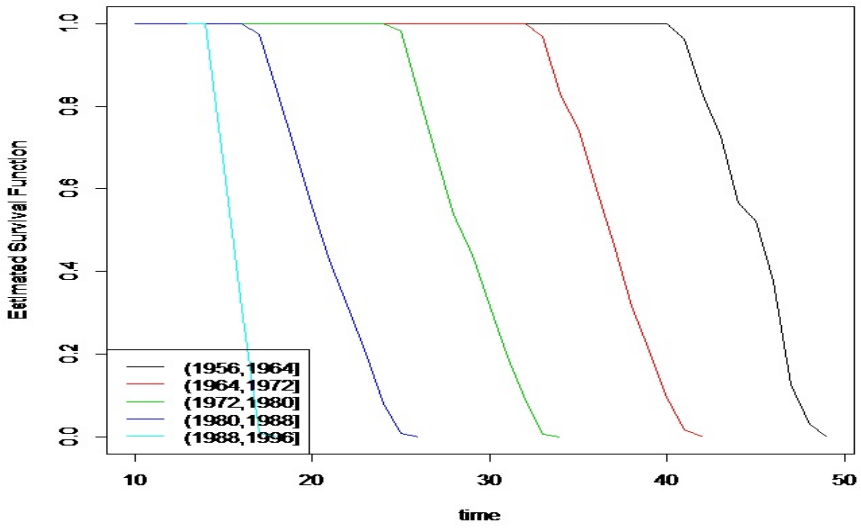

Figure 5.11: Estimated survival function for age cohorts.

5.5 Multivariate analysis

5.5.1 Introduction

Here we will compare and interpret different models according to different link functions and according to Swaziland Demographic and Health Survey data. Below is Table 5.5 indicating variables used with categories and reference baseline categories.

Table 5.5: Variables considered in the study with categories and reference baseline categories.

Variables	Categories	Reference
Place of residence	Rural, Urban	Urban
Region	Hhohho, Manzini, Shiselweni, Lubombo	Hhohho
Education	No education, Primary, Secondary, Higher	No education
Age cohort	1956-1964, 1964-1972, 1972-1980, 1980-1988, 1988-1996	1956-1964

5.5.2 Specification of the model

Two variables were created which are: time indicating each i^{th} interval and censoring-time coded as 1 if the event of interest occurred and 0 if the event did not occur within the interval. For example, if during the observational period a woman was 38 years old and she indicated that she was married at the age of 19, the variable time will take on the value 19 and censoring-time will take the value 1 since the woman have experienced an event. However if during the observational period a woman was 38 years old and she indicated that she was never married, the variable time will take the value 38 and censoring-time will take 0 since the woman did not experience the event. Below is Table 5.6 which shows the total number and percentages of married and unmarried women as of the time of the survey.

Table 5.6: Censoring table.

Censoring	Total	Percentage
0	2486	49.8
1	2501	50.2

Variable time and censoring time are then combined to form time to age at first marriage. Time to age at first marriage is grouped into seventeen duration dummies to avoid having too many dummy variables. All the models used in this study include the duration dummies. Models with frailty include family as the random effect.

The value of ξ for Pareto model with family parameter ξ is estimated at regular fixed values of ξ starting from 0.01 to 20 moving in steps of 0.1 allowing R software programme to select the position where the smallest value of AIC is from the model using 'gamlss.mx' package. When R is done selecting the position it is then allowed

to select the value in that particular position and the value is taken as the value of the family parameter ξ (Hess et al., 2014; Stasinopoulos and Rigby, 2012).

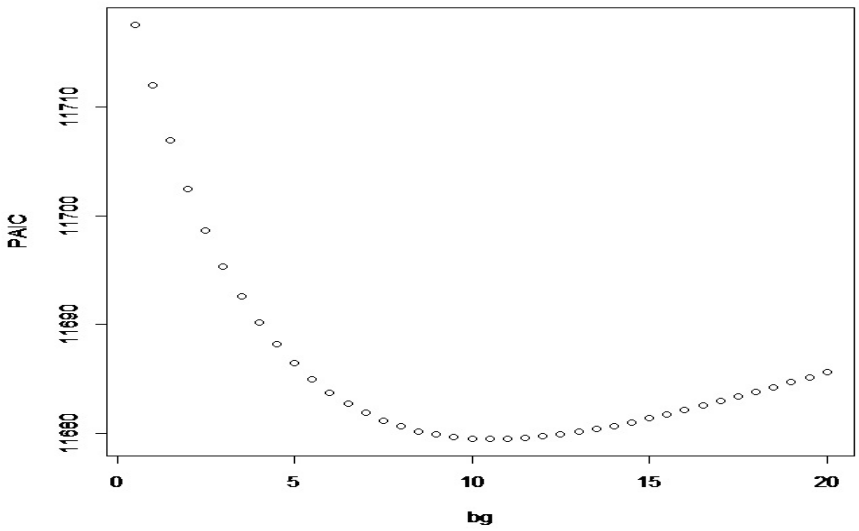

Figure 5.12: Akaike information criterion (AIC).

Parameters a and b for the Pregibon model are estimated in the same way as family parameter ξ of the Pareto model. However the only difference is that for the Pregibon family we have to estimate two parameters starting from 0.01 to 20 moving in steps of 0.1.

5.5.3 Results after fitting models with several link functions

Here we present the estimated coefficients and the p-values found after fitting the discrete multilevel models which are: the Pareto ($\xi = \xi$), logit, cloglog, probit and the Pregibon model. The family parameter (ξ) represents the shape parameter in the Pareto model. The $p-$values measure the statistical significance of an explanatory variable, indicating wether the effect of an explanatory variable on the hazard is systematically different from zero or not. The results for models without frailty are displayed in Table 5.7 and the results for models with frailty are not shown due to the fact that the random effect is insignificant across all models at 5% level of significance.

The value of ξ for Pareto model with family parameter ξ was found to be 10.5 and the results are displayed In Figure 5.12.

Table 5.7: Estimation Results for the 2006-2007 Swaziland Demographic and Health Survey data with $p-$values inside the brackets.

Explanatory variables	Pareto($\xi = \xi$)	Logit	Cloglog	Probit	Pregibon
(Place of residence)					
Rural	0.472 (0.000***)	0.313 (0.000***)	0.301 (0.000***)	0.159 (0.000***)	0.286 (0.000 ***)
(Region)					
Manzini	-0.221 (0.024*)	-0.127 (0.031*)	-0.115 (0.034 *)	-0.06760 (0.029 *)	-0.119 (0.021 *)
Shiselweni	-0.552 (0.009***)	-0.337 (0.014***)	-0.277 (0.014 ***)	-0.163 (0.006 ***)	-0.284 (0.009 ***)
Lubombo	0.061 (0.532)	0.045 (0.466)	0.040 (0.478)	0.023 (0.489)	0.041 (0.478)
(Level of education)					
Primary	0.503 (0.000***)	0.291 (0.000 ***)	0.249 (0.000 ***)	0.152 (0.000 ***)	0.267 (0.000 ***)
Secondary	-0.275(0.041*)	-0.100 (0.191)	-0.101 (0.144)	-0.06371 (0.115)	-0.104 (0.144)
Higher	-0.431 (0.009**)	-0.167 (0.091 ·)	-0.16092 (0.076 ·)	-0.104 (0.043 *)	-0.171 (0.061 ·)
(Year of birth)					
$(1964, 1972]$	-0.0163 (0.893)	0.050 (0.434)	0.086 (0.138)	0.021 (0.543)	0.043 (0.479)
$(1972, 1980]$	-0.571 (0.006***)	-0.242 (0.000 ***)	-0.177 (0.002 **)	-0.139 (0.030 ***)	-0.234 (0.057 ***)
$(1980, 1988]$	-0.972 (0.000***)	-0.511 (0.000 ***)	-0.433 (0.000 ***)	-0.275 (0.000 ***)	-0.478 (0.000 ***)
$(1988, 1996]$	-2.672 (0.000***)	-1.835 (0.000 ***)	-1.659 (0.000 ***)	-0.874 (0.000 ***)	-1.605 (<0.000 ***)
(Shape Pameter)					
ξ	10.5	1	0		
(Number of observations)					
N	22099	22099	22099	22099	22099
AIC	14450.11	14485.26	14497.94	14473.68	14478.88
Parameter estimates					
a					0.07
b					0.01

5.5.4 Comparison of results according to all link functions used

In this section we compare the results in Table 5.7 according to different link functions used in this study.

Here different models give different results. The effect of covariates obtained from different models exhibit substantial difference. The covariate effect of the Pareto model differs markedly with those of the remaining models. This suggests that the estimated effects of explanatory variables are affected by the link function chosen when specifying the discrete hazard.

Explanatory variables are not changing signs across all models except for the category 1964 to 1972 for variable year of birth, and it is found to be insignificant at 5% level of significance across all models($p-$values are greater than 0.05). Variables place of residence (rural), level of education (primary) and age cohort (1980 to 1988 and 1988 to 1996) are found to be significant with $p-$values equal to zero across all models. The Pareto model exhibits extremely large and extremely small values of covariates effect through out.

When comparing the AIC values of various models which are displayed in Table 5.7, the smallest value of AIC is found for the Pareto model with family parameter ξ (AIC=14450,11). This suggests that the Pareto model best fits the SDHS data. The AIC of the Pregibon model is found to be 14478,88, which is greater than the AIC of the Pareto model with family parameter ξ and the probit model (AIC=14473.68). The AIC of the Pregibon model is however less than the AIC of the logit (AIC=14485.26)

and cloglog model (AIC=14497.94). This suggests that the Pregibon model outperforms the cloglog and logit on the analysis of SDHS data based on the AIC values.

It is some what surprising that the second smallest value of AIC is found for the probit model which is not nested in the Pareto family of link functions.

5.6 Interpretation of the results of the discrete survival model based on the Pareto model with family parameter ξ

In this section we employ the Pareto model with family parameter ξ, since it is found to be the best model based on the value of AIC.

Table 5.8: Estimated coefficients, standard errors,t values, two-Tailed $p-$values variables signicant at the 5% level in Pareto with family parameter ξ.

	Estimate	Std. Error	t value	Pr($>|t|$)	Hazard ratio(HR)
Place of residence (Rural)	0.472	0.084	5.628	0.001 ***	1.603
Region (Manzini)	-0.221	0.102	-2.263	0.024 *	0.802
Region (Shiselweni)	-0.552	0.113	-4.895	0.009 ***	0.576
Region (Lubombo)	0.061	0.112	0.624	0.532	1.063
Level of education (Primary)	0.503	0.131	3.605	0.000 ***	1.654
Level of education (Secondary)	-0.275	0.135	-2.039	0.041 *	0.751
Level of education (Higher)	-0.431	0.169	-2.598	0.009 **	0.641
Year of birth (1964,1972]	-0.016	0.121	-0.135	0.893	0.984
Year of birth (1972,1980]	-0.571	0.116	-4.984	0.006 ***	0.565
Year of birth (1980,1988]	-0.972	0.123	-7.874	0.000 ***	0.378
Year of birth (1988,1996]	-2.671	0.297	-8.998	$<$ 0.000 ***	0.069

Table 5.8 shows that women who reside in rural areas have a hazard ratio of (HR=1.603). The hazard of early first marriage for women residing in rural areas is 60.3% higher than the hazard for women residing in urban areas. The positive sign on the category rural indicates that women residing in rural areas have a higher risk of experiencing their first marriages early.

From Table 5.8 the hazard ratios (HRs) for regions Manzini and Shiselweni are found to be 0.802 and 0.576 respectively. The hazard of early first marriage for women in the Manzini and Shiselweni regions are 19.8% and 42.4% respectively lower than

the hazard for those in the Hhohho region. The sign of the coefficients are negative (-0.221 and -0.552 respectively) which implies that the risk of early first marriages is lower for women in the Manzini and Shiselweni regions as compared to Hhohho region. The hazard of early first marriage in the Lubombo region is not different from that of Hhohho region.

With regard to the educational level of the woman, risk of early first marriage was signicantly lower among those women with higher education (HR=0.641) compared to those with no education. The positive sign on the category primary (0.503) shows that women who are not well educated tend to marry at an early age. The hazard of early first marriage for women with higher education is 35.9% higher than the hazard for women with no formal education.

The hazard of early first marriages for women who were born from the year 1988 to 1996 is 1.6% lower than the hazard for women who were born from the year 1956 to 1964. This suggests that younger women are getting married at a later age compared to older women.

5.6.1 Hazard functions for the fitted Pareto model with family parameter ξ

This section presents the hazard functions of all categories used in this study based on the fitted Pareto model with family parameter ξ. The time interval is reduced from 17 to 16 due to data sparsity, since at the end of the interval there are fewer people and as a result predictions are affected.

Figure 5.13: Estimated discrete hazard rate of place of residence.

Figure 5.13 shows the estimated discrete hazard function of women's place of residence. It is revealed that the hazard of early first marriages is considerably higher among women who reside in rural areas than those in urban areas. This confirms the positive coefficient for category rural in Table 5.7.

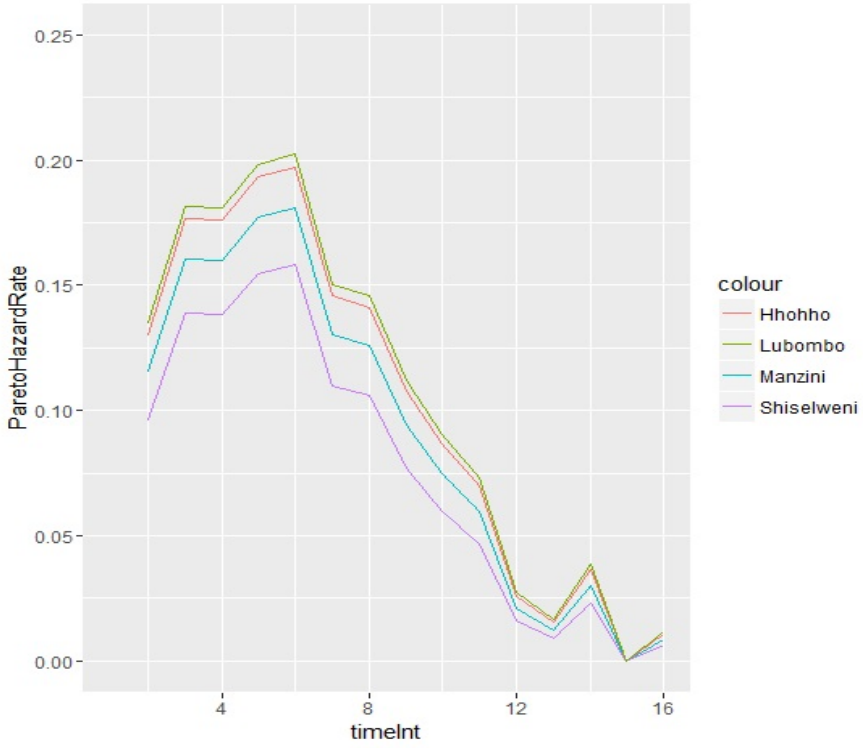

Figure 5.14: Estimated discrete hazard rate of region.

Figure 5.14 shows the fitted discrete hazard function for regions. It is seen that the risk of early first marriages is higher for women who stay in the Lubombo and Hhohho regions in comparison to those who stay in the Manzini and Shiselweni regions. The negative coefficients (-0.221 and -0.552) for the categories Manzini and Shiselweni in Table 5.7 give strong evidence that women from the Manzini and Shiselweni regions get married at a late age as compared to women from Lubombo and Hhohho regions.

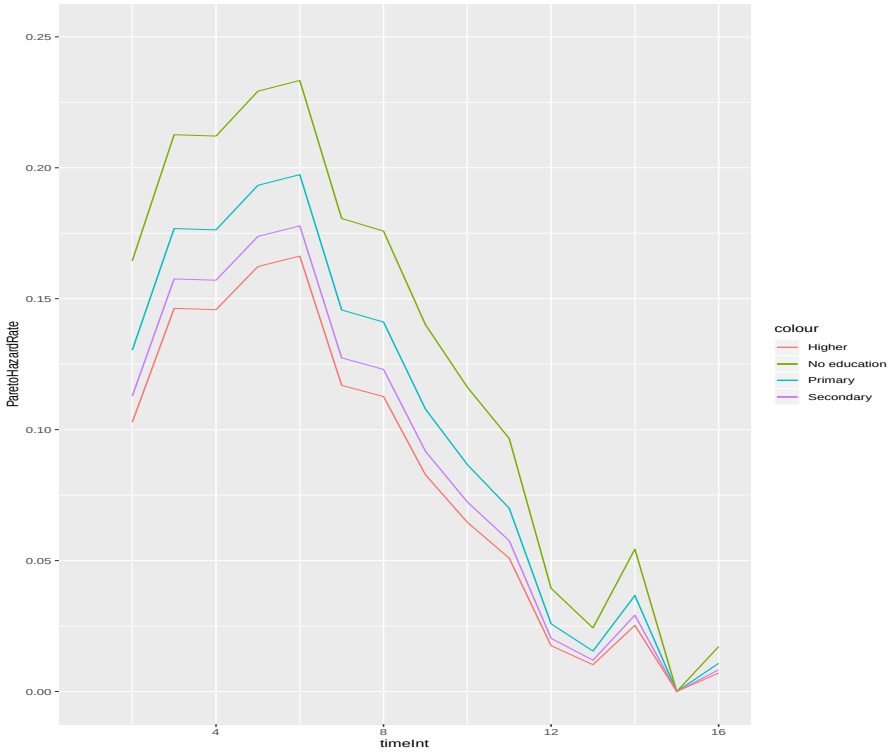

Figure 5.15: Estimated discrete hazard rate of level of education.

Figure 5.15 presents the fitted discrete hazard function of women's level of education. It is revealed that the hazard for women with no education is estimated with extremely higher risk of early first marriages in comparison of those with secondary and higher education. The huge gap between no education and higher education shows that education has a strong effect on women's age at first marriage.

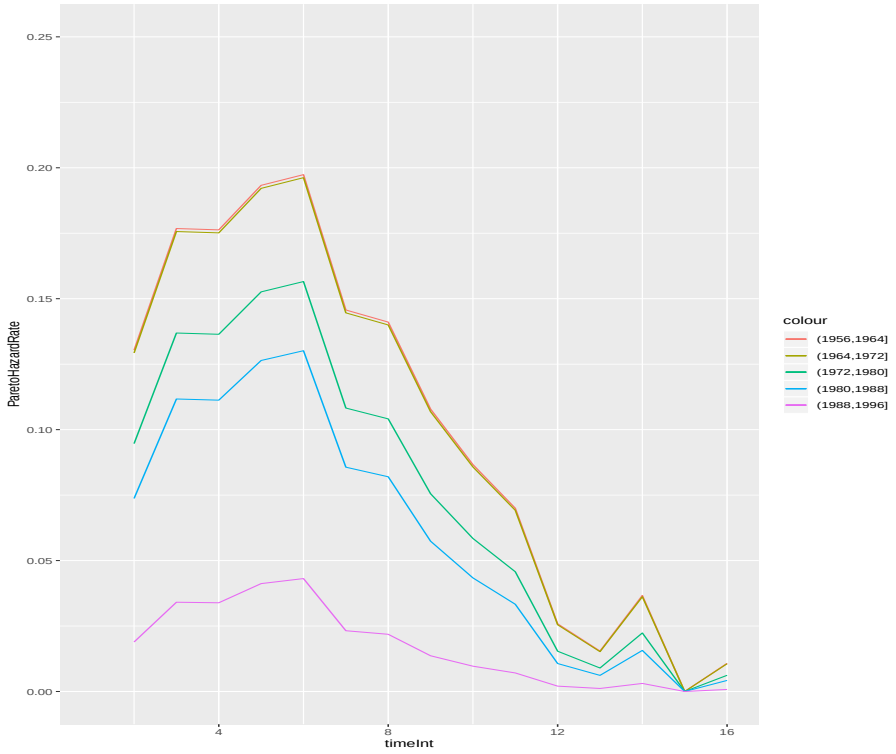

Figure 5.16: Estimated discrete hazard rate of age cohorts.

Figure 5.16 illustrates the fitted discrete hazard function of women's age cohorts. It is seen that the hazard rate for women who were born from the year 1988 to 1996 is extremely lower in comparison to all the remaining categories. This confirms the findings in Table 5.7, where the category 1988 to 1996 has a negative coefficient (-2.672) and a p-value of zero, which shows that women who were born from the year 1988 to 1996 get married at a later age compared to women who were born from the year 1956 to 1964 and 1964 to 1972. A huge gap between women who were born from the year 1988 to 1996 and other categories is because they were still teenagers when the 2007 Swaziland Demographic and Health Survey was conducted.

5.7 Conclusion

With regard to real data analysis, the Pareto model with family parameter ξ out performed all the remaining models which are: the logit, cloglog, probit and the Pregibon model since it was the most accurate in terms of recovering the coefficients of the model.

Regarding the Swaziland Demographic and Health Survey data, women who reside in rural areas are at lower risk of early first marriages. The hazard of early first marriages decreases with level of education. Regarding age cohorts, the hazard of early first marriages was found to increase with age.

Chapter 6

Conclusion

6.1 Conclusion and discussion

This study presented a comparison of flexible family of link functions which are: the Pareto hazard model, Pregibon and the Gosset model and the commonly used link functions which are: the logit, cloglog and the probit links. The link functions were compared through a simulation exercise and application to Swaziland Demographic and Health survey data.

Simulation studies based on a simulated data sets which consist of 1500 and 3000 time intervals were carried out. Two simulation experiments were performed. In the first experiment models without unobserved heterogeneity were considered and in the second experiment models with unobserved heterogeneity were considered.

The results from the simulation studies revealed that, when an incorrect link function is used, the covariate effects become biased. The performance of the Pareto model with family parameter ξ was found to be quite good throughout the simulation study

compared to the remaining models (probit, cloglog, Gosset and the Pregibon) since it gave the lowest AIC. Incresing the sample size did not help on reducing the biasness, it however helped to make the standard errors of the parameter estimates smaller.

The Pregibon family of link functions was found to be the second one on good performance, followed by the Gosset family of link functions compared to the traditional cloglog and probit links.

Introducing the random effect in the model did not help reducing the biasness, it caused the covariate effects to be under estimated. This finding is consistent with that of Hess et al. (2014).

Real data analysis in this study was based on a data set of women obtained from SDHS2006-07 with an aim of exploring the use of flexible families of link functions for discrete survival models in practice and to examine women's age at first marriage. Out of a total of 4987, about 49.8% experienced an event (got married) and 50.2% were not married for a different age between 15 and 49 years.

Women residing in rural areas were found to be at higher risk of early first marriages using urban as a reference category.

The findings of this study showed that women in the Manzini and Shiselweni regions prolonged age at first marriage by the factors of (HR = 0.802 and 0.576, respectively) when women from Hhohho region were used as the reference group.

The findings of this study revealed that educational level of women had a significant effect on the survival of age at first marriage with 5% level of significance since the category primary had a $p-$value of zero. It was also indicated through the hazard function that there was a significantly higher risk of early first marriages among those women with no formal education compared to those with higher education. A study conducted in Ethiopian regions by Tessema et al. (2015) investigated the determinants of time-to-age at first marriage and their results also indicated that women who had higher education were found to survive early first marriages more than those who are uneducated and primary education. This might be due to the fact that highly educated women take education seriously than marriage. These findings are also in line with results from other studies (Agaba et al., 2010; Kumchulesi et al., 2011; Haloi and Limbu, 2013).

The difference among birth cohorts was found to be very high, with women of younger cohorts getting married late compared to elderly women. These results confirm the findings of the US Bureau of the census (1996) where they performed a study on comparative marriage rates in sub-Saharan Africa and it was found that the proportion of marriages is declining. This finding is also in line with the results from Manda and Meyer (2005), DVera et al. (2011), Fry (2014) and Chizuko (1998).

In particular, the findings of this study shows that misspecification of the link function causes severe bias on the estimation of results. It was then proved in this study through simulation study that, biasness can be avoided by making use of flexible

families of link functions.

Regarding marriage, in particular the timing of marriage among women was influenced by educational level, place of residence, year of birth and region. Women who are highly educated, tend to live in the Manzini and Shiselweni regions, those who reside in urban areas were more likely to marry later compared to their counterparts in Swaziland. This might be due to the fact that most families in rural areas are poor and they find it hard to choose between sending sons and daughters to school, they end up sending their sons often. Government should make education free and compulsory, to eliminate gender disparity in education and employ more women teachers who can serve as role models.

6.2 Limitation of the study

The data set used in the simulation study was generated from the logit link which is a member of the Log Burr family of link functions, Gosset and the Pregibon family, and no effort was made to asses the performance of link function using data generated from a link function which is not a member of a particular family. The results of the Preginon model with unobserved heterogeneity in the simulation exercises was simulated 20 times. The Pregibon goodness of link test which asses the suitability of the link function, was not used in this study.

6.3 Future research

Future researchers may conduct a study where they generate data from a link function which is not a member of the families they are using on model estimation, and compare the results with the one they get after estimating the model using the data generated from a member of the family. A study which assess the suitability of link function using the Pregibon goodness of link test is needed. Additionally, a study which include both men and women on the timing to age at first marriage need to be conducted in Swaziland.

References

[1] P. Agaba, L. Atuhaire, and G. Rutaremwa, *Determinants of age at first marriage among women in Western Uganda*, Acta. Math. (September 2010), 1–4, Presented in European Population Conference 2010 Vienna,.

[2] T. Amemiya, *Qualitative response models: A survey*, Journal of economic literature **19** (1981), no. 4, 1483–1536.

[3] F. J. Aranda-Ordaz, *An extension of the proportional-hazards model for grouped data*, Biometrics (1983), 109–117.

[4] T. R. Aryal, *Age at first marriage in Nepal: differentials and determinants*, Journal of Biosocial Science **39** (2007), no. 05, 693–706.

[5] M. Baker and A. Melino, *Duration dependence and nonparametric heterogeneity: a Monte Carlo study*, Journal of Econometrics **96** (2000), no. 2, 357–393.

[6] I. Baldi, M. Maule, R. Bigi, L. Cortigiani, Simona Bo, and D. Gregori, *Some notes on parametric link functions in clinical research*, Statistical methods in medical research **18** (2009), no. 2, 131–144.

[7] N. E. Breslow and D. G. Clayton, *Approximate inference in generalized linear mixed models*, Journal of the American statistical Association **88** (1993), no. 421, 9–25.

[8] J. Buckley and C. Westerland, *Duration dependence, functional form, and corrected standard errors: Improving eha models of state policy diffusion*, State Politics & Policy Quarterly **4** (2004), no. 1, 94–113.

[9] U. Chizuko, *The declining birthrate: Whose problem*, Review of Population and Social Policy **7** (1998), 103–128.

[10] D. G. Clayton, *A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence*, Biometrika **65** (1978), no. 1, 141–151.

[11] D. Collett, *Modelling survival data in medical research*, Chapman and Hall & CRC, 2015.

[12] M. Cordeiro and M. de Castro, *A new family of generalized distributions*, Journal of statistical computation and simulation **81** (2011), no. 7, 883–898.

[13] D. R. Cox, *Analysis of survival data*, Routledge, 2018.

[14] D. R. Cox and D. Oakes, *Analysis of survival data (vol. 21).*, CRC Press, 1984.

[15] C. Czado and A. E. Raftery, *Choosing the link function and accounting for link uncertainty in generalized linear models using Bayes factors*, Statistical Papers **47** (2006), no. 3, 419–442.

[16] C. Czado and T. J. Santner, *The effect of link misspecification on binary regression inference*, Journal of statistical planning and inference **33** (1992), no. 2, 213–231.

[17] P. Dierckx, *On calculating normalized Powell-Sabin B-splines*, Computer Aided Geometric Design **15** (1997), no. 1, 61–78.

[18] C. DVera, W. Wang, and G. Livingston, *Barely half of US adults are married–a record low*, Pew Research Social & Demographic Trends (2011).

[19] T. Ermans, *Multilevel survival analysis of determinants of age at first marriage among women living in Nigeria*, Master's thesis, 2010-2011.

[20] L. Fahrmeir and G. Tutz, *Models for multicategorical responses: Multivariate extensions of generalized linear models*, Multivariate statistical modelling based on generalized linear models, Springer, 2001, pp. 69–137.

[21] A. A. Fisher and A. A. Way, *The demographic and health surveys program: An overview*, International Family Planning Perspectives (1988), 15–19.

[22] R. Fry, *No reversal in decline of marriage*, Religion (2014).

[23] A. Groll and G. Tutz, *Variable selection in discrete survival models including heterogeneity*, Lifetime data analysis **23** (2016), no. 2, 1–34.

[24] R. U. Gromping, W. H. Finch, J. E. Bolin, and K. Kelley, *Multilevel modeling using R*, Citeseer, 2015.

[25] G. Guo, *Event-history analysis for left-truncated data*, Sociological methodology (1993), 217–243.

[26] A. Haloi and D. K. Limbu, *Socio-economic factors influence the age at first marriage of muslim women of a remote population from north-east india.*, Sociology **75** (2013), 79.

[27] W. Hess and M. Persson, *The duration of trade revisited*, Empirical Economics **43** (2012), no. 3, 1–25.

[28] W. Hess, G. Tutz, and J. Gertheiss, *A flexible link function for discrete-time duration models*, Jahrbücher für Nationalökonomie und Statistik **236** (2016), no. 4, 455–481.

[29] L. Ikamari, *The effect of education on the timing of marriage in Kenya* , A peer-reviewed, open-access journal of population sciences **12** (2005), 1–28.

[30] S. P. Jenkins, *Survival analysis*, Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK **42** (2005), 54–56.

[31] X. Jiang, D.K. Dey, R. Prunier, A. M. Wilson, and Kent. E. Holsinger, *A new class of flexible link functions with application to species co-occurrence in cape floristic region*, The Annals of Applied Statistics **7** (2013), no. 4, 2180–2204.

[32] M. Jones, *Families of distributions arising from distributions of order statistics*, Test **13** (2004), no. 1, 1–43.

[33] J. P. Klein and M. L. Moeschberger, *Survival analysis: techniques for censored and truncated data*, Springer Science & Business Media, 2006.

[34] R. Koenker and J. Yoon, *Parametric links for binary choice models: A Fisherian–Bayesian colloquy*, Journal of Econometrics **152** (2009), no. 2, 120–130.

[35] G. Kumchulesi, M. Palamuleni, and I. Kalule-Sabiti, *Factors affecting age at first marriage in Malawi*, Proceedings of the 6^{th} African Population confrence, Ouagadougou, Burkina Faso, 5-9 December 2011.

[36] S. Laaksonen, *Does the choice of link function matter in response propensity modelling?*, Model Assisted Statistics and Applications **1** (2005), no. 2, 95–100.

[37] X. Lin and N. E. Breslow, *Analysis of correlated binomial data in logistic-normal models*, Journal of Statistical Computation and Simulation **55** (1996), no. 1-2, 133–146.

[38] B. K. Mallick and A. E. Gelfand, *k/ld 2f3 7i*, (1994).

[39] S. Manda and R. Meyer, *Age at first marriage in Malawi: a Bayesian multilevel analysis using a discrete time-to-event model*, Journal of the Royal Statistical Society: Series A (Statistics in Society) **168** (2005), no. 2, 439–455.

[40] M. Mills, *Introducing survival and event history analysis*, Sage Publications, 2011.

[41] V. MR. Muggeo and G. Ferrara, *Fitting generalized linear models with unspecified link function: A p-spline approach*, Computational Statistics & Data Analysis **52** (2008), no. 5, 2529–2537.

[42] C. W. Nam, T. S. Kim, N. J. Park, and H. K. Lee, *Bankruptcy prediction using a discrete-time duration model incorporating temporal and macroeconomic dependencies*, Journal of Forecasting **27** (2008), no. 6, 493–506.

[43] C. Nicoletti and C. Rondinelli, *The (mis) specification of discrete duration models with unobserved heterogeneity: a Monte Carlo study*, Journal of Econometrics **159** (2010), no. 1, 1–13.

[44] US Department of Commerce and Statistics Staff, *Statistical abstract of the united states, 1996: The national data book*, Bernan Reprints, 1996.

[45] K. H. Pollock, S. R. Winterstein, and M. J. Conroy, *Estimation and analysis of survival distributions for radio-tagged animals*, Biometrics (1989), 99–109.

[46] D. Pregibon, *Goodness of link tests for generalized linear models*, Journal of the Royal Statistical Society: Series C (Applied Statistics) **29** (1980), no. 1, 15–24.

[47] R. L. Prentice, *A generalization of the probit and logit methods for dose response curves*, Biometrics **32** (1976), no. 4, 761–768.

[48] S. W. Raudenbush and A. S. Bryk, *Hierarchical linear models: Applications and data analysis methods*, vol. 1, Sage, 2002.

[49] J. D. Singer and J. B. Willett, *Its about time: Using discrete-time survival analysis to study duration and the timing of events*, Journal of Educational and Behavioral Statistics **18** (1993), no. 2, 155–195.

[50] J. D. Singer and J.B Willett, *Applied longitudinal data analysis: Modeling change and event occurrence*, Oxford university press, 2003.

[51] M. Stasinopoulos and B. Rigby, *gamlss. mx: A gamlss add on package for fitting mixture distributions*, R package version **4** (2012), 1–0.

[52] B. Tessema, S. Ayalew, and K. Mohammed, *Modeling the determinants of timeto-age at first marriage in Ethiopian women: Comparison of various parametric shared frailty models*, Science Journal of Public Health; 3 (5): 707 **718** (2015).

[53] G. Tutz and S. Petry, *Nonparametric estimation of the link function including variable selection*, Statistics and Computing **22** (2012), no. 2, 545–561.

[54] G. Tutz and M. Schmid, *Modeling discrete time-to-event data*, Springer, 2016.

[55] D. Wuertz, *Rmetrics core team members, uses code builtin from the following r contributed packages: gmm from pierre chauss, gld from robert king, gss from chong gu, nortest from juergen gross, hyperbolicdist from david scott, sandwich from thomas lumley, achim zeileis and fortran/c code from kersti aas. fbasics: Rmetrics-markets and basic statistics. r package version 2110.79*, R package version **2110** (2010).

[56] A. Zeileis, R. Koenker, and P. Doebler, *glmx: Generalized linear models extended*, R package version 0.1-0, URL http://CRAN. R-project. org/package= glmx (2013).

Apendix

```
##Generating data and fitting the logit model without frailty##


rm(list=ls())##cleaning the memory


set.seed(2311988)


BX1CL<-NULL

BX1CL2<-NULL

BX2CL<-NULL

BX2CL2<-NULL


nit<-10


cx1<-matrix(,nrow=nit,ncol=1)
```

```
cfz<-matrix(,nrow=nit,ncol=1)


cfd1<-matrix(,nrow=nit,ncol=1)

cfd12<-matrix(,nrow=nit,ncol=1)


sx1<-matrix(,nrow=nit,ncol=1)

sz<-matrix(,nrow=nit,ncol=1)

sH1<-matrix(,nrow=nit,ncol=1)

sH2<-matrix(,nrow=nit,ncol=1)


H1<-matrix(,nrow=nit,ncol=1)

H2<-matrix(,nrow=nit,ncol=1)


 for (l in 1:10){

print(l)


library(gamlss.mx)

burr <- function(xi=1)

{
```

```
  linkfun <- function(mu) log(((1-mu)^(-xi)-1)/xi)
  linkinv <- function(eta) 1-(1+xi*exp(eta))^(-1/xi)
  mu.eta <- function(eta) exp(eta)*(1+xi*exp(eta))^(-1/xi-1)
  valideta <- function(eta) TRUE
  link <- paste("burr(", xi, ")", sep="")
  structure(list(linkfun = linkfun, linkinv = linkinv,
                 mu.eta = mu.eta, valideta = valideta, name = link),
            class = "link-gamlss")
} #closing burr function

n<-3000

p<-12

k<-n*p
  id<-rep(1:n,each=p)
  t<-rep(1:p,times=n)

x1<-rep(rnorm(n,mean=0,sd=1.5),each=p)
 #b1<-rep(rnorm(n,mean=0,sd=0.1),each=p)
```

```
x2<-rep(rgamma(n,shape=1)-1,each=p)


 ##the family

  #lambda<-1-(1 + 5*exp(b1+ (-log(t)) + 4*x1))^(-1/5)


##logit

  lambda<-exp(-log(t) + 0.5*x1 + 0.5*x2)/(1+exp(-log(t) +

   0.5*x1 + 0.5*x2))


##GENERATING UNIFORM NUMBERS##

  u <- runif(k)


  ##CREATING DUMMY VARIABLES

  y<-NULL

  D12<-NULL

  D2<-NULL

  D3<-NULL

  D4<-NULL

  D5<-NULL
```

```
D6<-NULL

D7<-NULL

D8<-NULL

D9<-NULL

D10<-NULL

D11<-NULL

for (j in 1:k){ #opening the j loop

  y[j]<-ifelse(lambda[j]> u[j], 1, 0)


  D12[j]<-ifelse(t[j]==12, 1, 0)

  D2[j]<-ifelse(t[j]==2,1,0)

  D3[j]<-ifelse(t[j]==3, 1, 0)

  D4[j]<-ifelse(t[j]==4,1,0)

  D5[j]<-ifelse(t[j]==5, 1, 0)

  D6[j]<-ifelse(t[j]==6,1,0)

  D7[j]<-ifelse(t[j]==7, 1, 0)

  D8[j]<-ifelse(t[j]==8,1,0)

  D9[j]<-ifelse(t[j]==9, 1, 0)

  D10[j]<-ifelse(t[j]==10,1,0)

  D11[j]<-ifelse(t[j]==11, 1, 0)
```

```
}#closing the j loop

DFF<-as.data.frame(cbind(t,x1,x2,lambda,y,D12,D2,D3,D4,D5,D6,D7
,D8,D9,D10,D11))

#creating temporary memory storage
TS<-data.frame(matrix(0,ncol = 16, nrow = 1))

n <- c("t", "x1","x2","lambda","y","D12","D2","D3","D4","D5",
"D6","D7","D8","D9","D10","D11")

colnames(TS) <- n

m<-seq(0,k,p)
for(i in 1: (length(m)-1)){ #opening the i loop

  LL<-m[i]+1
  UL<-m[i+1]
```

```
NDF<-DFF[LL:UL,]


CR<-min(which(NDF$y== 1))



if ( is.infinite(CR)==TRUE) {

  TU<-NDF


} else if ( CR==12) {

  TU<-NDF



} else{


  TU<-NDF[-c((CR+1):12),]

}


TS<-rbind(TS,TU)
```

```
  }#closing the i loop


  TSNW<-TS[-1,]


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


##Fiting logit model
  lgt <- gamlssNP(y ~ D2+D3+D4+D5+D6+D7+D8+D9+D10+D11+D12+x1+x2, random=~1|id,
   family=BI(mu.link="logit"),
                      data=TSNW, K=1, mixture="gq", tol=0.5,
                       control = NP.control(EMcc=0.001, trace=FALSE),
                      g.control = gamlss.control(trace=FALSE))


  summary(lgt)
#print(summary(lgt))


##LGT
cx1[l]<-coef(lgt)[13]
sx1[l]<-coef(lgt)[13]
```

```
cfz[l]<-coef(lgt)[14]

sz[l]<-coef(lgt)[14]


cfd1[l]<-coef(lgt)[1]

cfd12[l]<-coef(lgt)[10]


#period1


HRFx0<-exp(cfd1)/(1+exp(cfd1))

HRFx1<-exp(cfd1+ cx1)/(1+exp(cfd1+ cx1))

H1[l]<-HRFx1/HRFx0

sH1[l]<-HRFx1/HRFx0


##Period9


HRFx000<-exp(cfd1+ cfd12 )/(1+exp(cfd1+ cfd12 ))

HRFx111<-exp(cfd1+cfd12+ cx1 )/(1+exp(cfd1+cfd12+ cx1))

H2[l]<-HRFx111/HRFx000

sH2[l]<-HRFx111/HRFx000
```

```
}#closing the for loop1


cx1<-mean(cx1)

sx1<-sd(sx1)

cfz<-mean(cfz)

sz<-sd(sz)


H1<-mean(H1)

H2<-mean(H2)

sH1<-sd(sH1)

sH2<-sd(sH2)


print(cx1)

print(cfz)

print(sx1)

print(sz)


print(H1)

print(H2)

print(sH1)
```

```
print(sH2)


##Generating data and fitting the logit model with frailty##


rm(list=ls())##cleaning the memory


set.seed(2311988)


nit<-2

cx1<-matrix(,nrow=nit,ncol=1)


cfz<-matrix(,nrow=nit,ncol=1)


cfd1<-matrix(,nrow=nit,ncol=1)


cfd12<-matrix(,nrow=nit,ncol=1)


sx1<-matrix(,nrow=nit,ncol=1)

sz<-matrix(,nrow=nit,ncol=1)

sH1<-matrix(,nrow=nit,ncol=1)
```

```
sH2<-matrix(,nrow=nit,ncol=1)


H1<-matrix(,nrow=nit,ncol=1)

H2<-matrix(,nrow=nit,ncol=1)


for (l in 1:100){

print(l)


library(gamlss.mx)

burr <- function(xi=1)

{

  linkfun <- function(mu) log(((1-mu)^(-xi)-1)/xi)

  linkinv <- function(eta) 1-(1+xi*exp(eta))^(-1/xi)

  mu.eta <- function(eta) exp(eta)*(1+xi*exp(eta))^(-1/xi-1)

  valideta <- function(eta) TRUE

  link <- paste("burr(", xi, ")", sep="")

  structure(list(linkfun = linkfun, linkinv = linkinv,

                 mu.eta = mu.eta, valideta = valideta, name = link),

            class = "link-gamlss")

} #closing burr function
```

```
n<-1500

p<-12

k<-n*p
  id<-rep(1:n,each=p)
  t<-rep(1:p,times=n)

x1<-rep(rnorm(n,mean=0,sd=1.5),each=p)
 b1<-rep(rnorm(n,mean=0,sd=0.1),each=p)

#x2<-rep(rgamma(n,shape=1)-1,each=p)

 ##the family
  #lambda<-1-(1 + 5*exp(b1+ (-log(t)) + 4*x1))^(-1/5)

##logit
  lambda<-exp(b1+ (-log(t)) + 0.5*x1)/(1+exp(b1+ (-log(t)) + 0.5*x1))
```

```
##GENERATING UNIFORM NUMBERS##
  u <- runif(k)

  ##CREATING DUMMY VARIABLES
  y<-NULL
  D12<-NULL
  D2<-NULL
  D3<-NULL
  D4<-NULL
  D5<-NULL
  D6<-NULL
  D7<-NULL
  D8<-NULL
  D9<-NULL
  D10<-NULL
  D11<-NULL
  for (j in 1:k){ #opening the j loop
    y[j]<-ifelse(lambda[j]> u[j], 1, 0)
```

```
  D12[j]<-ifelse(t[j]==12, 1, 0)

  D2[j]<-ifelse(t[j]==2,1,0)

  D3[j]<-ifelse(t[j]==3, 1, 0)

  D4[j]<-ifelse(t[j]==4,1,0)

  D5[j]<-ifelse(t[j]==5, 1, 0)

  D6[j]<-ifelse(t[j]==6,1,0)

  D7[j]<-ifelse(t[j]==7, 1, 0)

  D8[j]<-ifelse(t[j]==8,1,0)

  D9[j]<-ifelse(t[j]==9, 1, 0)

  D10[j]<-ifelse(t[j]==10,1,0)

  D11[j]<-ifelse(t[j]==11, 1, 0)
}#closing the j loop


DFF<-as.data.frame(cbind(t,id,x1,lambda,y,D12,D2,D3,D4,D5,D6,

D7,D8,D9,D10,D11))

#creating temporary memory storage

TS<-data.frame(matrix(0,ncol = 16, nrow = 1))



n <- c("t", "x1","id","lambda","y","D12","D2","D3","D4","D5",
```

```
"D6","D7","D8","D9","D10","D11")


colnames(TS) <- n


m<-seq(0,k,p)
for(i in 1: (length(m)-1)){ #opening the i loop


  LL<-m[i]+1
  UL<-m[i+1]


  NDF<-DFF[LL:UL,]


  CR<-min(which(NDF$y== 1))




  if ( is.infinite(CR)==TRUE) {
    TU<-NDF


  } else if ( CR==12) {
    TU<-NDF
```

```
    } else{

      TU<-NDF[-c((CR+1):12),]
    }


    TS<-rbind(TS,TU)

  }#closing the i loop

  TSNW<-TS[-1,]


#fiting LOGIT#

lgt <- gamlssNP(y ~D2+D3+D4+D5+D6+D7+D8+D9+D10+D11+D12+x1, random=~1|id,
 family=BI(mu.link="logit"),
                            data=TSNW, K=9, mixture="gq", tol=0.5,
                             control = NP.control(EMcc=0.001, trace=FALSE),
```

```
                          g.control = gamlss.control(trace=FALSE))

summary(lgt)

#print(summary(lgt))


##LGT

cx1[l]<-coef(lgt)[13]

sx1[l]<-coef(lgt)[13]

cfz[l]<-coef(lgt)[14]

sz[l]<-coef(lgt)[14]


cfd1[l]<-coef(lgt)[1]

cfd12[l]<-coef(lgt)[10]


HRFx0<-1-(1 + 5*exp(cfd1[l]))^(-1/5)

HRFx1<-1-(1 + 5*exp( cfd1[l] + cx1[l]  ))^(-1/5)

H1[l]<-HRFx1/HRFx0

sH1[l]<-HRFx1/HRFx0


##period10
```

```
HRFx000<-1-(1 + 5*exp(cfd1[l]+ cfd12[l] ))^(-1/5)

HRFx111<-1-(1 + 5*exp( cfd1[l]+ cfd12[l]+ cx1[l] ))^(-1/5)

H2[l]<-HRFx111/HRFx000

sH2[l]<-HRFx111/HRFx000



}#closing the for loop1


cx1<-mean(cx1)

sx1<-sd(sx1)

cfz<-mean(cfz)

sz<-sd(sz)


H1<-mean(H1)

H2<-mean(H2)

sH1<-sd(sH1)

sH2<-sd(sH2)


print(cx1)
```

```
print(cfz)

print(sx1)

print(sz)


print(H1)

print(H2)


print(sH1)

print(sH2)

##Real data code##

rm(list=ls())##cleaning the memory


library(glmx)

library(gamlss.mx)

burr <- function(xi=1)

{

  linkfun <- function(mu) log(((1-mu)^(-xi)-1)/xi)

  linkinv <- function(eta) 1-(1+xi*exp(eta))^(-1/xi)

  mu.eta <- function(eta) exp(eta)*(1+xi*exp(eta))^(-1/xi-1)

  valideta <- function(eta) TRUE
```

```
  link <- paste("burr(", xi, ")", sep="")
  structure(list(linkfun = linkfun, linkinv = linkinv,
                 mu.eta = mu.eta, valideta = valideta, name = link),
            class = "link-gamlss")
}#closing burr function


library(foreign)
library(survival)
library(discSurv)
library(mgcv)
library(gamlss.mx)
library(mgcv)
library(MASS);library(nlme)
library(glmmLasso)
library(Hmisc)
library(nlme)
library(bshazard)
library(sjlabelled)
library(labelled)
```

```
#file.choose()

dataset=read.spss( "G:\\szir51fl.sav",use.value.labels=FALSE

,to.data.frame=TRUE)##, use.missings = to.data.frame

attach(dataset)


#creating cens andtime

#cens<-NULL

#for(i in 1:length(newdataset$AFM)){cens[i]<-ifelse

(is.na(newdataset$AFM[i])==TRUE,0,1)}

#time<-NULL

#for(i in 1:length(newdataset$AFM)){time[i]<-ifelse

(is.na(newdataset$AFM[i])==TRUE,newdataset$CurrentAge[i],newdataset$AFM[i])}

#observation_long<-data.frame(newdataset$CurrentAge,newdataset$AFM,time,cens)


### LOG-RENK TEST

#survdiff(Surv(time,cens)~dataset$YearOfBirth,rho=0)


#creating cens andtime

cens<-NULL

time<-NULL
```

```
Ctime<-NULL



for(i in 1:length(dataset$AFM)){

  cens[i]<-ifelse(is.na(dataset$AFM[i])==TRUE,0,1)
  time[i]<-ifelse(is.na(dataset$AFM[i])==TRUE,dataset$CurrentAge[i],
  dataset$AFM[i])

  if ((time[i]>=15 & time[i]<17)){
    Ctime[i]<-1

  }else if((time[i]>=17 & time[i]<19)){
    Ctime[i]<-2



  }else if((time[i]>=19 & time[i]<21)){
    Ctime[i]<-3
```

```
}else if((time[i]>=21 & time[i]<23)){

  Ctime[i]<-4



}else if((time[i]>=23 & time[i]<25)){

  Ctime[i]<-5



}else if((time[i]>=25 & time[i]<27)){

  Ctime[i]<-6



} else if((time[i]>=27 & time[i]<29)){

  Ctime[i]<-7


} else if((time[i]>=29 & time[i]<31)){

  Ctime[i]<-8


} else if((time[i]>=31 & time[i]<33)){

  Ctime[i]<-9
```

```
} else if((time[i]>=33 & time[i]<35)){
  Ctime[i]<-10

} else if((time[i]>=35 & time[i]<37)){
  Ctime[i]<-11
} else if((time[i]>=37 & time[i]<39)){
  Ctime[i]<-12
} else if((time[i]>=39 & time[i]<41)){
  Ctime[i]<-13

} else if((time[i]>=41 & time[i]<44)){
  Ctime[i]<-14
} else if((time[i]>=44 & time[i]<45)){
  Ctime[i]<-15
} else if((time[i]>=45 & time[i]<47)){
  Ctime[i]<-16

}else{
  Ctime[i] <-17
```

```
  }

  }

NYROFBRTH <- cut(YearOfBirth, g=5,seq(1956,2000,8))
table(NYROFBRTH )
COHORT<-factor(NYROFBRTH)
print(COHORT)

newdataset<-as.data.frame(cbind(cens,time,AFM,Ctime,
Residence,Region,Edu,COHORT))
attach(newdataset)
##############################################################################
##Ploting life table and extracting the hazard graph

LFTB <- lifeTable (dataSet=newdataset, timeColumn="time", censsColumn="cens")
head(LFTB$Output, 50)
plot(x=1:dim(LFTB$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])



#Region
ts0 <- survfit( Surv(time, 1-cens)~ 1, conf.type="none",
subset=(Region==1), data=newdataset)
ts1 <- survfit( Surv(time, 1-cens)~ 1, conf.type="none",
subset=(Region==2), data=newdataset)
ts2 <- survfit( Surv(time, 1-cens)~ 1, conf.type="none",
subset=(Region==3), data=newdataset)
ts3 <- survfit( Surv(time, 1-cens)~ 1, conf.type="none",
subset=(Region==4), data=newdataset)


#ploting survivor function


plot(ts0$time, ts0$surv, type="l", ylab="Estimated Survival
Function", xlab="time",
     ylim=c(0.0, 1.0), xlim=c(1, 49), col=c(1))
```

```
par(new=T)
plot(ts1$time, ts1$surv, type="l", ylab=" ", ylim=c(0.0, 1.0), xlim=c(1, 49)
, xlab="", col=c(2))
abline(lty=2)
par(new=T)
plot(ts2$time, ts2$surv, type="l", ylab=" ", ylim=c(0.0, 1.0), xlim=c(1, 49)
, xlab="", col=c(3))
abline(lty=2)
par(new=T)
plot(ts3$time, ts3$surv, type="l", ylab=" ", ylim=c(0.0, 1.0), xlim=c(1, 49)
, xlab="", col=(4))
abline(lty=2)
par(new=T)
legend("topright",c("Hhohho","Manzini","Shiselweni","Lubombo"),lty=1,
col=c(1,2,3,4))


 #place of residence
 ts0 <- survfit( Surv(time, 1-cens)~ 1, conf.type="none",
 subset=(Residence==1), data=newdataset)
  ts1 <- survfit( Surv(time, 1-cens)~ 1, conf.type="none",
```

```
   subset=(Residence==2), data=newdataset)


  plot(ts0$time, ts0$surv, type="l", ylab="Estimated Survival
  Function", xlab="time",
  ylim=c(0.0, 1.0), xlim=c(1, 49), col=c(1))
  par(new=T)
   plot(ts1$time, ts1$surv, type="l", ylab=" ", ylim=c(0.0, 1.0),
    xlim=c(1, 49), xlab="", col=c(2))
    abline(lty=2)
    par(new=T)
    legend("bottomleft",col=c(1,2),c("Urban","Rural"),lty=c(1,1))
#Edu
ts0 <- survfit( Surv(time, 1-cens)~ 1, conf.type="none",
 subset=(Edu==1), data=newdataset)
ts1 <- survfit( Surv(time, 1-cens)~ 1, conf.type="none",
 subset=(Edu==2), data=newdataset)
ts2 <- survfit( Surv(time, 1-cens)~ 1, conf.type="none",
 subset=(Edu==3), data=newdataset)
ts3 <- survfit( Surv(time, 1-cens)~ 1, conf.type="none",
subset=(Edu==4), data=newdataset)
```

```
plot(ts0$time, ts0$surv, type="l", ylab="Estimated
 Survival Function", xlab="time",
     ylim=c(0.0, 1.0), xlim=c(1, 49), col=c(1))
par(new=T)
plot(ts1$time, ts1$surv, type="l", ylab=" ", ylim=c(0.0, 1.0),
 xlim=c(1, 49), xlab="", col=c(2))
abline(lty=2)
par(new=T)
plot(ts2$time, ts2$surv, type="l", ylab=" ", ylim=c(0.0, 1.0),
 xlim=c(1, 49), xlab="", col=c(3))
abline(lty=2)
par(new=T)
plot(ts3$time, ts3$surv, type="l", ylab=" ", ylim=c(0.0, 1.0),
 xlim=c(1, 49), xlab="", col=c(4))
abline(lty=2)
par(new=T)
legend("bottomleft",col=c(1,2,3,4),c("No education",
"Primary","Secondary","Higher"),lty=c(1,1))
```

```
#Cohort
ts0 <- survfit( Surv(time, 1-cens)~ 1, conf.type="none",
 subset=(COHORT=="(1956,1964]"))

ts1 <- survfit( Surv(time, 1-cens)~ 1, conf.type="none",
 subset=(COHORT=="(1964,1972]"))

ts2 <- survfit( Surv(time, 1-cens)~ 1, conf.type="none",
 subset=(COHORT=="(1972,1980]"))

ts3 <- survfit( Surv(time, 1-cens)~ 1, conf.type="none",
 subset=(COHORT=="(1980,1988]"))

ts4 <- survfit( Surv(time, 1-cens)~ 1, conf.type="none",
subset=(COHORT=="(1988,1996]"))


plot(ts0$time, ts0$surv, type="l", ylab="Estimated Survival
```

```
Function", xlab="time",
     ylim=c(0, 1.0), xlim=c(10, 49), col=c(1))
par(new=T)
plot(ts1$time, ts1$surv, type="l", ylab=" ", ylim=c(0, 1.0),
 xlim=c(10, 49), xlab="", col=c(2))
abline(lty=2)
par(new=T)
plot(ts2$time, ts2$surv, type="l", ylab=" ", ylim=c(0, 1.0),
 xlim=c(10, 49), xlab="", col=c(3))
abline(lty=2)
par(new=T)
plot(ts3$time, ts3$surv, type="l", ylab=" ", ylim=c(0, 1.0),
 xlim=c(10, 49), xlab="", col=c(4))
abline(lty=2)
par(new=T)
plot(ts4$time, ts4$surv, type="l", ylab=" ", ylim=c(0, 1.0),
 xlim=c(10, 49), xlab="", col=c(5))
abline(lty=2)
par(new=T)
legend("bottomleft",col=c(1,2,3,4,5),c("(1956,1964]","(1964,1972]",
```

```
"(1972,1980]","(1980,1988]","(1988,1996]"),lty=c(1,1))


##############################################################################
##converting data to long fomat
#is.na(NYROFBRTH)
NRES <- factor(newdataset$Residence,levels=c(1:2),
labels=c("Rural","Urban"))
NREG <- factor(newdataset$Region,levels=c(1:4),
labels=c("Hhohho","Manzini","Shiselweni","Lubombo"))
NEDU <- factor(newdataset$Edu,levels=c(0:3),
labels=c("No education","Primary","Secondary","Higher"))
print(NEDU)


newdataset<-as.data.frame(cbind(cens,time,Ctime,NRES,
NREG,NEDU,COHORT,Family))



dataLongpo <- as.data.frame(dataLong(dataSet=newdataset,
 timeColumn="Ctime", censColumn="cens"))
dataLongpo<-na.omit(dataLongpo)
```

```
#print(head(datLongpo))

#sum(any(is.na(datLongpo)))

#datLongpo<-datLongpo[-c(5)]

#print(head(datLongpo,100))

#na_vec<-which(!complete.cases(datLongpo))

#datLongpo_no_na<-datLongpo[-na_vec,]


#is.na(NYROFBRTH)

dataLongpo$NRES <- factor(dataLongpo$NRES,levels=c(1:2),

labels=c("Rural","Urban"))

dataLongpo$NREG <- factor(dataLongpo$NREG,levels=c(1:4),

labels=c("Hhohho","Manzini","Shiselweni","Lubombo"))

dataLongpo$NEDU <- factor(dataLongpo$NEDU,levels=c(1:4),

labels=c("No education","Primary","Secondary","Higher"))

dataLongpo$COHORT<-factor(dataLongpo$COHORT,levels=c(1:5),

labels=c("(1956,1964]","(1964,1972]","(1972,1980]",

"(1980,1988]","(1988,1996]"))
```

```
#fiting logit model


lgt <- gamlssNP(y ~ as.factor(timeInt)+as.factor(NRES)+

as.factor(NREG)+as.factor(NEDU)

+as.factor(COHORT),random=NULL, family=BI(mu.link=logit),

                data=dataLongpo, K=1, mixture="gq", tol=0.5,

                control = NP.control(EMcc=0.001, trace=FALSE),

                g.control = gamlss.control(trace=FALSE))

summary(lgt)


#predictions from the model


EF<-coef(lgt)

sEF<-EF[1:17]

sEF[2:17]<-sEF[2:17]+sEF[1]

hEF<-1/(1+exp(-(sEF)))


plot(c(1:17),hEF, ylim=c(0, 0.3))


##Placeof RES Hazard GGPLOT2
```

```
Testdata<-data.frame(timeInt=c(1:17),NRES="Urban",NREG="Hhohho",
NEDU ="No education",COHORT="(1956,1964]")
Testdata1<-data.frame(timeInt=c(1:17),NRES="Rural",NREG="Hhohho",
NEDU ="No education",COHORT="(1956,1964]")
print(Testdata1)

LogitHazardRate<-predict(lgt, newdata=Testdata, type = "response",
data=dataLongpo,se.fit=TRUE)
print(Urban)
plot(ap)
Rural<-predict(lgt, newdata=Testdata1, type = "response",
data=dataLongpo,se.fit=TRUE)
print(Rural)

predicts<-data.frame(Testdata,LogitHazardRate,Rural,hEF)
#predicts1<-data.frame(Testdata1,Rural,hEF)

##GGPLOT FOF RES
ggplot(predicts, aes(timeInt)) +
  geom_line(aes(y =LogitHazardRate, colour = "Urban")) +
```

```
  geom_line(aes(y = Rural, colour = "Rural"))+xlim(1,16)+ylim(0,.25)


#p<-ggplot(predicts, aes(x=timeInt,y=ap,)) + geom_line()


#p<-p+geom_line(predicts1,aes(x=timeInt,y=ap1))
#p
#ggplot(predicts,aes(x=as.numeric(timeInt), y=fit))+
geom_line(aes(color="Urban"))+
  #geom_line(data=predicts1,aes(color="Rural"))+
  #labs(color="Legend text") +xlab("time interval (Age)") +
  #ylab("Estimated logit Hazard Rate")


##Region Hazard GGPLOT2
Testdata<-data.frame(timeInt=c(1:17),NREG="Hhohho",NRES="Rural",
NEDU ="No education",COHORT="(1956,1964]")
Testdata1<-data.frame(timeInt=c(1:17),NREG="Manzini",NRES="Rural",
NEDU ="No education",COHORT="(1956,1964]")
Testdata2<-data.frame(timeInt=c(1:17),NREG="Shiselweni",NRES="Rural",
NEDU ="No education",COHORT="(1956,1964]")
Testdata3<-data.frame(timeInt=c(1:17),NREG="Lubombo",NRES="Rural",
```

```
NEDU ="No education",COHORT="(1956,1964]")


LogitHazardRate<-predict(lgt, newdata=Testdata,
 type = "response",data=dataLongpo,se.fit=FALSE)
Manzini<-predict(lgt, newdata=Testdata1, type = "response"
,data=dataLongpo,se.fit=FALSE)
Shiselweni<-predict(lgt, newdata=Testdata2, type = "response",
data=dataLongpo,se.fit=FALSE)
Lubombo<-predict(lgt, newdata=Testdata3, type = "response",
data=dataLongpo,se.fit=FALSE)


predicts<-data.frame(Testdata,LogitHazardRate,Manzini,
Shiselweni,Lubombo,hEF)
#predicts1<-data.frame(Testdata1,Rural,hEF)


##GGPLOT FOF REG
ggplot(predicts, aes(timeInt)) + geom_line
(aes(y = LogitHazardRate, colour = "Hhohho")) +
  geom_line(aes(y = Manzini, colour = "Manzini"))+geom_line
  (aes(y = Shiselweni, colour = "Shiselweni"))+geom_line
```

```
  (aes(y = Lubombo, colour = "Lubombo"))+xlim(1,16)+ylim(0,.25)


##Edu Hazard GGPLOT2

Testdat<-data.frame(timeInt=c(1:17),NEDU="No education",

NREG="Hhohho",NRES="Rural",COHORT="(1956,1964]")

Testdat1<-data.frame(timeInt=c(1:17),NEDU="Primary",NREG="Hhohho",

NRES="Rural",COHORT="(1956,1964]")

Testdat2<-data.frame(timeInt=c(1:17),NEDU="Secondary",NREG="Hhohho",

NRES="Rural",COHORT="(1956,1964]")

Testdat3<-data.frame(timeInt=c(1:17),NEDU="Higher",NREG="Hhohho",

NRES="Rural",COHORT="(1956,1964]")


LogitHazardRate<-predict(lgt, newdata=Testdat,

 type = "response",data=dataLongpo,se.fit=FALSE)

Primary<-predict(lgt, newdata=Testdat1, type = "response",

data=dataLongpo,se.fit=FALSE)

Secondary<-predict(lgt, newdata=Testdat2, type = "response",

data=dataLongpo,se.fit=FALSE)

Higher<-predict(lgt, newdata=Testdat3, type = "response",

data=dataLongpo,se.fit=FALSE)
```

```
predicts<-data.frame(Testdata,LogitHazardRate,Primary,Secondary,Higher,hEF)
#predicts1<-data.frame(Testdata1,Rural,hEF)


##GGPLOT FOF EDU
ggplot(predicts, aes(timeInt)) + geom_line(aes
(y = LogitHazardRate, colour = "No education")) +
  geom_line(aes(y = Primary, colour = "Primary"))+geom_line
  (aes(y = Secondary, colour = "Secondary"))+geom_line(
  aes(y = Higher, colour = "Higher"))+xlim(1,16)+ylim(0,.25)


##cohort Hazard GGPLOT2
Testdat<-data.frame(timeInt=c(1:17),COHORT="(1956,1964]",
NEDU="No education",NREG="Hhohho",NRES="Rural")
Testdat1<-data.frame(timeInt=c(1:17),COHORT="(1964,1972]",
NEDU="No education",NREG="Hhohho",NRES="Rural")
Testdat2<-data.frame(timeInt=c(1:17),COHORT="(1972,1980]",
NEDU="No education",NREG="Hhohho",NRES="Rural")
Testdat3<-data.frame(timeInt=c(1:17),COHORT="(1980,1988]",
NEDU="No education",NREG="Hhohho",NRES="Rural")
```

```
Testdat4<-data.frame(timeInt=c(1:17),COHORT="(1988,1996]",
NEDU="No education",NREG="Hhohho",NRES="Rural")


LogitHazardRate<-predict(lgt, newdata=Testdat,
 type = "response",data=dataLongpo,se.fit=FALSE)
SixtyFourtoSeventyTwo<-predict(lgt, newdata=Testdat1,
 type = "response",data=dataLongpo,se.fit=FALSE)
SeventyTwotoeighty<-predict(lgt, newdata=Testdat2,
 type = "response",data=dataLongpo,se.fit=FALSE)
eightyToEightyEight<-predict(lgt, newdata=Testdat3,
type = "response",data=dataLongpo,se.fit=FALSE)
EightyEighttoNinentySix<-predict(lgt, newdata=Testdat4,
 type = "response",data=dataLongpo,se.fit=FALSE)


predicts<-data.frame(Testdata,LogitHazardRate,
SixtyFourtoSeventyTwo,
SeventyTwotoeighty,eightyToEightyEight,EightyEighttoNinentySix,hEF)
#predicts1<-data.frame(Testdata1,Rural,hEF)


##GGPLOT of COHORT
```

```
ggplot(predicts, aes(timeInt)) + geom_line(aes(y = LogitHazardRate,
 colour = "(1956,1964]")) +
  geom_line(aes(y = SixtyFourtoSeventyTwo, colour = "(1964,1972]"))+
  geom_line(aes(y = SeventyTwotoeighty, colour = "(1972,1980]"))+g
  eom_line(aes(y =eightyToEightyEight, colour = "(1980,1988]"))+
  geom_line(aes(y = SeventyTwotoeighty, colour = "(1972,1980]"))+
  geom_line(aes(y =EightyEighttoNinentySix, colour =
  "(1988,1996]"))+xlim(1,16)+ylim(0,.20)
##############################################################################


##Fitting burr e=e


bg<-seq(10,20,1)


PAIC<-NULL
PSBC<-NULL
PGDC<-NULL


for (k in 1:length(bg)){ #opening the loop
  print(k)
```

```
  xi<-bg[k]
  #print(xi)
  fml1<- gamlssNP(y ~ as.factor(timeInt)+as.factor(NRES)+as.factor(NREG)+
  as.factor(NEDU)+as.factor(COHORT), random=~NULL, family=BI(burr(xi=10.5)),
  data=dataLongpo, K=1, mixture="gq", tol=0.5, control = NP.control
                  (EMcc=0.001, trace=FALSE),
  g.control = gamlss.control(trace=FALSE))


  summary(fml1)

  PAIC[k]<-fml1$aic
  PSBC[k]<-fml1$sbc

  PGDC[k]<-fml1$G.deviance

}#closing the loop
par(mfrow=c(2,2))
plot(bg,PAIC)
```

```
plot(bg,PSBC)

#plot(bg,PGDC)


print(PAIC)

print(PSBC)

print(PGDC)


i1<-which.min(PAIC)

#print(i1)

i2<-which.min(PSBC)

#print(i2)

i3<-which.min(PGDC)

#print(i3)



print(c(i1,i2,i3))


#bg<-seq(0.5,10,0.5)

xi<-bg[i1]

print(xi)
```

```
fml <- gamlssNP(y ~ pb(as.numeric(timeInt))+
as.factor(NRES)+as.factor(NREG)+
as.factor(NEDU)+as.factor(COHORT), random=~1|Family,
 family=BI(burr(xi=10.5)),
                data=dataLongpo, K=9, mixture="gq", tol=0.5,
                control = NP.control(EMcc=0.001, trace=FALSE),
                g.control = gamlss.control(trace=FALSE))

summary(fml)

#predictions fom the model

EF<-coef(fml)
sEF<-EF[1:17]
sEF[2:17]<-sEF[2:17]+sEF[1]
hEF<-1/(1+exp(-(sEF)))

plot(c(1:17),hEF, ylim=c(0, 0.3))
```

```
#rESIDENCE

LR<-NULL

LU<-NULL


for(i in 2:17){


  p<-summary(fml)[1,1]

  u<-summary(fml)[18,1]



  LR[i]<-1-(1 + xi*exp(p+summary(fml)[i,1]))^(-1/xi)


  LU[i]<-1-(1 +xi*exp(p+summary(fml)[i,1]+u ))^(-1/xi)

}

print(LR)

print(LU)


#RES

Testdata<-data.frame(timeInt=c(1:17),
```

```
NRES="Urban",NREG="Hhohho",
NEDU ="No education",COHORT="(1956,1964]")
Testdata1<-data.frame(timeInt=c(1:17),
NRES="Rural",NREG="Hhohho",
NEDU ="No education",COHORT="(1956,1964]")
print(Testdata1)


predicts<-data.frame(Testdata,LR,LU,hEF)
#predicts1<-data.frame(Testdata1,Rural,hEF)


##GGPLOT FOF RES
ggplot(predicts, aes(timeInt)) +
  geom_line(aes(y =LU, colour = "Urban")) +
  geom_line(aes(y = LR, colour = "Rural"))+
  xlim(1,16)+ylim(0,.25)


#Region


ParetoHazardRate<-NULL
LM<-NULL
```

```
LS<-NULL

LL<-NULL

for(i in 2:17){


  p<-summary(fml)[1,1]

  M<-summary(fml)[19,1]

  S<-summary(fml)[20,1]

  L<-summary(fml)[21,1]


  ParetoHazardRate[i]<-1-(1 + xi*exp

  (p+summary(fml)[i,1]))^(-1/xi)


  LM[i]<-1-(1 +xi*exp(p+summary(fml)[i,1]+M ))^(-1/xi)

  LS[i]<-1-(1 +xi*exp(p+summary(fml)[i,1]+S ))^(-1/xi)

  LL[i]<-1-(1 +xi*exp(p+summary(fml)[i,1]+L ))^(-1/xi)

}

print(ParetoHazardRate)

print(LM)

print(LS)

print(LL)
```

```
predicts2<-data.frame(Testdata,ParetoHazardRate,LM,LS,LL,hEF)

#predicts1<-data.frame(Testdata1,Rural,hEF)


##GGPLOT FOF REG

ggplot(predicts2, aes(timeInt)) + geom_line(aes(y = ParetoHazardRate,

 colour = "Hhohho")) +

  geom_line(aes(y = LM, colour = "Manzini"))+geom_line(aes(y = LS,

   colour = "Shiselweni"))+geom_line(aes(y = LL, colour = "Lubombo"))

   +xlim(1,16)+ylim(0,.25)


#EDUCATION


ParetoHazardRate<-NULL

LP<-NULL

LSEC<-NULL

LHI<-NULL


for(i in 2:17){


  p<-summary(fml)[1,1]
```

```
 PR<-summary(fml)[22,1]
  SE<-summary(fml)[23,1]
  HI<-summary(fml)[24,1]

  ParetoHazardRate[i]<-1-(1 + xi*exp(p+summary(fml)[i,1]))^(-1/xi)

  LP[i]<-1-(1 +xi*exp(p+summary(fml)[i,1]+PR))^(-1/xi)
  LSEC[i]<-1-(1 +xi*exp(p+summary(fml)[i,1]+SE ))^(-1/xi)
  LHI[i]<-1-(1 +xi*exp(p+summary(fml)[i,1]+HI ))^(-1/xi)
}
print(ParetoHazardRate)
print(LP)
print(LSEC)
print(LHI)
predicts2<-data.frame(Testdata,ParetoHazardRate,LP,LSEC,LHI,hEF)
#predicts1<-data.frame(Testdata1,Rural,hEF)

##GGPLOT FOF edu
ggplot(predicts2, aes(timeInt)) + geom_line(aes(y
= ParetoHazardRate, colour = "Secondary")) +
```

```
  geom_line(aes(y = LP, colour = "Higher"))
  +geom_line(aes(y = LSEC,
   colour = "Primary"))+geom_line(aes(y = LHI,
   colour = "No education"))+xlim(1,16)+ylim(0,.25)


ParetoHazardRate<-NULL
SEV2<-NULL
EIGHTY<-NULL
EIGTYEIGT<-NULL
NINESX<-NULL
for(i in 2:17){
  p<-summary(fml)[1,1]
  Y<-summary(fml)[25,1]
YY<-summary(fml)[26,1]
O<-summary(fml)[27,1]
OO<-summary(fml)[28,1]
ParetoHazardRate[i]<-1-(1 + xi*exp(p+summary(fml)[i,1]))^(-1/xi)
SEV2[i]<-1-(1 +xi*exp(p+summary(fml)[i,1]+Y))^(-1/xi)
EIGHTY[i]<-1-(1 +xi*exp(p+summary(fml)[i,1]+YY ))^(-1/xi)
EIGTYEIGT[i]<-1-(1 +xi*exp(p+summary(fml)[i,1]+O ))^(-1/xi)
```

```
NINESX[i]<-1-(1 +xi*exp(p+summary(fml)[i,1]+OO ))^(-1/xi)

}

print(ParetoHazardRate)

print(SEV2)

print(EIGHTY)

print(EIGTYEIGT)

print(NINESX)


predicts3<-data.frame(Testdata,ParetoHazardRate,

SEV2,EIGHTY,EIGTYEIGT,NINESX,hEF)


##GGPLOT of COHORT

ggplot(predicts3, aes(timeInt)) + geom_line(aes(y = ParetoHazardRate,

 colour = "(1956,1964]",linetype = "dashed")) + geom_path(aes(y = SEV2,

  colour = "(1964,1972]"),linetype = "twodash")+geom_line(aes(y = EIGHTY,

   colour = "(1972,1980]",linetype = "solid"))+ geom_line(aes(y =EIGTYEIGT,

   colour = "(1980,1988]",linetype="dotted"))+geom_line(aes(y =NINESX,

   colour = "(1988,1996]",linetype="longdash"))+xlim(1,16)+ylim(0,.25)


#fiting cloglog model
```

```
clglg <- gamlssNP(y ~ as.factor(timeInt)+as.factor(NRES)+as.factor(NREG)+
as.factor(NEDU)+as.factor(COHORT), random=NULL, family=BI(mu.link=cloglog),
                  data=dataLongpo, K=1, mixture="gq", tol=0.5,
                   control = NP.control(EMcc=0.001, trace=FALSE),
                  g.control = gamlss.control(trace=FALSE))



summary(clglg)


#fiting probit model


prbt <- gamlssNP(y ~ as.factor(timeInt)+as.factor(NRES)
+as.factor(NREG)+as.factor(NEDU)
+as.factor(COHORT), random=NULL, family=BI(mu.link=probit),
                  data=dataLongpo, K=1, mixture="gq", tol=0.5,
                  control = NP.control(EMcc=0.001, trace=FALSE),
                  g.control = gamlss.control(trace=FALSE))


summary(prbt)
```

```
library(glmx)

gossbin <- function(nu) binomial(link = gosset(nu))

m0 <- glmx(y ~ as.factor(timeInt)+as.factor

(NRES)+as.factor(NREG)+as.factor(NEDU)

+factor(NYROFBRTH), data = dataLongpo,

           family = gossbin, xstart = 0, xlink = "log")

summary(m0)
```