MODELLING CHILDREN UNDER FIVE MORTALITY IN SOUTH AFRICA USING COPULA AND FRAILTY SURVIVAL MODELS

by

TSHILIDZI BENEDICTA MULAUDZI

THESIS

Submitted in fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

STATISTICS

in the

FACULTY OF SCIENCE AND AGRICULTURE
(School of Mathematical and Computer Sciences)

at the

UNIVERSITY OF LIMPOPO

PROMOTER: Prof. Yehenew G. Kifle (University of Maryland Baltimore County, USA)

CO-PROMOTERS: Prof. Roel Braekers (University of Hasselt, Belgium)
Prof. M Lesaoana

2022

Declaration

I, **Tshilidzi Benedicta Mulaudzi**, declare that the thesis hereby submitted to the University of Limpopo, for the degree of Doctor of Philosophy in Statistics has not previously been submitted by me for a degree at this or any other university; that it is my work in design and in execusion, and that all material contained herein has been duly acknowledged.

....

Mulaudzi,T.B **Date**

Abstract

This thesis is based on application of frailty and copula models to under five child mortality data set in South Africa. The main purpose of the study was to apply sample splitting techniques in a survival analysis setting and compare clustered survival models considering left truncation to the under five child mortality data set in South Africa. The major contributions of this thesis is in the application of the shared frailty model and a class of Archimedean copulas in particular, Clayton-Oakes copula with completely monotone generator, and introduction of sample splitting techniques in a survival analysis setting.

The findings based on shared frailty model show that clustering effect was significant for modelling the determinants of time to death of under five children, and revealed the importance of accounting for clustering effect. The conclusion based on Clayton-Oakes model showed association between survival times of children from the same mother. It was found that the parameter estimates for the shared frailty and the Clayton-Oakes models were quite different and that the two models cannot be comparable. Gender, province, year, birth order and whether a child is part of twin or not were found to be significant factors affecting under five child mortality in South Africa.

Keywords: Frailty models, Archimedean copula, left truncation, penalised likelihood, clustered survival models.

Dedication To

My husband Ramudzuli Mulaudzi

My mother Phyllis Netshivhuyu and my siblings Rendani and Thabelo

My children :
Asakundwi, Arehone and Uafulufhedzea

Acknowledgments

Completion of this doctoral thesis was possible with the support of many people. I take this opportunity to extend my sincere gratitude and appreciation to all those who directly and indirectly made this thesis possible.

First of all, I greatly appreciate my supervisors Prof. Yehenew G. Kifle, Prof. Roel Braekers and Prof. Maseka Lesaoana for their guidance, intellectual support and encouragement throughout my PhD studies.

A sincere word of gratitude is extended to Statistics South Africa (Stats SA) for providing data used for this thesis. I would like to thank NRF-TDG reference number APP-TDG-059 for funding my trip to University of Hasselt (Belgium) to work with my supervisor, Prof. Roel Braekers.

The financial assistance of the Flemish interuniversity council, Institutional University corporation (VLIR-IUC) VLIR-IUC programme of the University of Limpopo towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at are those of the author and not necessarily be attributed to the VLIR-IUC programme.

The thesis would not have come to a successful completion without the help and motivation from my colleagues in the department of Statistics at University of Venda and the assistance of Mr. Nqobile Muleya (University of Venda) and Mr. Akalu Banbeta (Jimma University, Ethiopia).

To my wonderful husband Mr. Ramudzuli Mulaudzi and my lovely children Asakundwi, Arehone and Uafulufhedzea, I thank you for your patience, support and encouraging me throughout the writing of this thesis. I would like to thank my mother Mrs. Phyllis Netshivhuyu, and my siblings for encouraging me to continue until the end of the journey.

Above all, I owe it all to Almighty God for giving me strength and wisdom to undertake this research task until its completion.

Thank You All

Contents

List of Symbols

β	Column vector of regression coefficients
δ	Censoring indicator
f	Generic symbol for a probability density function
F	Generic symbol for a cumulative density function
$\Gamma(.)$	Gamma function
κ	Smoothing parameter
$\mathcal{L}$	Laplace transform
$h(t)$	Baseline hazard function
$H(t)$	Cumulative hazard function
$\hat{H}_l$	Converged Hessian matrix
L_{ij}	Left truncated times
$pl(\hat{\Phi}_k)$	Penalized log-likelihood
τ	Kendall's tau
T	Survival time
ρ	Pearson correlation coefficient
θ	Association parameter
S_e	Standard error
$S(.)$	Survival function
λ_L	Lower tail dependence
λ_U	Upper tail dependence
Q	Number of knots

List of Abbreviations

AIC	Akaike Information criterion
AIDS	Acquired Immune Deficiency Syndrome
C	Copula
E-step	Expectation step
EM	Expectation-Maximisation
GLM	Generalised Linear Model
HR	Hazard Ratio
K-M	Kaplan-Meier
LCV	Likelihood Cross Validation
LRT	Likelihood Ratio Test
MDG	Millenium development goal
MLE	Maximum likelihood estimator
M-step	Maximisation step
OR	Odds Ratio
PH	Proportional Hazard
RSA	Republic of South Africa
SDG	Sustainable Development Goal
Stats SA	Statistics South Africa
U5CM	Under five child mortality

Research Outputs

The following section gives a list of workshops and events attended related to this thesis.

Workshops and related events

1. NRF-TDG (reference no: APP-TDG-059) funded a trip to Belgium to interact and collaborate with international experts, 25 November - 17 December 2017, hosted by Hasselt University.

2. The Flemish interuniversity council, Institutional University corporation (VLIR-IUC) funded two trips to Belgium to interact and collaborate with international experts, 01 February - 30 April 2019 and 01 September - 30 October 2019, hosted by Hasselt University.

3. Attended an article writing workshop, hosted by University of Venda at 2 ten hotel, South Africa, 22-26 March 2021. During this workshop, the following draft paper was produced:

   Mulaudzi, TB., Braekers, R., Kifle YG. and Lesaoana, M. (2021).
   Modelling under five mortality: application of shared frailty model to left truncated data.

Chapter 1

Introduction and background

1.1 Introduction

Under five child mortality is still a problem in Sub-Saharan Africa. The probability of the death of children under the age of five in Sub-Saharan Africa is more than 14 times the probability of the death of children in developed regions (Munyamahoro, 2016). One of the millennium developmental goals (MDG-4) was to decrease under five child mortality cases by two-thirds between 1990-2015 (Bryce et al., 2006). The aim of the Proposed Sustainable Development Goal (SDG) is to end under five child mortality and deaths of new born babies in 2030. The whole world is aiming to decrease deaths of babies born within the first 28 days of their life by at least 12 per 1000 live births and also to decrease mortality rate of children under the age of five by at least 25 per 1000 live births (UNDP, 2019).

South Africa is one of the countries in Sub-Saharan Africa which is also ready to decrease under five mortality rates in agreement with the SDG targets. Pol-

icy makers and health officials require information about causes of child deaths so that they can keep a track of child health and service delivery (Bamford et al., 2018). The above-mentioned concern is one of the reasons why under five mortality has attracted many researchers in order to identify causes of high deaths of under five children. It is important to use correct statistical methods to determine factors that are strongly associated with child mortality to come up with intervention strategies (Munyamahoro, 2016). Children belonging to the same family share the same environment and are often exposed to the same conditions in terms of parental care. They also share the same genes and socio-economic position. One or more covariates shared by children of the same mother will induce a correlation between their mortality risks. The risks of mortality of siblings are related and therefore standard estimation procedures can produce faulty results (Cesar et al., 1997).

This thesis investigates clustered survival models for analysing time until the death of children under five years of age born in South Africa. Often, the time to event is right-censored, which occurs when an individual leaves the study before an event happens, for example due to drop-out. The time to event can also be left truncated for some individuals in the study. This happens when an individual is not observed if the event happens before a certain period or date.

A survival study can involve grouped data such as children from the same mother as it applies to our study. Since grouped study items share common traits, their event times show within-cluster correlation. Popular survival models that account for association in grouped survival data are the frailty and the copula models. The difference between these models is that a frailty model is a hazard model with a cluster-specific random term called frailty. A copula model describes the joint survival function of the survival times using the marginal survival functions and a dependence function called copula. Both survival mod-

els provide measures for the strength of association between event times next to the estimated covariates effects. This thesis has considered both right censoring and left truncation to explore shared frailty and copula models.

Further, in many instances, researchers come across large data sets that are difficult to analyse at once. This thesis proposes a sample splitting technique in a survival analysis setting that can partition large data sets into sub-samples, analyse each sub-sample separately and properly combine estimates into one.

1.2 Background of the study area

South Africa officially known as the Republic of South Africa (RSA) is one of the countries in Sub-Saharan Africa with a population of about 59 million people and a land area of about 1220813 squares kilometres (Mabin et al., 2021). Its neighbouring countries are Zimbabwe, Botswana, Namibia, Mozambique and Lesotho. South Africa has nine provinces namely: Western Cape, Eastern Cape, Northern Cape, North West, Free State, KwaZulu Natal, Gauteng, Limpopo and Mpumalanga. The map of South Africa showing the location of all nine provinces is given in Figure 1.1.

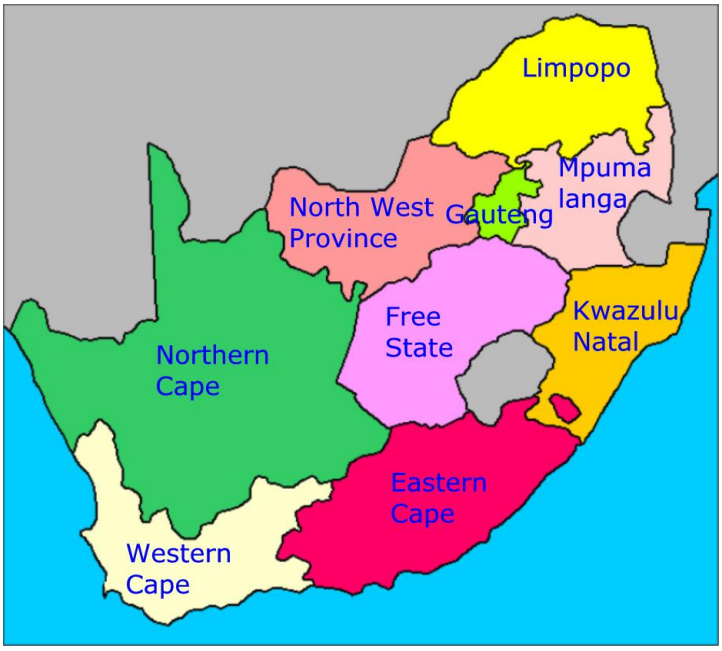

Figure 1.1: The map of South Africa with its provinces

Table 1.1 gives the 2019 mid-year provincial population estimates by Statistics South Africa (Stats SA).

Table 1.1: 2019 mid-year provincial population estimates

Rank	Province	Population	Percentage
1	Gauteng	15 176 115	25.8%
2	KwaZulu Natal	11 289 086	19.2%
3	Western Cape	6 844 272	11.6%
4	Eastern Cape	6 712 276	11.4%
5	Limpopo	5 982 584	10.2%
6	Mpumalanga	4 592 187	7.8%
7	North West	4 027 160	6.9%
8	Free State	2 887 465	4.9%
9	Northern Cape	1 263 875	2.2%
	Total	58 775 022	100%

Statistics South Africa

Table 1.1 shows the estimated percentages in 2019 of the entire population in each of the nine provinces of South Africa (SA, 2019). Gauteng province has the largest portion of population with about 15.2 million (25.8%) people, followed by KwaZulu Natal with close to 11.3 million (19.2%) people. Thus, nearly half of

South Africa's population reside in these two provinces. Northern Cape is the province with the smallest portion of the population with about 1.26 million (2.2%) people.

1.3 Statement of the problem

In most cases survival data are in groups of clusters such as children from the same mother, husband and wife, society and geographical divisions. Observations taken from subjects belonging to the same cluster are likely to be related just because they share something in common. As an example, children of the same mother share similar familial background and environment that contribute towards their survival rate. Most of researchers in the literature admit that the correlation should be considered, but less has been done so far (Guo and Rodriguez, 1992). Siblings share certain unobserved characteristics (heterogeneity) which may not be described enough by covariates in the models and neglecting such heterogeneity association may give rise to estimates which are unreliable (Guo and Rodriguez, 1992). In most applications of survival analysis, only few and known covariates such as gender, age and marital status are included in the analysis (Li and Wu, 2018). There are many other factors that influence survival such as family diet, life style, health status and smoking status which are unknown and end up not being included in the analysis. It is not always possible to add all important covariates in the model because of many reasons. For example, some covariates are not added because we do not know that they are important, while others are not added due to reasons of economic, ethical and practical nature. Eliminating those important covariates can create some unobserved heterogeneity between subjects (Vaupel et al., 1979).

In most cases, researchers do not analyse survival data of related individuals in an optimal way, mainly because of ignoring the association which resulted

in a biased estimate. Most of the limited studies conducted on child mortality used logistic regression and Cox proportional hazard models to estimate the significance of the risk factors on child mortality. Logistic regression is not an optimal method because it does not consider the time to event variable (Schober and Vetter, 2018). Cox proportional hazard model on the other hand assumes that survival times of individuals are independent, and this requires a population of the same kind, but in most instances the population is not alike. In some situations, researchers ignore left truncation, which needs to be considered to avoid getting misleading results and conclusions. The main purpose of the study is to model under five child mortality in South Africa by applying frailty and copula survival models that consider clustering and left truncation. These models provide unbiased estimates as they take association of siblings into consideration. Due to the size of our data sets, sample splitting technique was used to partition the data into sub-samples, analyse each sub-sample separately and then combine estimates of sub-samples.

1.4 Motivation of the study

Standard regression models are not appropriate to analyse clustered survival data and therefore techniques that consider clustering need to be considered (Bouwmeester et al., 2013). The use of standard logistic regression and survival models such as Cox model is not appropriate and can cause biased estimates because of the association of survival times in the data. Even though survival analysis has been studied extensively, frailty and copula models have not yet been applied and tested to the left truncated under five child mortality data set with clusters of large and unequal sizes in South Africa. To make efficient and valid inferences, we need statistical methods that can account for associations among observations within clusters. Furthermore, sample splitting technique has not yet been applied in survival analysis. In this study, we

used appropriate techniques designed for left truncated data sets which also considers dependence between observations within a cluster. The survival of subjects in a cluster such as children from the same family in this case depends on the value of frailty. Frailty information of those subjects with left truncated survival times needs to be considered when analysing the data (Jensen et al., 2004). Ignoring truncation and censoring will make our estimates of population parameters to be inconsistent. The use of sample splitting technique also makes this study unique.

1.5 Purpose of the study

1.5.1 Research aim

The aim of the study is to apply sample splitting techniques in a survival analysis setting and compare clustered survival models considering left truncation to the under five child mortality data set in South Africa.

1.5.2 Objectives

The following objectives have been followed to achieve the aim of the study:

- to compare survival curves using non-parametric tests;
- to analyse the under five child mortality data set using marginal survival model, where the dependence structure between the event times is taken care of via a robust standard error estimation technique;
- to compare Cox proportional and shared frailty models;
- to model the association of individuals within a cluster using frailty models that also consider left truncation;
- to explore association within a cluster by using copula models;

- to apply sample splitting techniques in a survival analysis setting;
- to compare Clayton-Oakes copula and shared frailty models with respect to how they handle association within a cluster;
- to assess if there are unobserved genetic and environmental factors that aggravate under five child mortality;
- to recommend to decision makers and programme managers in the child health sector on how the application of advanced survival models can be useful to improve situation of child mortality in South Africa.

1.6 General literature review

In many studies, there is a natural grouping of individuals in such a way that survival times of individuals belonging to the same group may be related (Martinussen and Scheike, 2007). Ignoring such correlation may produce estimates that are faulty and unreliable. There are two approaches that are mainly used to analyse correlated survival time data. These are frailty and copula models. In the present study, the two models were investigated with respect to the manner with which they handle association within clusters. Previous investigators such as Zhenzhen (2000), Moerbeek et al. (2003), and Islam et al. (2010), ignored association within clusters. It was shown by Bouwmeester et al. (2013) that a model with a random effect was better compared to the standard logistic regression model. There are researchers who have recognised that ignoring association between related individuals in the survival studies would produce faulty and unreliable results. These include: Vaupel et al. (1979), Guo and Rodriguez (1992), Sastry (1997), and Mahmood et al. (2013). It has been pointed out by Sainani (2010) that application of many statistical tests to observations that are correlated will overestimate p-values in situations where we consider within-subject or within-cluster effect and underestimate p-values

in cases where we consider between-cluster effects.

1.7 Introducing under five child mortality data

This study is conducted using mortality and causes of death data set and recorded live birth data set obtained from Stats SA. Since 2006, Stats SA has been collecting data on deaths using forms containing death information obtained from the Department of Home Affairs for data processing. The data sets are openly available and can be downloaded from Stats SA website. Mortality and causes of death data set and recorded live birth data set were merged so that different survival analysis techniques can be possible to apply. The two merged data sets were in SPSS file formats and analysis was done using R software. A brief description of the two data sets is given.

1.7.1 Mortality and causes of death data set

The form which is used to notify death of a person in South Africa confirms legally that the death occurred, and the same form is also used to prepare mortality and causes of death statistics. A death certificate is issued after the death registration process is completed. Stats SA collects all forms on a regular basis to capture, process, analyse and disseminate data sets containing mortality and causes of death. During data processing, forms are coded according to the year of death and are also given different numbers to identify them. Socio-demographic variables and causes of death are coded. Stats SA processes death certificates from the Department of Home Affairs and publishes statistical release called mortality and causes of deaths on annual basis (Stats SA, 2017). Together with the statistical release, raw data are also made available for public use. The limitations and the under reporting are well documented in the statistical release and data are declared adequate to use for research purposes. Contents of the mortality and causes of deaths data include geo-

graphic information, dates of birth, dates of death, demographic information of the deceased, particulars of next of kin, educational information, occupation as well as broad description of causes of deaths. The earliest unit record data published was from year 2006 and the most recent is year 2015. Earlier years, i.e., 2006 to 2008 contained less information regarding dates and years as compared to years from 2009 to 2015. As a result of the above limitation, only years from 2009 to 2015 were extracted. The death data set was reduced by selecting children who died from 2009 to 2015 and that resulted to 348941 deaths.

1.7.2 Recorded live births data set

Birth of a child in South Africa is recorded in the population register and the process is coordinated by Department of Home Affairs in South Africa. Children are supposed to be registered within 30 days of their birth date. It is also allowed to register them after 30 days as long as strong reasons for not registering them in time are given (Stats SA, 2017). Birth month, birth year, birth registration month and birth registration year are included when reporting birth cases. The background characteristics include gender, geographic information such as province of birth and district municipalities. There were 23 601 976 people in total in the births data set recorded from 1998 to 2015. These people were recorded from birth registration forms by Stats SA provided by the Home Affairs Department. There are five categories of recorded live birth files. These are: birth cases happened within 30 days of birth date; birth cases happened between 30 days and 15 years; births cases happened in health care facilities and birth cases happened outside South Africa to South Africans. To provide comparable births data set, only children born between 2009 to 2015 were selected and that resulted to 7020612 births.

1.7.3 Matching of birth and death data sets

Variables such as birth province, year of birth, month of birth and gender were used to match in the causes of death data with those in recorded live births data set. Observations not matching were not included in the final data set. The final data set containing children born between 2009 and 2015 had a total of 7020612 children.

1.7.4 Data sets used in the study

Two data sets were established from the final data set containing 7020612 children which was created after merging birth and death data sets. The two data sets were established so that different survival techniques can be possible to apply. The two data sets used in the study to demonstrate the developed methodology are described in the following sections.

1.7.4.1 Data set 1

This data set was established after 18214 children born outside South Africa and two observations that were wrongly captured were excluded. The total number of children resulted to 7002396 of which 294507 (4.2%) were indicated as died and 6707889 (95.8%) alive. From the total of 7002396 children, 3473972 (49.6%) were females and 3528424 (50.4%) were males. This data set contains stillbirths, babies who survived for less than a day (24 hours) and also babies who survived more than 24 hours. Stillbirths are regarded as babies born dead after at least 26 weeks of gestation. The response variable in this case survival time was obtained by calculating the difference (in days) from date of birth until date of death or censoring date if the child survived after the end of the study in this case, 31 December 2015. Observations are right-censored if no death had happened at the time the data are analysed. All those who survived for less than a day were given a survival time of 0.5 which is equivalent to half

a day and all stillbirths were given a survival time of zero. A contingency table including year, gender and province is given in Table 1.2.

1.7.4.2 Data set 2

This data set was established so that frailty and copula models can be applied. Children included in this data set were selected from data set 1, but excluded stillborn babies and children with missing mother's identity number because the mother's identity number was used to identify children from the same mother to form clusters. This data set was not a random sample of the entire population of children in data set 1. Only children with mother's identity number and non-stillbirths were selected. This resulted to a total of 2072621 children of which 25055 (1.2%) were indicated as died and 2047566 (98.8%) as alive. There were 1028141 (49.6%) females and 1044480 (50.4%) males. The recording of death information was problematic because only those who died between 2013 and 2015 had their death information recorded. This resulted to missing death data for children that died between 2010 and 2012. Due to missing death information in this data set, survival models that consider left truncation need to be used in order to get reliable results.
A contingency table including year, gender and province is given in Table 1.3.

Table 1.2: Contingency table including year, gender and province for dataset 1

		Year							
Gender	Province	2009	2010	2011	2012	2013	2014	2015	Total
	Limpopo	60225(13.8%)	61523(14.1%)	63804(14.6%)	63963(14.6%)	63685(14.6%)	63685(14.6%)	60492(13.8%)	437377(100%)
	Eastern cape	63779(15.1%)	61377(14.5%)	62373(14.8%)	61418(14.5%)	59353(14.1%)	59619(14.1%)	54228(12.8%)	422147(100%)
	Free State	27219(14.9%)	27713(15.1%)	26983(14.7%)	26612(14.5%)	25318(13.8%)	25854(14.1%)	23539(12.8%)	183238(100%)
	Gauteng	99589(14.2%)	99492(14.2%)	99466(14.2%)	99922(14.2%)	105313(15.0%)	102916(14.7%)	95386(13.6%)	702084(100%)
Female	Kwazulu Natal	111268(15.1%)	109166(14.8%)	108518(14.7%)	105730(14.4%)	106140(14.4%)	103837(14.1%)	91384(12.4%)	736043(100%)
	Mpumalanga	42995(14.8%)	41600(14.3%)	43416(14.9%)	43644(15.0%)	40554(14.0%)	41497(14.3%)	36775(12.7%)	290481(100%)
	North West	40169(15.3%)	39386(15.0%)	39615(15.1%)	39779(15.2%)	35204(13.4%)	34928(13.4%)	33004(12.6%)	262085(100%)
	Northern Cape	12141(14.2%)	12064(14.1%)	119161(14.0%)	12569(14.7%)	12334(14.4%)	12538(14.6%)	1200(14.0%)	85607(100%)
	Western Cape	53293(15.0%)	52275(14.7%)	51689(14.6%)	51094(14.4%)	48402(13.6%)	50246(14.2%)	47911(13.5%)	354910(100%)
	Total	510678(14.7%)	504596(14.5%)	507825(14.6%)	504731(14.5%)	49630(14.3%)	495120(14.3%)	454719(13.1%)	3473972(100%)
	Limpopo	60758(13.7%)	62359(14.1%)	65457(14.8%)	64696(14.6%)	64247(14.5%)	64282(14.5%)	61481(13.9%)	443280(100%)
	Eastern cape	64786(15.1%)	62513(14.5%)	63324(14.7%)	62592(14.6%)	60754(14.1%)	61136(14.2%)	54982(12.8%)	430087(100%)
	Free State	27783(15.0%)	27526(14.8%)	27038(14.6%)	27242(14.7%)	25981(14.0%)	26294(14.2%)	23934(12.9%)	185798(100%)
	Gauteng	100704(14.1%)	101565(14.2%)	101622(14.2%)	101961(14.3%)	107439(15.0%)	104497(14.6%)	97053(13.6%)	714841(100%)
Male	Kwazulu Natal	111735(15.0%)	110367(14.8%)	109066(14.7%)	107043(14.4%)	106933(14.4%)	105274(14.2%)	92841(12.5%)	743259(100%)
	Mpumalanga	42949(14.6%)	42283(14.4%)	43891(15.0%)	44113(15.0%)	41224(14.1%)	41923(14.3%)	36911(12.6%)	293294(100%)
	North West	40884(15.4%)	40563(15.2%)	40325(15.1%)	40002(15.0%)	35628(13.4%)	35629(13.4%)	33250(12.5%)	266281(100%)
	Northern Cape	12342(14.1%)	12361(14.1%)	12481(14.3%)	12636(14.4%)	12720(14.5%)	12645(14.5%)	12310(14.1%)	87495(100%)
	Western Cape	54750(15.0%)	53500(14.7%)	52854(14.5%)	52556(14.4%)	49890(13.7%)	51824(14.2%)	48715(13.4%)	364089(100%)
	Total	516691(14.6%)	513037(14.5%)	516058(14.6%)	512841(14.5%)	504816(14.3%)	503504(14.3%)	461477(13.1%)	3528424(100%)

Table 1.3: Contingency table including year, gender and province for dataset 2

		Year						
Gender	Province	2010	2011	2012	2013	2014	2015	Total
	Limpopo	841(0.6%)	1314(1.0%)	4309(3.2%)	60339(44.9%)	7586(5.6%)	59980(44.6%)	134369(100%)
	Eastern cape	1609(1.3%)	2233(1.8%)	5765(4.6%)	55699(44.1%)	7486(5.9%)	53528(42.4%)	126320(100%)
	Free State	351(0.7%)	433(0.8%)	1067(2.1%)	24185(47.5%)	1879(3.7%)	23028(45.2%)	50943(100%)
	Gauteng	1668(0.8%)	2194(1.0%)	5611(2.5%)	103647(46.6%)	10363(4.7%)	98893(44.5%)	222376(100%)
Female	Kwazulu Natal	2669(1.2%)	4174(1.9%)	10822(4.9%)	97466(44.1%)	15160(6.9%)	90582(41.0%)	220873(100%)
	Mpumalanga	815(1.0%)	1246(1.5%)	3440(4.1%)	37928(44.9%)	4851(5.7%)	36237(42.9%)	84517(100%)
	North West	599(1.0%)	994(1.6%)	2796(4.6%)	27497(22.2%)	2909(4.7%)	27412(44.1%)	62207(100%)
	Northern Cape	137(0.5%)	178(0.7%)	553(2.2%)	11649(46.3%)	797(3.2%)	11866(47.1%)	25180(100%)
	Western Cape	602(0.6%)	795(0.8%)	2049(2.0%)	46243(45.6%)	4417(4.4%)	47250(46.6%)	101356(100%)
	Total	9291(0.9%)	13561(1.3%)	36412(3.5%)	464653(45.2%)	55448(5.4%)	448776(%)	1028141(100%)
	Limpopo	858(0.6%)	1421(1.0%)	4401(3.2%)	60839(44.8%)	7388(5.4%)	60928(44.9%)	135835(100%)
	Eastern cape	1555(1.2%)	2254(1.7%)	5719(4.4%)	57150(44.4%)	7815(6.1%)	54330(42.2%)	128823(100%)
	Free State	323(0.6%)	449(0.9%)	1176(2.3%)	24785(47.7%)	1796(3.5%)	23418(45.1%)	51947(100%)
	Gauteng	1704(0.8%)	2192(1.0%)	5776(2.6%)	105411(46.6%)	10639(4.7%)	100499(44.4%)	226221(100%)
Male	Kwazulu Natal	2585(1.2%)	4311(1.9%)	11148(5.0%)	98123(43.8%)	15932(7.1%)	91962(41.0%)	224061(100%)
	Mpumalanga	790(0.9%)	1239(1.5%)	3429(4.0%)	38474(45.3%)	4752(5.6%)	36339(42.7%)	85023(100%)
	North West	684(1.1%)	940(1.5%)	2684(4.3%)	27804(44.4%)	2916(4.7%)	27556(44.4%)	62584(100%)
	Northern Cape	153(0.6%)	176(0.7%)	590(2.3%)	12098(46.6%)	802(3.1%)	12157(46.6%)	25976(100%)
	Western Cape	608(0.6%)	838(0.8%)	2190(2.1%)	47685(45.8%)	4600(4.4%)	48090(46.2%)	104012(100%)
	Total	9260(0.9%)	13820(1.3%)	37113(3.6%)	472369(45.2%)	56640(5.4%)	455279(43.6%)	1044480(100%)

Mothers of children were considered as the clustering variable. This resulted to a total of 1945471 different mothers (clusters) varying in sizes between 1 and 5. The distribution of cluster sizes according to the number of clusters is given in Table 1.4.

Table 1.4: Distribution of cluster sizes according to the number of clusters

Cluster size	No. of clusters
1	1822361
2	119200
3	3785
4	120
5	5
Total	1945471

There were 1822361 clusters of size 1, 119200 clusters of size 2, 3785 clusters of size 3, 120 clusters of size 4 and 5 clusters of size 5.

The data of children under the age of five from the study appear in Table 1.5.

Table 1.5: South Africa under five child mortality data set

id	clusterid	gender	Province	Year	twin	order	status	Time	Trunctime
1	1	Female	Western Cape	2013	0	0	0	1003	0
2	2	Male	Western Cape	2013	0	0	0	931	0
3	3	Female	Kwazulu Natal	2012	0	0	0	1393	299
...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...
2072626	1945477	Female	Western Cape	2015	1	0	0	219	0
2072626	1945477	Female	Western Cape	2015	1	0	0	219	0

The first column in Table 1.5 contains the unique child identification number. The second column contains the cluster identification number. The third column gives the gender of children. The fourth column contains the province in South Africa where a child was born. The fifth column is the year of birth of children. The sixth column is the twin identifier. The seventh column contains the birth order which gives previous number of children that the mother had. The eighth column is the status of an event, taking the value 1 if the child died

and zero if censored or still alive and the ninth column gives the time (in days) to death or censoring. The last column gives the time for truncation.

Due to the limitations associated with data provided by Stats SA, the covariates that are available and expected to affect the survival of under five children are year, gender, province, twin and order. These covariates are included in different models in the subsequent chapters. The variables used in the study are fully described and summarised in Table 1.6.

Table 1.6: Description of variables used in the analysis

Variables	Description	Codes/Values
Gender	Gender of a child	0=Female, 1=Male
Province	Province of birth	0=Limpopo 1=Eastern Cape, 2= Free State,3= Gauteng, 4= Kwazulu Natal,5= Mpumalanga,6= North West, 7 =Northern Cape, 8=Western Cape
Year	Year of birth	0=2009,1=2010,2=2011,3=2012,4=2013,5=2014,6=2015
Twin	Twin identifier	0 =not part of a twin, 1= part of a twin
Order	Previous number of living children	0= Eldest, 1= Second, 2= Third, 3= Fourth, 4= Fifth
Status	Survival status	0 =alive/censored, 1= dead
Time	Follow-up time	Number of days between day of birth and day of death or censoring
Trunctime	Time for left truncation	Number of days between day of birth and day of truncation (31/12/2012)
clusterid	Cluster variable identifying children from the same mother	

1.8 Overview of thesis

This thesis is divided into six chapters. The main aim of the first chapter is to describe data sets used in the study, to give a background of the study and to outline the purpose of the study, which includes research aim and objectives. Chapter 2 deals mainly with univariate survival analysis without considering clustering. This chapter contains some basic concepts of survival analysis, notations, theorems and basic results on which the methodology developed in this thesis is based. In Chapter 3 we show that left truncated shared frailty models are well suited for the left truncated data set with clusters of unequal sizes. Chapter 4 tackles the problem of modelling multivariate survival data that are grouped in clusters of different sizes through Archimedean copulas. We fit the

data with the Clayton-Oakes model and model the marginal survival functions by two stage estimation approach to obtain parameter estimates and introduce sample splitting technique. In Chapter 5 we compare shared frailty and copula models with respect to how they handle association within clusters. We also used the sample splitting technique introduced in Chapter 4 to partition our data set. Finally, in Chapter 6, we give the general summary of all chapters and main findings in the thesis and possible work to consider in future.

Chapter 2

Univariate survival analysis

In survival analysis, it is advisable to first consider simple univariate analysis without clustering before delving into more complicated models. In this thesis, univariate survival methods were considered first in this chapter before advanced clustered survival models were introduced in Chapter 3 and Chapter 4. In this chapter, different analysis techniques such as logistic regression and Cox proportional hazard models are considered to establish factors influencing probability of stillbirths and also to investigate the effects of covariates upon time to death of these children. Non-parametric procedures for survival data such as Kaplan-Meier estimator and Log-rank test are discussed and analysed. All analyses in this chapter were done using data set 1 described in Chapter 1, Section 1.7.4.1.

2.1 Introduction to survival analysis

Survival analysis is a branch of statistics used to analyse time to events data. In survival analysis, the event may be time until recovery from accident, time until marriage, time to an epileptic seizure, time it takes for a patient to re-

spond to a therapy, or death as it applies in our study. The time of entry is not always at the same time point for all individuals in the study and this results in what we call staggered entry. Staggered entries do not affect the analysis. What matters most is the total duration of time individuals spent in the study. Each individual's time to event is measured from the date of entry to the time of event. The reason why standard statistical techniques are not appropriate in the analysis of survival data is because survival data are positively skewed and not symmetrically distributed as normal distributions. Another reason is that survival times are often censored and truncated, which makes analysis to be more complicated (Schober and Vetter, 2018). Censoring and truncation are clearly defined in the next section.

2.1.1 Special features of survival data

Survival data are difficult to analyse due to censoring and truncation. These two features make survival analysis different from other areas of statistics. Both censoring and truncation need to be taken into consideration when dealing with survival data to avoid getting misleading results (Schober and Vetter, 2018).

2.1.1.1 Censoring

Censoring occurs when we cannot fully observe a time until an event, but only observe some boundaries for this time. Three different types of censoring schemes are right censoring, left censoring and interval censoring.

1. **Right censoring** happens when an individual leaves the study before an event of interest occurs. This occurs when an individual is lost due to follow-up or when the study ends, and the individual has not experienced the event of interest.

There are three types of right censoring, namely:

- Type I or fixed censoring
- Type II censoring
- Type III or random censoring

(a) **Type I or fixed censoring**

Let $t_c \in \mathcal{R}$ be a fixed time point and take a sample of survival times $T_1, ..., T_n$.

We only observe a survival time T_i if it is smaller than t_c, or else we get a fixed time point. Hence we get a sample $Y_1, ..., Y_n$ where

$$Y_i = \begin{cases} T_i, & \text{if } T_i \leq t_c \qquad i = 1, 2, ..., n \\ t_c, & \text{if } T_i > t_c \end{cases}$$

(b) **Type II censoring**

Suppose that $s < n$ and let $T_{(1)}, ..., T_{(n)}$ be the ordered survival times. We observe until the $s - th$ system has failed. Hence we get

$$Y_{(i)} = \begin{cases} T_{(i)}, & \text{if } T_{(i)} \leq T_{(s)} \qquad i = 1, 2, ..., n \\ T_{(s)}, & \text{if } T_{(i)} > T_{(s)} \end{cases}$$

(c) **Type III or random censoring**

Let $C_1, ..., C_n$ be a sample of censoring times. We observe a sample of couples, $(Y_1, \delta_1), ..., (Y_n, \delta_n)$ where, for $i = 1, 2, ..., n$,

$$Y_i = \min(T_i, C_i) = \begin{cases} T_i, & \text{if } T_i \leq C_i \\ C_i, & \text{if } T_i > C_i \end{cases}$$

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

We assume that T_i and C_i are independent for $i = 1, ..., n$.

2. **Left censoring** happens when the event of interest has already occurred to the individual prior to the start of the study. If the event of interest is death as it applies in our case, the data cannot be left censored.
   We observe a sample $(Y_1, \delta_1), ..., (Y_n, \delta_n)$, where, for $i = 1, ..., n$,

$$Y_i = \max(T_i, C_i) = \begin{cases} T_i, & \text{if } T_i \geq C_i \\ C_i, & \text{if } T_i < C_i \end{cases}$$

$$\delta_i = I(T_i \geq C_i) = \begin{cases} 1, & \text{if } T_{(i)} \geq C_i \\ 0, & \text{if } T_{(i)} < C_i \end{cases}$$

   As an example, if we are following individuals until they become HIV positive, we may record a failure when an individual first tests positive for the virus. We may not know the exact time when he or she was exposed to the virus and hence the survival time is left censored because the true survival time ending at exposure is shorter than the follow-up time ending when an individual tests positive (Kleinbaum, 1998).

3. **Interval censoring** happens when an event of interest has occurred within some interval. In the interval censoring, we get an interval in which the event occurred for each individual instead of survival times $T_{(1)}, ..., T_{(n)}$. In this case we get $(L_1, U_1), (L_2, U_2), ..., (L_n, U_n)$, where L_i is the

lower limit of the interval and U_i the upper limit of the interval.

An example to illustrate right censoring and left censoring schemes is a study age at which children learn a certain task. Some already knew a task (left-censored) and some had not learnt by end of a study (right censored).

2.1.1.2 Truncation

Truncation happens when we only observe certain individuals whose event times are within certain intervals. Truncation schemes are left truncation and right truncation. **Left truncation** happens when some individuals are not observed if the event happens before a certain date. This is when individuals come under observation only some known time after the natural time origin of the phenomenon under study. Individuals only show up in the data set when their event of interest happens later than a certain boundary. Otherwise they are missed completely. **Right truncation** happens when we include in the study, individuals who have experienced the event of interest by a specified time. This thesis dealt with left truncation because of the missing death data for children who have died between 2010 and 2012.

The illustration of these special features of survival data using a sample of six subjects is given in Figure 2.1. The horizontal axis indicates the time of event (in months). Capital letter X indicates the exact time at which the event of interest is observed. Capital letter O indicates that the subject is right censored at that point and finally T indicates left truncated point.

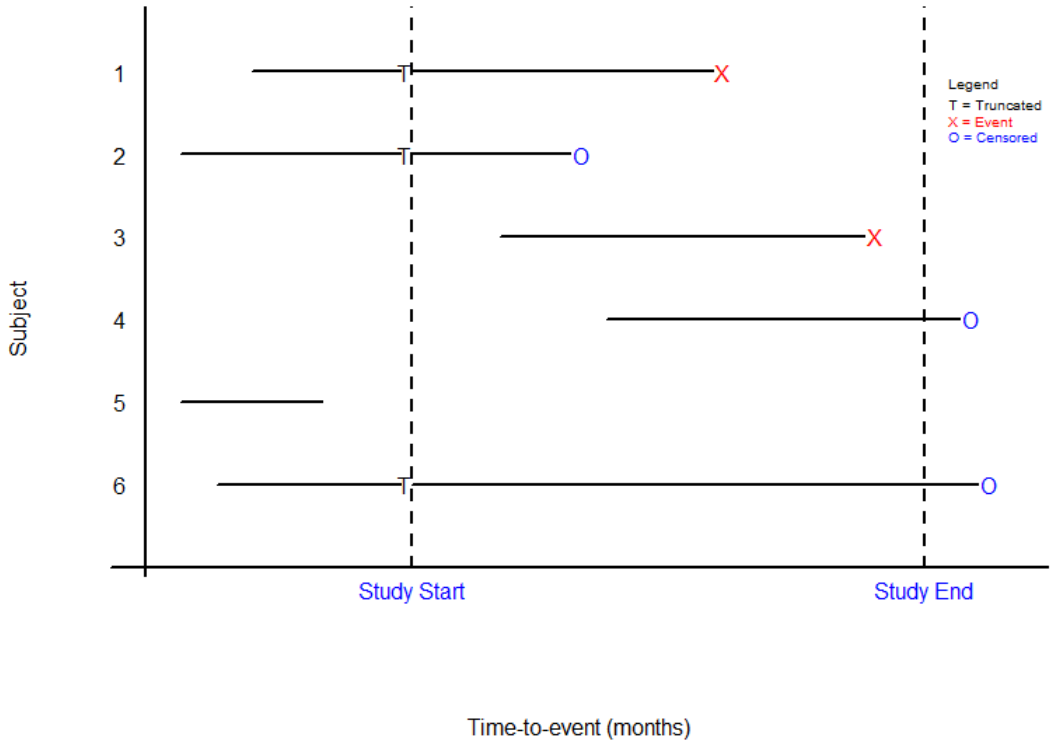

Figure 2.1: Illustration of special features of survival data using a sample of six subjects

We can see from Figure 2.1 that subjects 1 and 2 were exposed to the risk for some time. Both subjects were left truncated. Subject 1 experienced event of interest whereas subject 2 was right censored due to loss of follow-up. Subject 3 was followed until an event of interest occurred. Subject 4 was right censored due to termination of the observation period. This individual was still alive at the end of the study. Subject 5 was left censored. In this particular case we do not have X at the right end of the time line for this subject because we do not have any information as to when the event of interest occurred. Finally, Subject 6 was both left-truncated and right censored.

2.1.2 Basic concepts of univariate survival analysis

This section provides a brief outline of the basic concepts of univariate survival analysis.

Suppose we have right censored survival times of n independent individuals. We observe for individual i time $Y_i = \min(T_i, C_i)$, where T_i is the survival time and C_i is the censoring time.

We define T as a positive random variable representing the survival time, i.e., $T \geq 0$, and it has the probability density function denoted by $f(t)$ and cumulative density function denoted by $F(t)$.

The cumulative density function is mathematically represented as:

$$F(t) = \int_0^t f(s)ds = P(T \leq t),$$

where t is the specific value of T and $P(T \leq t)$ is the probability that the survival time until the event happens is less than or equal to the specific value of T. $F(t)$ gives us the probability that an event of interest has occurred by time t. To get $f(t)$, we differentiate $F(t)$ with respect to t, as follows:

$$f(t) = \frac{dF(t)}{dt} = F'(t) \Rightarrow f(t) = \lim_{dt \to 0} \frac{F(t + dt) - F(t)}{dt},$$

where dt represents the small-time interval. Relating this to the current study, $f(t)$ will give us the unconditional instantaneous probability that a child died in the interval (t, dt) and it is formally written as follows:

$$f(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt)}{dt}.$$

The two basic functions that are very important in survival analysis are survival and hazard functions. The survival function $S(t)$ focuses on surviving and the hazard function $h(t)$ focuses on failing of the subject. The two concepts are

mathematically presented in the following way:

$S(t) = P(T > t) = \int_t^\infty f(x)dx = 1 - F(t)$
and
$h(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt | T \geq t)}{dt} = \frac{f(t)}{S(t)}$.

The survival function $S(t)$ gives the probability that a subject or item will survive beyond a specified time t. The hazard function $h(t)$ is the instantaneous failure rate that a subject survived up to time t.

Two properties of the survival function $S(t)$ are:

- $S(t)$ is a decreasing function on $[0, \infty]$
- $S(0) = 1$ and $S(\infty) = 0$.

The survival, density and hazard functions have the following relationships:
$f(t) = -\frac{dS(t)}{dt}$
$h(t) = \frac{f(t)}{S(t)} = -\frac{dlogS(t)}{dt}$
$S(t) = exp(-H(t))$ where $H(t) = \int_0^t h(s)ds = -logS(t)$.
$H(t)$ is the cumulative hazard function.
Suppose we have a random sample of pairs (T_j, d_j), $j = 1, 2, ..., n$. The likelihood function is written as
$L = \prod_{j=1}^{n} [S(t_j)]^{1-d_j} [f(t_j)]^{d_j}$,

where d_j is the number of observed events at time t_j.

2.2 Research methodology

This section presents the univariate survival analysis methods used to analyse data set 1 described in Chapter 1. The methods include logistic regression model, Kaplan-Meier estimator, log-rank test and Cox PH model.

2.2.1 The logistic regression model

This section provides a brief background on the statistical technique called logistic regression model used to predict the probability of stillbirths. The target by World Health Assembly-endorsed Every Newborn Action Plan which is to reduce stillbirth cases from 18.4 in 2015 to 12.0 in 2030 per 1000 birth cases (Bamford et al., 2018). About 2,6 million stillbirth cases occur every year across the world. Most cases of stillbirths occur in low- and middle-income countries because causes of stillbirths are not often investigated (Madhi et al., 2019). According to Aminu et al. (2014), the following were reported as risk factors connected with stillbirth in developing countries: gender, age of the mother, gestation age at birth, insufficient antenatal care, weight at birth and physical or mental illnesses of the mother during pregnancy. It has been pointed out by Aminu et al. (2014) that most of stillbirth deaths can be prevented. It is very important to investigate factors associated with stillbirths so that stillbirth cases can be reduced.

In this thesis, logistic regression model was employed to predict the probability of stillbirths and also to determine whether there is an association between being a stillborn baby and factors included in the model which are birth province, birth year, and gender. Logistic regression model is used to model the relationship between multiple independent variables and a categorical dependent variable. In most cases, the outcome variable in logistic regression is a binary event like alive versus dead, fail versus pass, win versus lose and stillbirth versus non-stillbirth. The independent variables can be binary or continuous. The difference between this model and the Proportional hazard (PH) model is that a logistic regression model calculates the probability of an event happening based on the factors included in the model. A Cox PH model is used to explore the relationship between the survival of a subject and the explanatory variables. The Cox PH model takes into account time to event which is not the

case in a logistic regression model.

2.2.1.1 Odds

Logistic regression determines the chance of an event to occur over the chance of an event not to occur. The effect of covariates is usually explained in terms of odds. The odds of an event are ratio of the likelihood that the event will happen to the likelihood that event will not happen. Suppose that p is the probability that an event will occur, and $1-p$ is the probability that it will not occur (Park, 2013).

The odds of an event are given by:

$$Odds = \frac{p}{1-p}.$$

The effect of covariates is usually explained in terms of odds. The natural log odds are modelled as a linear function of the independent variables as follows:

$$logit(y) = \ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x, \tag{2.1}$$

where p is the probability that an event will occur, x is the covariate, β_0 and β_1 are parameters of the logistic regression. We can predict the occurrence by taking antilog in simple logistic regression as follows:

$$p = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}.$$

The logit in 2.1 can be extended to include multiple covariates as follows:

$$logit(y) = \ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k. \tag{2.2}$$

The probability of occurrence based on equation 2.2 is given by:

$$p = \frac{exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)}{1 + exp(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k)}.$$

2.2.1.2 Odds ratio

The odds ratio (OR) is a measure to compare two odds relative to different events, say A and B. As an example, the odds of event A occurring relative to event B occurring is:
Odds ratio $= \frac{OddsA}{OddsB} = \frac{p_A(1-p_A)}{p_B(1-p_B)}$.

The odds ratio measures the association between an exposure and an outcome. It represents the odds that an outcome will occur given a particular exposure compared to the odds of the outcome occurring in the absence of that exposure.

2.2.1.3 Overall model evaluation

The overall fit of the statistical model gives us an idea of the strength of a relationship between all covariates included in the model. The logistic regression model with all k covariates (full model) is considered a better fit to the data if we see an improvement over the model without covariates (null model) (Park, 2013). The overall fit can be examined via the LRT and test the null hypothesis.
$H_0 = \beta_1 = \beta_2 = ... = \beta_k = 0$.

This hypothesis says that there is no independent variable that affects the prediction of the outcome. The goodness of fit index denoted by G is a χ^2 statistic with k degrees of freedom is given by:

$$G = \chi^2 = (-2 \text{ log likelihood of null model}) - (-2 \text{ log likelihood of full model}).$$

The likelihood of the null model is the likelihood of obtaining the observation if the covariates had no effect on the outcome and the likelihood of the full model is the likelihood of obtaining the observations with all independent variables included in the model.
If the p-value for the overall model fit statistic is less than 0.05, we reject H_0 and conclude that there is evidence that at least one of the covariates affects the prediction of the outcome.

Statistical significance of individual regression coefficient
We next determine the importance of each covariate included in the model. The likelihood ratio test can be used to assess the contribution of covariates that are included in the model. The LRT for a particular parameter compares the likelihood of getting the data when the parameter is zero (L_0) with the likelihood (L_1) of getting the data evaluated at the MLE of the parameter. The statistic is calculated as follows:

$$G = -2ln\frac{L_0}{L_1} = -2(lnL_0 - lnL_1).$$

An alternative way to assess the contribution of individual covariate is by using the Wald statistic. The Wald statistic is the square of the regression coefficient to the square of the standard error of the coefficient as follows:

$$W_j = \frac{\beta_j^2}{SE(\beta_j^2)}.$$

Each Wald statistic is compared with a Chi square with 1 degree of freedom.

2.2.2 Non-parametric procedures

Non-parametric procedures for survival data, namely, Kaplan-Meier estimator and Log-rank test are presented in this section.

2.2.2.1 The Kaplan-Meier estimator

The most common method of estimating the survival function $S(t)$ is the Kaplan-Meier (K-M) estimator also called the product limit estimator. It is a non-parametric statistic that gives us the probability that an individual included in the study will survive past a particular time *t*. One of the advantages of this statistic is that it is not based on the assumption of the underlying probability distribution and that is good because the distribution of survival data is skewed. To estimate the proportion of individuals who are still alive at a given time point, K-M estimator uses information of those who have died and also for those who have survived. The estimator is plotted over time and the curve is called the K-M curve. The K-M curve is a series of horizontal steps of declining magnitude that, when a large enough sample is taken, approaches the true survival function for that population. It is a step function that jumps at the observed event times. The K-M estimator shows what the probability of an event is at a particular time point. It is the method for estimating the survival function $S(t)$.

Recall that $S(t) = P(T > t)$ is the probability of survival beyond time *t*. The survival function $S(t)$ can be estimated as follows:

Suppose that p individuals have failures in a group of individuals.
Let the ordered event times be given by $t_1, t_2, ..., t_p$. We will denote the size of the risk set by r_j, the number of censored observations between the j^{th} and $j^{th}+1$ failure times. Let the number of observed events at time t_j be denoted by d_j. The risk set includes all individuals who have survived just before time t_j.

The conditional probability of an individual surviving past time t_j given survival to that time is estimated by $\frac{(r_j-d_j)}{r_j}$. Thus the unconditional probability of surviving past any time t is estimated by

$$\hat{S}(t) \approxeq \prod_{j:t_j<t} \frac{(r_j-d_j)}{r_j},$$

where $j = 1, 2, ..., p$.
$\hat{S}(t)$ is the K-M estimate.

A table to obtain K-M estimator is shown with the survival function estimate in Table 2.1.

Table 2.1: K-M estimator table

j	t_j	d_j	c_j	r_j	$r_j - d_j$	$\frac{(r_j-d_j)}{r_j}$	$\hat{S}(t)$
0	t_0	d_0	c_0	r_0	$r_0 - d_0$	$\frac{(r_0-d_0)}{r_0}$	$\frac{(r_0-d_0)}{r_0}$
1	t_1	d_1	c_1	r_1	$r_1 - d_1$	$\frac{(r_1-d_1)}{r_1}$	$\frac{(r_0-d_0)}{r_0} * \frac{(r_1-d_1)}{r_1}$
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
p	t_p	d_p	c_p	r_p	$r_p - d_p$	$\frac{(r_p-d_p)}{r_p}$	$\frac{(r_0-d_0)}{r_0} * \frac{(r_1-d_1)}{r_1} * ... * \frac{(r_p-d_p)}{r_p}$

$\hat{S}(t)$ only goes to zero if the last observation is uncensored. If there is no censoring, the K-M estimator equals the empirical survival estimate.

2.2.2.2 Log-rank test

The Log-rank test is the most often used statistical test to compare survival distributions of two or more groups such as treated versus control group in a randomised trial. The test is named after Nathan Mantel and David Cox, hence is called Mantel-Cox test. We use it to test the null hypothesis that there is no difference between the survival curves.

The Log-rank test statistic for two groups is as follows:

$$\chi^2 = \frac{(O_i - E_i)^2}{Var(O_i - E_i)^2} \sim \chi^2_1.$$

Mathematically, the log-rank formula for more than two groups is very complicated in such a way that a computer is needed. In general, the test statistic is approximately chi-square in large samples with $k - 1$ degrees of freedom, where k is the number of groups being compared (Kleinbaum, 1998). Other alternatives to the log-rank test are Wilcoxon, the Tarone-Ware, the Peto and the Flemington-Harrington tests. In this thesis, the log-rank test will be applied to check if survival curves are the same or not.

2.2.3 The Cox Proportional hazard model

One of the main goals of the survival analysis is to investigate if there are factors (covariates) that affect the risk of an event of interest. The aim in survival analysis is to obtain some measure of effect describing the relationship between covariates and time to event. The effect of the covariates is often measured using Proportional hazard (PH) model (Kristiansen, 2012). The Cox Proportional Hazard model proposed by Cox (1972) is the most commonly used model for the analysis of censored survival data in the presence of covariates. In this model, time is treated as continuous and it makes no assumption about the shape of the hazard function. The effect of the covariates is assumed to act multiplica-

tively on the baseline hazard rate and the ratio of the hazards is constant over survival time (Basar, 2017). There are two assumptions that must be satisfied before one can apply this technique. The first assumption is that there is a linear relationship between log hazard and a covariate. The second assumption is that the hazard ratios are constant over time. The second assumption means that the hazard for one individual is proportional to the hazard for any other individual. If these two assumptions are violated, they can lead to biased results (Basar, 2017). The strength of the Cox PH model lies in the ability to model and test many inferences without making any specific assumptions about the form of the life distribution model (Hanagal, 2011).

The Cox Proportional hazard model is used to assess the effects of the covariates on the hazard function and does not take into consideration clustering acting as if the event times are independent of each other, even if they belong to the same cluster (Gachau, 2014). The analysis in the Cox PH model assumes that individuals in the study are all homogeneous in the sense that they are prone to experience events in the same way, but in reality, some individuals are more frail and thus, more likely to experience an event of interest.
The conditional hazard of an individual given the covariate values $X_1, X_2, ..., X_p$ is given by

$$h(t|X) = h_0(t)\mathbf{exp}(\beta^{'}X), \tag{2.3}$$

with $h_0(t)$ the baseline hazard function at time *t* and β the regression coefficients. The baseline hazard function $h_0(t)$ is the hazard function of individuals whose covariate values equal to 0. The good thing about this model is that $h_0(t)$ can be left unspecified or take a parametric form. A popular choice is the Weibull baseline hazard which takes the form $h_0(t) = \lambda\rho t^{\rho-1}$ with $\lambda > 0$ a scale parameter and $\rho > 0$ a shape parameter.

Suppose that $h_a(t)$ and $h_b(t)$ are the hazard functions at time *t* of the a^{th} and b^{th}

individuals respectively. Then the ratio of the two hazard functions is given by

$$\frac{h_a(t)}{h_b(t)} = \exp[\beta^{'}(X_a - X_b)]. \tag{2.4}$$

Equation 2.4 implies that the hazard ratio of two individuals is constant over time, and that explains the notion proportional hazards model. The hazard ratio which can be denoted by HR is the measure of the effect of the given covariates on survival time.
$HR = 1$ implies that individuals in the two groups are at the same risk of getting the event of interest. $HR > 1$ implies that the event is happening faster for the treatment group than for the control group and $HR < 1$ implies that the event is happening slower for the treatment group than for the control group.

2.2.3.1 Estimation in Cox Proportional Hazard model

By fitting the Cox Proportional hazard model, we wish to estimate model parameters or vector of regression coefficients β. The parameters can be estimated using the method of partial likelihood developed by Cox (1972) which considers only the probability of individuals that have experienced the event of interest.

To construct the partial likelihood function, we let $t_1, t_2, ..., t_n$ be the observed survival times of n individuals in the study. Furthermore, let the ordered death times of p individuals be $t_{(1)} < t_{(2)} ... < t_{(p)}$ and let $R(t_i)$ be the risk set just before $t_{(i)}$. The risk set includes those individuals alive and not censored at a time just before $t_{(i)}$.

The conditional probability that the j^{th} individual from the risk set dies at time $t_{(i)}$ is:
P(individual j dies at $t_{(i)}$)/P(one death from $R(t_i)$ at $t_{(i)}$)

$$
\begin{aligned}
&= \frac{P(\text{individual } j \text{ dies } t_{(i)})}{P(\text{one death at } t_{(i)})} \\
&= \frac{h_j(t_i)}{\sum_{k \epsilon R(t_{(i)})} h_k(t_{(i)})} \\
&= \frac{h_o(t_{(i)})\exp(\beta' x_j(t_{(i)}))}{\sum_{k \epsilon R(t_{(i)}} h_o(t_{(i)})\exp(\beta' x_k(t_{(i)}))} \\
&= \frac{\exp(\beta' x_j(t_{(i)}))}{\sum_{k \epsilon R(t_{(i)})} \exp(\beta' x_k(t_{(i)}))}.
\end{aligned}
\tag{2.5}
$$

Then the partial likelihood function for the Cox PH model is given by:

$$
L(\beta) = \prod_{i=1}^{p} \frac{\exp(\beta' x_j(t_{(i)}))}{\sum_{k \epsilon R(t_{(i)})} \exp(\beta' x_k(t_{(i)}))}. \tag{2.6}
$$

The likelihood in equation 2.6 is only for individuals not censored, $x_j(t_{(i)})$ is a vector of covariate values for individual j who dies at $t(i)$. Let δ_i be the censoring or event indicator. The event indicator is equal to zero if the j^{th} survival time is censored and one otherwise. The likelihood function can now be written as:

$$
L(\beta) = \prod_{j=1}^{n} \frac{\exp(\beta' x_j(t_{(j)}))}{\sum_{k \epsilon R(t_{(j)})} \exp(\beta' x_k(t_{(j)}))}. \tag{2.7}
$$

$R(t_{(j)})$ is the risk set at time t_j. The partial likelihood in equation 2.7 is valid only if there are no ties in the data, i.e., when we do not have two individuals with the same event time.

2.3 Data analysis and results

In this section, methods and models discussed in te previous sections were used to perform data analysis. In all the analyses, the data set described in Section 1.7.4.1 was used.

2.3.1 Descriptive statistics for categorical variables

A total of 7002396 children were included in the analysis. A summary of descriptive statistics is presented in Table 2.2.

Table 2.2: Distribution of births and deaths by some of survival determinants

Factors	Level	Total (%)	Death N (%)	Censored N (%)
Gender	Female	3473972 (49.6%)	135604(3.9%)	3338368 (96.1%)
	Male	3528424 (50.4%)	158903(4.5%)	3369521(95.5%)
Province	Limpopo	880657 (12.6%)	33889(3.8%)	846768(96.2%)
	Eastern Cape	852234 (12.2%)	24846 (2.9%)	827388(97.1%)
	Free State	369036 (5.3%)	27109(7.3%)	341927(92.7%)
	Gauteng	1416925 (20.2%)	61996(4.4%)	1354929(95.6%)
	KwaZulu-Natal	1479302(21.1%)	60597(4.1%)	1418705(95.9%)
	Mpumalanga	583775(8.3%)	23858(4.1%)	559917(95.9%)
	North West	528366(7.5%)	29148(5.5%)	499218(94.5%)
	Northern Cape	173102(2.5%)	9637(5.6%)	163465(94.4%))
	Western Cape	718999 (10.3%)	23427(3.3%)	695572(96.7%)
Year	2009	1027369(14.7%)	53631(5.2%)	973738 (94.8%)
	2010	1017633 (14.5%)	48982(4.8%)	968651(95.2%)
	2011	1023883(14.6%)	43581(4.3%)	980302(95.7%)
	2012	1017572(14.5%)	42561(4.2%)	975011(95.8%)
	2013	1001119(14.3%)	41015(4.1%)	960104(95.9%)
	2014	998624(14.3%)	36338(3.6%)	962286 (96.4%)
	2015	916196(13.1%)	28399(3.1%)	887797(96.9%)

Table 2.2 shows the frequency table of some of the covariates included in the study. Out of the total number of 70002396 children, 294507 (4.2%) were reported dead and 6707889 (95.8%) were still alive on the date of the survey. A higher percentage of death was observed among male children (4.5%) compared to female children (3.9%). The mortality rates of children under the age of five varied from one province to another in South Africa. The highest percentage of deaths was observed in Free State (7.3%), followed by Northern Cape (5.6%); while the lowest percentage of deaths was recorded in Eastern Cape (2.9%), followed by Limpopo (3.8%).

Across birth year, a highest death rate (5.2%) was recorded in 2009 and the

lowest (3.1%) was recorded in 2015. The declining of death rates over the years might be due to improvement in the health institutions, like for example, proper administration of Antiretroviral (ARV) treatment that can prevent mother-to child transmission of HIV.

2.3.2 Proportion of stillbirths

Children were grouped into two categories namely, stillbirths and non-stillbirths to investigate their survival rate separately. Stillbirths are regarded as babies born dead after at least 26 weeks of gestation and non-stillbirths are those that were born alive. The proportions of the two groups are given in Table 2.3.

Table 2.3: Percentages of stillbirths and non-stillbirths

Factors	Level	Stillbirths (%)	Non-stillbirths (%)	Total
Gender	Female	41090(1.2%)	3432882 (98.8%)	3473972
	Male	49766 (1.4%)	3478658 (98.6%)	3528424
	Total	90856(1.3%)	6911540 (98.7%)	7002396
Province	Limpopo	8100(0.9%)	872557 (99.1%)	880657
	Eastern Cape	4626(0.5%)	847608(99.5%)	852234
	Free State	8118 (2.2%)	360918 (97.8%)	369036
	Gauteng	20541(1.4%)	1396384(98.6%)	1416925
	KwaZulu-Natal	22935(1.6%)	1456367 (98.4%)	1479302
	Mpumalanga	6813(1.2%)	576962 (98.8%)	583775
	North West	7302(1.4%)	521064(98.6%)	528366
	Northern Cape	2563(1.5%)	170539(98.5%)	173102
	Western Cape	9858 (1.4%)	709141(98.6%)	718999
	Total	90856(1.3%)	6911540 (98.7%)	7002396
Year	2009	11755 (1.1%)	1015614 (98.9%)	1027369
	2010	13733 (1.3%)	1003900 (98.7%)	1017633
	2011	12866 (1.3%)	1011017 (98.7%)	1023883
	2012	13180 (1.3%)	1004392 (98.7%)	1017572
	2013	13672 (1.4%)	987447 (98.6%)	1001119
	2014	13093 (1.3%)	985531 (98.7%)	998624
	2015	12557 (1.4%)	903639 (98.6%)	916196
	Total	90856(1.3%)	6911540(98.7%)	7002396

The total number of stillbirths for this study was 90856 (1.3%), of which 41090

(1.2%) were females and 49766 (1.4%) were males. On the other hand, the total number of non-stillbirths were 6911540 (98.7%), with 3432882 females and 3478658 males. Out of 6911540 non-stillbirths, 203651 (2.9%) were reported dead and 6707889 (97.1%) were still alive. The probability of dying given that a child is not a stillbirth is 0.029 and the probability of being alive given that a child is not stillbirth is 0.971. It can be shown from Table 2.3 that the majority of stillbirth cases occurred in Free State province (2.2%), followed by KwaZulu Natal province (1.6%). The lowest percentage of stillbirth cases was in Eastern Cape province with 0.5% followed by Limpopo with 0.9%. Majority of stillbirths were born in 2013 and 2015.

The predicted probabilities of stillbirths in each of the gender groups for all provinces were calculated. These are probabilities that children are stillborn given predictor values. Figure 2.2 shows the predicted probabilities of stillbirths in each of the gender groups for all nine provinces in South Africa.

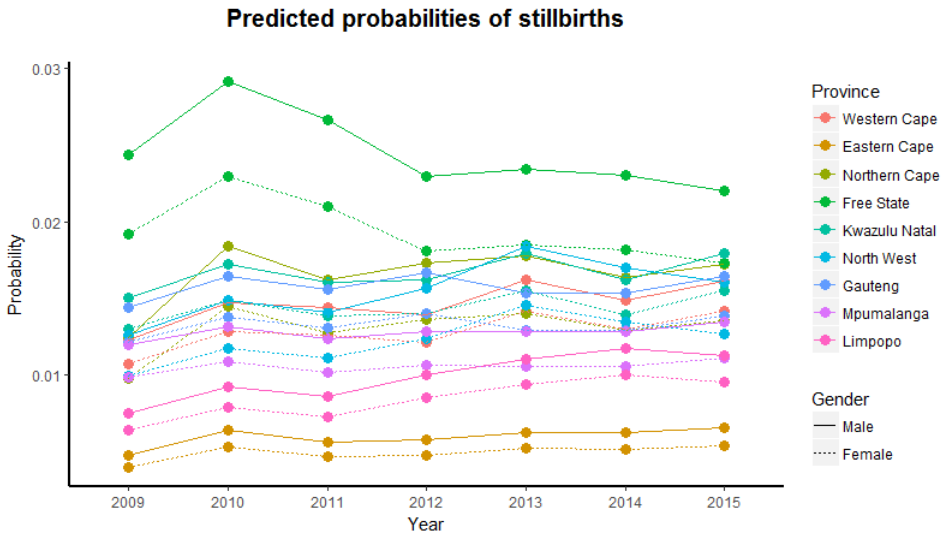

Figure 2.2: Predicted probability of stillbirths in each of the gender groups for all provinces

The probabilities of stillbirth for both males and females in Free State province are the highest compared to other provinces. On the other hand, probabilities

for both males and females in Eastern Cape province are the lowest.

2.3.3 Results from non-parametric procedures

2.3.3.1 The overall K-M survivor curve

The overall K-M survivor curve to show the probability of survival on a certain day is given in Figure 2.3.

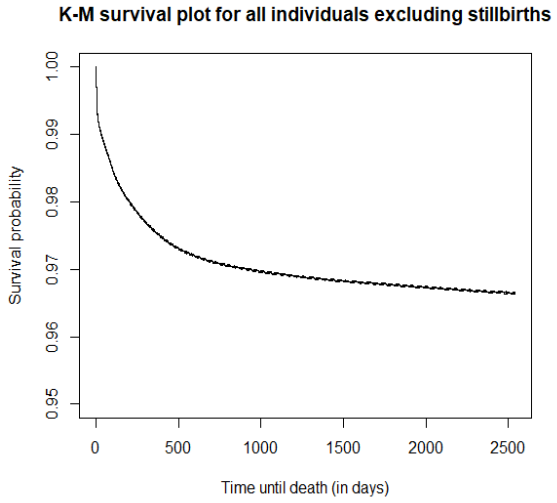

Figure 2.3: K-M survival curve for all individuals excluding stillbirths

According to the K-M survivor curve in Figure 2.3, the probability of under five children surviving was high at the first few days and then decreased as follow-up time increased. We see a drop of survival probability from 100% at the beginning of the study to 97.9% at about day 600 and after that it became constant.

2.3.3.2 Comparison of survival curves

Figures 2.4 (a) - (c) present K-M survival curves to show the risk of death of under five children as distributed across the categories of gender, year and province.

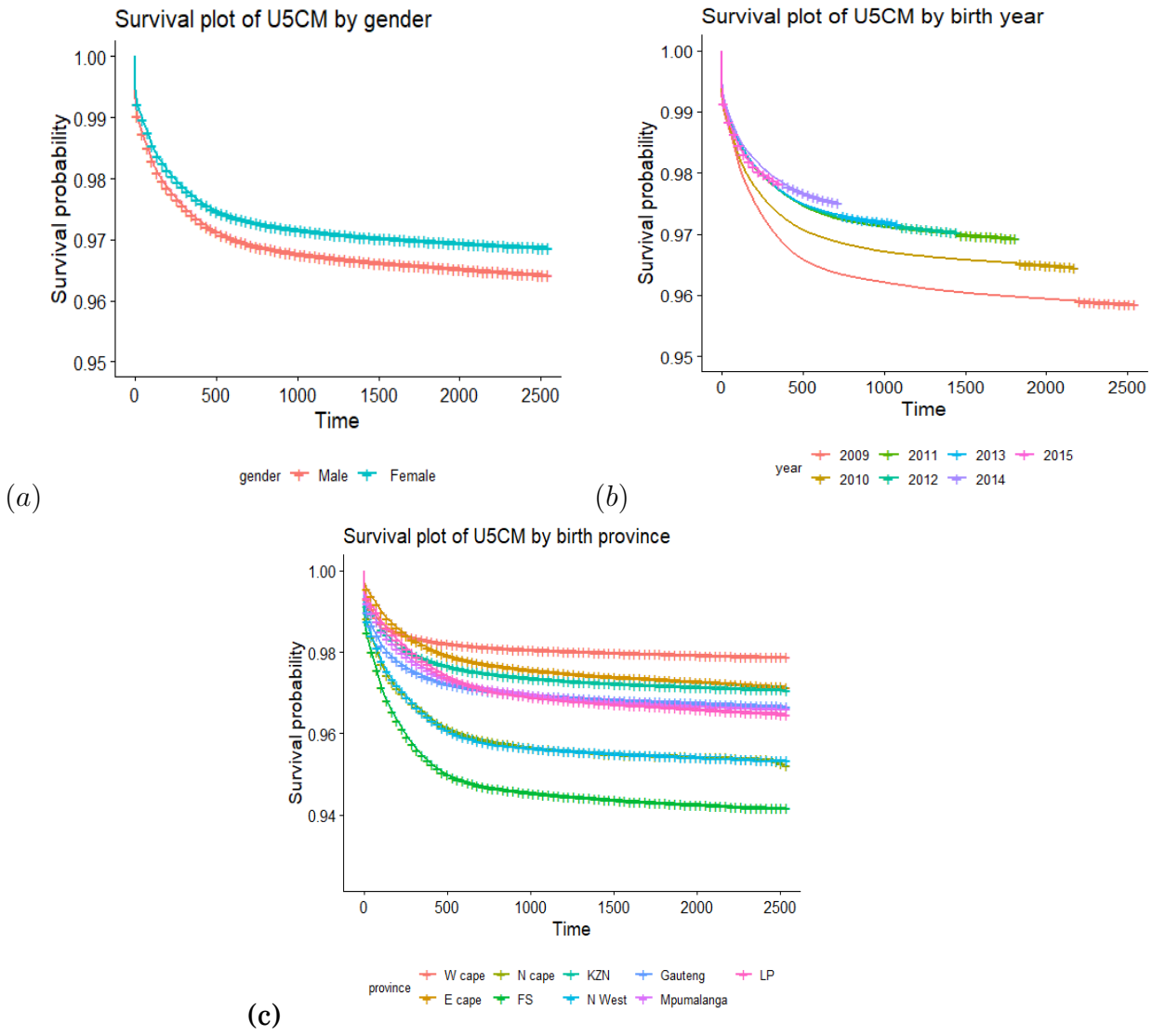

Figure 2.4: Survival plots of U5CM by (a) gender (b) year (c) province

Figure 2.4 (a) shows the K-M survival plot of under five mortality by gender. The figure shows that both groups have a similar pattern of survival. We see a rapidly descending estimated survival function within the first 400 days and then survival became constant thereafter. The line for female children lies above that for male children, which indicates that female children survived longer during the entire study period than male children.

Figure 2.4 (b) shows the K-M survival plot of under five mortality by year. The

line indicating 2014 birth year in the survival plot lies above all other lines with 2009 lowest. It displays clearly that children born in 2014 survived longer than other children. Figure 2.4 (c) shows the K-M survival plot of under five mortality by province. The plot shows that all children in all nine provinces have a similar pattern of survival. The line indicating Western Cape province is above all others with Free State the lowest. This shows that children born in Western Cape had a longer survival span than children born in other provinces.

2.3.3.3 Log-rank test

The log-rank test was performed to see whether there is a significant difference among survival experiences of two or more groups of the covariates included in the study. If p-value < 0.05, we reject the null hypothesis that groups are the same. The results of the log-rank test are given in Table 2.4.

Table 2.4: Results of log-rank test of equality of survival distribution for the three covariates

Covariates	Chi-square	d.f	p value
Gender	898	1	0.000
Province	14793	8	0.000
Year	3262	6	0.000

Based on Table 2.4, we found that the log-rank test is significant in survival experience of children in different categories of gender, province and year at 5% level of significance. We can conclude that the risk of child mortality for these three covariates varies significantly from group to group.

2.3.4 Results from the Cox PH model

The Cox PH model was fitted to the data to find factors affecting under five mortality in South Africa. The LRT was performed to check the overall effect of the variables. The results are shown in Table 2.5.

Table 2.5: Overall effect of the variable using the LRT

Covariates	*df*	Chi Square	Pr(> Chi)
Gender	1	1594.4	<0.0010
Province	8	16367	<0.0010
Year	6	1702.5	<0.0010
Gender * Province	8	27.612	0.0006
Gender * Year	6	13.361	0.0377
Province * Year	48	933.45	<0.0010

The LRT is highly significant for all covariates including the interaction terms. We would conclude that all the covariates included in Table 2.5 should remain in the model.

Tables 2.6 and 2.7 show the estimated β coefficients, standard errors, hazard ratios and p-values of the Cox PH model.

Table 2.6: Results of the Cox PH on the effect of covariates on under five mortality

Factors	Level	**Coef (β)**	**Hazard ratio**	$S_e(\beta)$	***z***	***p-value***
Gender	Female (Ref)					
	Male	0.13834	1.14837	0.01539	8.99	0.00000
Province	Limpopo (Ref)					
	Eastern Cape	-0.17993	0.83533	0.02351	-7.65	0.00000
	Free State	0.72024	2.05493	0.02351	30.64	0.00000
	Gauteng	0.08979	1.09395	0.02006	4.48	0.00000
	KwaZulu-Natal	-0.04975	0.95147	0.02023	-2.46	0.01394
	Mpumalanga	0.06324	1.06528	0.02459	2.57	0.01012
	North West	0.35412	1.42492	0.02315	15.30	0.00000
	Northern Cape	0.33617	1.39958	0.03449	9.75	0.00000
	Western Cape	-0.50967	0.60070	0.02727	-18.69	0.00000
Year	2009 (Ref)					
	2010	-0.10851	0.89717	0.02261	-4.80	0.00000
	2011	-0.18342	0.83242	0.02300	-7.98	0.00000
	2012	-0.16914	0.84439	0.02307	-7.33	0.00000
	2013	-0.17750	0.83736	0.02344	-7.57	0.00000
	2014	-0.24684	0.78126	0.02459	-10.04	0.00000
	2015	-0.27756	0.75763	0.02844	-9.76	0.00000
Gender* Province	(Ref:Female)* (Ref:Limpopo)					
	Male * Eastern Cape	-0.00281	0.99719	0.01883	-0.15	0.88126
	Male * Northern Cape	-0.03220	0.96831	0.02688	-1.20	0.23099
	Male * Free State	0.03453	1.03513	0.01919	1.80	0.07196
	Male * KwaZulu-Natal	0.01928	1.01947	0.01621	1.19	0.23427
	Male * North West	0.00813	1.00816	0.01843	0.44	0.65938
	Male* Gauteng	0.04268	1.04360	0.01591	2.68	0.00731
	Male * Mpumalanga	0.03022	1.03068	0.01980	1.53	0.12690
	Male *Western Cape	-0.02741	0.97296	0.02125	-1.29	0.19720
Gender*Year	Ref:Female* Ref:2009					
	Male * 2010	-0.03444	0.96614	0.01450	-2.38	0.01749
	Male * 2011	-0.03886	0.96189	0.01507	-2.58	0.00991
	Male * 2012	-0.01391	0.98619	0.01527	-0.91	0.36241
	Male * 2013	-0.01390	0.98620	0.01561	-0.89	0.37328
	Male * 2014	-0.02822	0.97217	0.01641	-1.72	0.08548
	Male * 2015	-0.01132	0.98875	0.01873	-0.60	0.54573
Province*Year	Ref:Limpopo* Ref:2009					
	Eastern Cape * 2010	0.01208	1.01215	0.03118	0.39	0.69849
	Northern Cape* 2010	-0.02531	0.97501	0.04631	-0.55	0.58474
	Free State * 2010	-0.04114	0.95970	0.03101	-1.33	0.18462
	KwaZulu-Natal* 2010	-0.03535	0.96527	0.02683	-1.32	0.18755
	North West * 2010	-0.02386	0.97643	0.03075	-0.78	0.43791
	Gauteng* 2010	-0.05125	0.95004	0.02661	-1.93	0.05410
	Mpumalanga* 2010	-0.03763	0.96307	0.03272	-1.15	0.25014
	Western cape * 2010	0.05841	1.06015	0.03607	1.62	0.10541

Table 2.7: Continuation of the results of the Cox PH on the effect of covariates on under five mortality

Factors	Level	**Coef (β)**	**Hazard ratio**	$S_e(\beta)$	***z***	***p-value***
	Ref: Limpopo* Ref:2009					
	Eastern Cape * 2011	-0.04142	0.95942	0.03200	-1.29	0.19548
	Northern Cape* 2011	-0.02382	0.97646	0.04735	-0.50	0.61494
	Free State * 2011	-0.10825	0.89740	0.03206	-3.38	0.00074
	KwaZulu-Natal * 2011	-0.10181	0.90320	0.02754	-3.70	0.00022
	North West* 2011	-0.07632	0.92652	0.03162	-2.41	0.01579
	Gauteng * 2011	-0.10916	0.89659	0.02726	-4.00	0.00000
	Mpumalanga * 2011	-0.14107	0.86843	0.03377	-4.18	0.00000
	Western cape* 2011	0.04372	1.04469	0.03702	1.18	0.23767
	Eastern Cape * 2012	-0.11293	0.89321	0.03243	-3.48	0.00050
	Northern Cape* 2012	0.05773	1.05943	0.04622	1.25	0.21168
	Free State * 2012	-0.28413	0.75267	0.03300	-8.61	0.00000
	Kwazulu-Natal* 2012	-0.14263	0.86708	0.02775	-5.14	0.00000
	North West * 2012	-0.08558	0.91798	0.03168	-2.70	0.00690
	Gauteng * 2012	-0.17260	0.84147	0.02745	-6.29	0.00000
	Mpumalanga * 2012	-0.14403	0.86586	0.03376	-4.27	0.00000
	Western cape * 2012	0.02267	1.02293	0.03721	0.61	0.54245
	Eastern Cape * 2013	-0.04640	0.95466	0.03272	-1.42	0.15616
	Northern Cape* 2013	0.10312	1.10863	0.04645	2.22	0.02642
	Free State * 2013	-0.29896	0.74159	0.03389	-8.82	0.00000
	KwaZulu-Natal* 2013	-0.22368	0.79957	0.02845	-7.86	0.00000
Province * Year	North West * 2013	0.04665	1.04776	0.03209	1.45	0.14600
	Gauteng * 2013	-0.24586	0.78203	0.02791	-8.81	0.00000
	Mpumalanga * 2013	-0.15836	0.85354	0.03477	-4.55	0.00000
	Western cape * 2013	0.09115	1.09543	0.03765	2.42	0.01547
	Eastern Cape * 2014	-0.07528	0.92749	0.03454	-2.18	0.02932
	Northern Cape* 2014	0.03305	1.03360	0.04939	0.67	0.50340
	Free State * 2014	-0.28979	0.74842	0.03552	-8.16	0.00000
	KwaZulu-Natal* 2014	-0.22973	0.79475	0.03001	-7.65	0.00000
	North West * 2014	-0.02276	0.97750	0.03413	-0.67	0.50489
	Gauteng * 2014	-0.17787	0.83706	0.02918	-6.09	0.00000
	Mpumalanga * 2014	-0.10795	0.89767	0.03609	-2.99	0.00278
	Western cape* 2014	0.09908	1.10415	0.03916	2.53	0.01141
	Eastern Cape* 2015	-0.13609	0.87277	0.04116	-3.31	0.00095
	Northern Cape* 2015	0.09171	1.09604	0.05609	1.64	0.10204
	Free State * 2015	-0.26621	0.76628	0.04177	-6.37	0.00000
	KwaZulu-Natal* 2015	-0.09528	0.90912	0.03466	-2.75	0.00598
	North West * 2015	0.13444	1.14390	0.03862	3.48	0.00050
	Gauteng * 2015	-0.00659	0.99343	0.03324	-0.20	0.84292
	Mpumalanga * 2015	-0.04666	0.95442	0.04220	-1.11	0.26895
	Western cape * 2015	0.41035	1.50735	0.04260	9.63	0.00000

The results show that gender, year and province were significantly associated with survival of under five children (p-value < 0). The results further revealed that male children were more likely to die as compared to female children. A hazard of dying of a male child under the age of five was 15% higher than that of a female child under the age of five ($HR = 1.148,$ p-value $= 0$). Children born in Eastern Cape, KwaZulu Natal and Western Cape were less likely to die as compared to children born in Limpopo. Children born from other provinces

such as Free State, Gauteng, Mpumalanga, North West and Northern Cape were more likely to die than those born in Limpopo. As an example, children born in Free State were 2.054 times more likely to die as compared to children born in Limpopo.

Children born between 2010 and 2015 were less likely to die than children born in 2009. We see a reduced risk of under five mortality between 2010 and 2015 as compared to 2009. The results suggest that there is no significant interaction between gender and birth province and birth province and birth year. That is, survival times for children under five years do not depend on the interaction between gender of a child and the province of birth and also not on the interaction between province of birth and year of birth.

2.3.5 Results from the logistic regression model

The logistic regression was considered to check factors influencing the probability of stillbirths. The overall effects of the variables was checked via the likelihood ratio test.
Table 2.8 shows the results of the overall effect of the variables.

Table 2.8: Overall effect of the variables

Factors	*df*	Deviance	Pr(>Chi)
Gender	1	709.50	0.0000
Province	8	9011.70	0.0000
Year	6	308.05	0.0000
Gender * Province	8	27.16	0.0007
Gender * Year	6	6.6973	0.3498
Province * Year	48	330.88	0.0000

The results in Table 2.8 show that the likelihood ratio test is highly significant for all the variables except for gender and year interaction. We conclude that gender, province, year, gender and province interaction, province and year interaction should remain in the model. The logistic regression model excluding

gender and birth year interaction is given in Table 2.9 and Table 2.10.

Table 2.9: Results of the logistic regression model

Factors	Level	Estimate	Se	z	p
Gender	(Intercept)	-5.045288	0.036655	-137.642	0.0000
	Ref: Female				
	Male	0.162109	0.022411	7.234	0.0000
Province	Ref: Limpopo				
	Eastern Cape	-0.482077	0.058283	-8.271	0.0000
	Northern Cape	0.429674	0.074568	5.762	0.0000
	Free state	1.110188	0.048575	22.855	0.0000
	KwaZulu natal	0.712860	0.041471	17.189	0.0000
	North West	0.446087	0.051250	8.704	0.0000
	Gauteng	0.643450	0.042246	15.231	0.0000
	Mpumalanga	0.439748	0.050970	8.628	0.0000
	Western Cape	0.522184	0.047717	10.943	0.0000
Year	(Ref: 2009)				
	2010	0.208946	0.046333	4.510	0.0000
	2011	0.132755	0.046665	2.845	0.00444
	2012	0.293395	0.045151	6.498	0.0000
	2013	0.390256	0.044348	8.800	0.0000
	2014	0.449826	0.043851	10.258	0.0000
	2015	0.406048	0.044629	9.098	0.0000
Gender*Province	Ref:Female* Ref:Limpopo				
	Male * Western Cape	-0.025873	0.030270	-0.855	0.39270
	Male * Eastern Cape	0.025802	0.037160	0.694	0.48747
	Male * Northern Cape	0.082439	0.045977	1.793	0.07296
	Female * Free State	0.081995	0.031844	2.575	0.01003
	Male * Kwazulu-Natal	-0.011037	0.026087	-0.423	0.67224
	Male * North West	0.077268	0.032658	2.366	0.01798
	Male * Gauteng	0.012443	0.026488	0.470	0.63853
	Male * Mpumalanga	0.032801	0.033201	0.988	0.32318
Province*Year	Ref:Limpopo* Ref:2009				
	Western Cape* 2010	-0.027371	0.060455	-0.453	0.65073
	Eastern Cape * 2010	0.088294	0.072909	1.211	0.22589
	Northern Cape* 2010	0.185386	0.091550	2.025	0.04287
	Free State * 2010	-0.023857	0.060944	-0.391	0.69546
	Kwazulu-Natal* 2010	-0.069508	0.052530	-1.323	0.18577
	North West * 2010	-0.044099	0.064837	-0.680	0.49641
	Gauteng * 2010	-0.074464	0.053501	-1.392	0.16397
	Mpumalanga * 2010	-0.114642	0.064997	-1.764	0.07776

Table 2.10: Continuation of the results of the logistic regression model

Covariate	Level	Estimate	Se	z	p
	Ref:Limpopo* Year Ref:2009				
	Eastern Cape * 2011	0.031461	0.074286	0.424	0.67193
	Northern Cape* 2011	0.133446	0.093496	1.427	0.15350
	Free State * 2011	-0.039856	0.061866	-0.644	0.51943
	KwaZulu-Natal * 2011	-0.065827	0.053048	-1.241	0.21464
	North West* 2011	-0.022891	0.065482	-0.350	0.72665
	Gauteng * 2011	-0.051534	0.053952	-0.955	0.33949
	Mpumalanga * 2011	-0.100980	0.065399	-1.544	0.12257
	Western Cape * 2011	0.028224	0.060895	0.463	0.64302
	Eastern Cape * 2012	-0.100354	0.073186	-1.371	0.17030
	Northern Cape* 2012	0.039336	0.091370	0.431	0.66682
Province * Year	Free State * 2012	-0.351825	0.061757	-5.697	0.0000
	KwaZulu-Natal* 2012	-0.215328	0.051756	-4.160	0.0000
	North West* 2012	-0.073887	0.063636	-1.161	0.24561
	Gauteng * 2012	-0.143556	0.052421	-2.739	0.00617
	Mpumalanga * 2012	-0.221704	0.063983	-3.465	0.00053
	Western Cape * 2012	-0.166434	0.059997	-2.774	0.00554
	Eastern Cape * 2013	-0.114524	0.072178	-1.587	0.11258
	Northern Cape* 2013	-0.029784	0.090659	-0.329	0.74251
	Free State * 2013	-0.428171	0.061392	-6.974	0.0000
	KwaZulu-Natal* 2013	-0.210854	0.050764	-4.154	0.0000
	North West * 2013	-0.006759	0.062784	-0.108	0.91427
	Gauteng * 2013	-0.325210	0.051827	-6.275	0.0000
	Mpumalanga * 2013	-0.322730	0.064013	-5.042	0.0000
	Western Cape * 2013	-0.106324	0.058795	-1.808	0.07055
	Eastern Cape * 2014	-0.183966	0.071917	-2.558	0.01053
	Northern Cape* 2014	-0.172610	0.091535	-1.886	0.05933
	Free State * 2014	-0.505864	0.061046	-8.287	0.0000
	KwaZulu-Natal* 2014	-0.373604	0.050685	-7.371	0.0000
	North West *2014	-0.146229	0.063017	-2.320	0.02032
	Gauteng * 2014	-0.384068	0.051484	-7.460	0.0000
	Mpumalanga *2014	-0.381209	0.063500	-6.003	0.0000
	Western Cape * 2014	-0.256285	0.058717	-4.365	0.0000
	Eastern Cape * 2015	-0.091094	0.059136	-2.146	0.03188
	Northern Cape* 2015	-0.077659	0.091683	-0.847	0.39697
	Free State * 2015	-0.509720	0.062752	-8.123	0.0000
	KwaZulu-Natal*2015	-0.225227	0.051441	-4.378	0.0000
	North West * 2015	-0.162133	0.064514	-2.513	0.01197
	Gauteng * 2015	-0.270924	0.052174	-5.193	0.0000
	Mpumalanga * 2015	-0.286693	0.064665	-4.434	0.0000
	Western Cape* 2015	-0.126899	0.059136	-2.146	0.03188

Results obtained from logistic regression analysis indicate that being a still-born is influenced by gender, province of birth and year of birth. The three predictors were found to be statistically significant at 0.05 level of significance. The results further revealed that male children were more likely to be born

as stillbirths than female children (Coef = 0.162). The coefficient for Eastern Cape province is negative (-0.482). This means that children born in Eastern Cape province were less likely to be born as stillbirths than children born in Limpopo province. With regard to birth year, those who were born after 2009 were more likely to be stillbirths than those who were born in 2009.

2.4 Discussion

Findings from this chapter are in line with findings from other countries in Sub-Saharan Africa. Our findings revealed a higher percentage of male stillbirths as compared to female stillbirths. These findings agree with those in (Madhi et al., 2019). The probability of stillbirths for both males and females were highest in Free State, a rural province in South Africa. A paper published by Bhattacharyya and Pal (2012) showed also that rural residence contributes to the risk of stillbirths. The logistic regression results revealed that gender, birth province and birth year were factors influencing probability of stillbirths. Although Feresu et al. (2004) found no statistical difference between the risk of stillbirths in female and male children, Madhi et al. (2019) found gender to be a factor associated with stillbirth in other developing countries. A study done by Graner et al. (2009) also showed that place of residence influence stillbirths. Women who delivered in rural areas were at higher risk of stillbirth compared to those delivered in urban areas due to unavailability of health facilities and trained health practitioners.

The study also investigated potential risk factors for under five mortality using the Cox model. Gender was found to be significant factor of under five mortality. Male children had higher risks of death than female children. These findings agreed with those in (Nasejje, 2013) and Ezeh et al. (2015). Another

factor contributing to under five mortality was found to be birth province. Results obtained by Zike et al. (2018) and Ezeh et al. (2015) also showed that residential area influence the survival of under five children. A study done by Worku (2009) showed that in South Africa, residential area plays a role in child survival and that most of the disadvantaged children are those residing in rural areas. Although his study was done in South Africa, his focus was not on clustered survival models and he also focused on different variables other than what we have used in our study. His focus was on variables such as education level of mother, accessibility to clean water, income status of the family, breast feeding duration, family planning methods usage, age and marital status of mother. The K-M and predicted probability plots showed that children born in Western Cape had better survival rates compared to other children. This might be due to the fact that Western Cape is much more advanced in terms of resources, number of well-trained health practitioners and access to health care facilities. It has also been found that children born in 2014 had the highest survival rate compared to children born in other years.

2.5 Summary of the chapter

The objective of this chapter was to identify factors influencing probability of stillbirths and to investigate predictors of under five child mortality. Logistic regression model was applied to data set 1 described in Chapter 1, Section 1.7.4.1 to identify factors influencing probability of stillbirths. Three factors contributing to stillbirth cases were found to be gender, birth province and birth year. K-M curves and log-rank test were used to compare survival distributions of two or more categories of covariates included in the analysis. The results of the Log-rank test showed significant differences in the occurrence of death of different categories of gender, birth province and birth year. The K-M curves revealed that female children survived longer than male children

and that those born in Western Cape survived longer than children in other provinces. The results further revealed higher survival of children born in 2014. The Cox PH model was applied to identify important predictors of under five mortality. The results showed that under five mortality was significantly influenced by three covariates, namely: gender, birth province and birth year.

Chapter 3

Shared frailty model for left truncated survival data

So far we have seen time to event data analysis without taking into account clustering and left truncation in all analyses. It is important to consider these two aspects now. In this chapter, the importance of considering clustering and left truncation is demonstrated.

3.1 General introduction

In many studies, natural clustering of subjects exists such that survival times within the same cluster may be correlated because of certain features such as shared environmental factors or genetic factors. Data containing clustered survival times also emanate from data involving multiple occurrences of an event from the same person, i.e., where an individual experiences the same type of event such as suicide attempts, more than once (Knox et al., 2013). These studies involve clusters of subjects or individuals who have common factors that might influence the results of interest. As an example, patients treated at

the same medical institution such as hospital are likely to have similar results than those treated at other medical institutions. Another example is that there may be an association in the times to events of a disease between children born from the same mother. Individuals within a group or cluster are likely to have similar results than subjects in other groups. There are two techniques that consider dependence between observations within a cluster that can be used to analyse this type of clustered survival data. These are frailty and copula models. Both models can give us estimates of the association between survival times in a cluster. The disadvantage of the marginal regression model approach such as the Cox model applied in Chapter 2 is that it does not provide us with information concerning the association between survival times of individuals in a cluster. The focus in this chapter is on the frailty model. The copula model is discussed in the next chapter, i.e., Chapter 4.

The main aim of this chapter is to analyse factors affecting under five child mortality taking consideration heterogeneity present in the data and left truncation. A shared frailty model using the gamma distribution as the choice of the frailty distribution is explored. In the sections that follow, introduction to frailty models, descriptive statistics, methodology and estimation methods are described. All analyses were done using data set 2 described in Chapter 1 Section 1.7.4.2. Clusters of size one were excluded in the analysis because they were in majority. For each cluster of size 1, we introduced a frailty term and it was not possible to include all of them in the model. Another reason is that our main focus was on links between siblings.

3.1.1 Introduction to frailty models

Frailty models are widely used to model association between failure times of the same cluster by including a random component called frailty term for the

hazard function in the model. Frailty can be defined as susceptibility to a certain event which can be individual or shared by different individuals in a cluster (Goethals, 2011). Frailty models are also called conditional models because covariate effects are specified conditionally on the frailty term. In frailty modelling, the assumption is that different members within the same cluster or group are related to each other conditionally on the frailty term(s). Conditional on the frailty, the survival times of children within a cluster are assumed to be independent. A frailty model is often used when we are interested in the association between failure times within the same cluster (Andersen, 2005). The purpose of the frailty model is to take care of the heterogeneity caused by covariates which were related to the event of interest, but were not measured. The likelihood is constructed by first considering the conditional contribution of a member within a cluster to the likelihood function and then integrate over the frailty distribution. The interpretation of covariate effect is at the conditional level. There are different frailty models, namely, univariate, correlated and shared frailty models.

In a univariate frailty model, each member in the study has its own frailty term and we assume that members who are most frail will experience event of interest at an earlier time than others. In the correlated frailty models, each member in the cluster has its own frailty term and that results to event times of members of the same cluster being correlated. In a shared frailty model, members belonging to the same cluster share the same frailty. In all these models, the frailty term is a positive random variable following some distribution such as gamma, positive stable or lognormal. Children in this study came from different households, raised by different mothers and hence clustered at mother's level. A shared frailty model is the appropriate model in this study. The frailty term in our study will take into consideration the situation where some of the children may be exposed to hazard of death more than others.

3.1.2 Left truncation

Left truncation in survival data exists when some of the individuals in the data are not observed from the time origin of interest (Jensen et al., 2004). In this study, death information for children who died between 2010 and 2012 was not captured in the two data sets provided by Stats SA and that resulted to left truncated data. The reasons for the missing information were not provided. The recording of death information was problematic because death information was recorded for only those who died between 2013 and 2015. Due to missing death information in this data set, survival models that consider left truncation need to be used in order to get reliable results. Individuals with left truncated survival times have survived until entry period and will carry information about their frailty value and this needs to be considered in the analysis (Jensen et al., 2004). When left truncation exists in the data set, the frailty distribution among those who survived in a cluster is different because the distribution will tend to have lower frailty values (Jensen et al., 2004).

3.1.3 Consequence of ignoring frailty

Analyses that ignore associations will overestimate the variability (Sainani, 2010). Ignoring heterogeneity in the data will produce incorrect estimated parameters (Abdulkarimova, 2013). Previous investigators such as Zhenzhen (2000), Moerbeek et al. (2003) and Islam et al. (2010), ignored association within clusters in their studies. On the other hand, researchers such as Vaupel et al. (1979), Guo and Rodriguez (1992), Sastry (1997), and Mahmood et al. (2013) have recognised that ignoring association between related individuals in the survival studies would lead to estimates that are inefficient and biased. It was shown by Bouwmeester et al. (2013) that a model developed with a random effect exhibited better discrimination than the standard logistic regression ap-

proach, if the cluster effects were used for risk prediction. It was also shown by Sainani (2010) that the application of many statistical tests to correlated observations will lead to the overestimation of p values in certain cases when one considers within-subject or within-cluster effect and underestimation in others when one considers between-cluster effects. Including the unobserved frailty term in the model avoids underestimation or overestimation of the model parameters (Gachau, 2014).

3.2 Methodology

In this section, the Cox PH model and the gamma shared frailty models with parameters estimated by penalised likelihood maximisation will be described.

3.2.1 The Cox PH model

The Cox PH model discussed in detail in Section 2.2.3 was analysed using *frailtypack* package. With *frailtypack*, it is possible to fit Cox PH model with parameters estimated by penalised likelihood estimation (Rondeau et al., 2012). The aim was to compare the Cox PH model and the shared frailty model for their performance. The Cox PH model is the same as the frailty model without including a random effect. This model can be extended to include a random effect term through a linear component considers the unobserved heterogeneity. The extended model is called the frailty model (Niragire et al., 2011).

3.2.2 Univariate frailty model

To illustrate the shared frailty model, let T be the survival time of an individual and Z be the frailty variable. The conditional hazard function for a given frailty variable $Z = z$ at time $t > 0$ is given by:
$h(t|z) = zh_0(t)\mathbf{exp}(X\beta)$.

In this model, $h_0(t)$ is the baseline hazard function and β is the column vector of regression coefficients. The conditional survival function for a given frailty at time $t > 0$ is given by

$$\begin{aligned} S(t|z) &= \exp\left[\int_0^t h(x|z)dx\right] \\ &= \exp[-zH_0(t)\exp(X\beta)]. \end{aligned} \tag{3.1}$$

The marginal survival function can be obtained by integrating over the range of frailty variable z as follows:

$$\begin{aligned} S(t) &= \int_0^\infty S(t|z)f(z)dz \\ &= \int_0^\infty \exp[-zH_0(t)\exp(X\beta)]f(z)dz \\ &= L[H_0(t)\exp(X\beta)], \end{aligned} \tag{3.2}$$

where $L(.)$ is the Laplace transformation of the distribution of Z.

3.2.3 The shared frailty model

In this section, the shared frailty model with a gamma distribution of the random effect is discussed. The model with the gamma distribution is very easy to fit and interpret in terms of the hazard ratios (Guo and Rodriguez, 1992).

In the shared frailty model, it is assumed that individuals belonging to the same cluster share the same frailty term (Duchateau and Janssen, 2007). This frailty term tells us that members in the same group behave in a similar but unknown manner. A shared frailty model can be regarded as a random effects model with two sources of variations, namely, cluster variation which is described by the frailty and individual variation described by the hazard function (Hougaard, 2012). A random effect is introduced for each cluster so that individuals from one cluster are more similar than individuals from different clusters. The random effect describes the unobserved influences common to all

individuals of that particular cluster (Gachau, 2014). The main assumption of shared frailty model is that all items or individuals in a cluster share the same value of frailty, which is the reason why the model is called shared frailty model. It is assumed that the frailties in different clusters are not related. Failure times of individuals in a cluster are dependent, while those across clusters are independent (Nguti, 2003). Frailty is assumed to be independent across the groups or clusters, while the survival times of individuals within the same group are conditionally dependent (Nasejje, 2013). The aim of the frailty model is to take care of the heterogeneity caused by variables not measured.

Let us assume that we have a total of n individuals coming from k different clusters. Let L_{ij} be the left truncated times. For each individual in the study, we observe $Y_{ij} = min(T_{ij}, C_{ij})$ and indicators showing censoring status $\delta_{ij} = I_{T_{ij} \leq C_{ij}}$. T_{ij} are survival times and C_{ij} the censoring times of individuals under study. The survival times are said to be left truncated in a situation where only individuals with $T_{ij} > L_{ij}$ are observed.

The shared frailty model which specifies that the hazard function conditional on the frailty is given by:

$$h_{ij}(t|z_i) = z_i h_0(t) exp(\beta^{'} X_{ij}), \tag{3.3}$$

where:
$h_0(t)$ is the baseline hazard at time t, X_{ij} is a vector of covariates for individual j in cluster i, β is a vector of regression coefficients, and z_i's are the frailties which are independently and identically distributed from a gamma distribution which has a mean of 1 and unknown variance θ. Individuals in cluster i share frailty z_i, and conditional on z_i their survival times are assumed to be independent.
The frailty for cluster i gives us an idea on how the hazard for that cluster

varies from the hazard of an event in general. Large values for the frailty, i.e., $z_i > 1$ indicate that the event is happening earlier as compared to clusters with $z_i < 1$ (Legrand et al., 2006). Subjects with a larger frailty are more frail and expected to die earlier than subjects with the same measured covariates. In case the event of interest is a positive outcome, like for example, recovery from a disease, individuals with larger frailty values are expected to experience the positive outcome earlier than others with the same covariates. In other words, they are expected to heal faster than others (Balan and Putter, 2020). In shared frailty models, the correlation between frailties of different groups is equal to zero and the correlation between subject frailties of members in a group is equal to one (Goethals et al., 2008).

In this study, an R package called *frailtypack* was used to fit Cox and shared frailty models. Estimation of unknown parameters in both models were done using the full penalised likelihood approach. *Frailtypack* can be used to estimate the parameters in a shared gamma frailty model with possibly right censored, left truncated and stratified survival data (Rondeau and Gonzalez, 2005). For a detailed description of *frailtypack*, see Rondeau and Gonzalez (2005) and Rondeau et al. (2012).

3.2.4 Heterogeneity parameter θ

It crucial to detect the presence of cluster effects or heterogeneity in a clustered survival data. When we estimate the heterogeneity parameter (variance θ), we get a better idea of the heterogeneity of the values of the random effect between clusters. The stratified Cox model, in which cluster to-cluster variability is treated as nuisance, does not provide a framework for describing heterogeneity (Glidden and Vittinghoff, 2004).

In a frailty model, θ is estimated to get an idea on heterogeneity in the outcome between clusters (Gachau, 2014). The variance parameter θ measures the degree of between-cluster variability (Glidden and Vittinghoff, 2004). If θ equals zero then there is no evidence of frailty, and that the frailty component does not contribute to the model (Mills, 2011). Thus, when $\theta = 0$, the frailties are independently equal to 1 and in that case the cluster effects are absent, and events are independent within and across clusters. This will be the same as using the Cox Proportional hazard model. On the other hand, a value of θ greater than 0 reflects heterogeneity between clusters and a strong association between members of the same cluster. Large values of θ reflect a greater degree of heterogeneity (Zike et al., 2018). The larger the value of θ, the larger would be the heterogeneity in outcome between clusters (Legrand et al., 2006).

3.2.5 Choice of frailty distribution

The distribution of the random effect can be chosen among several distributions (Mauguen, 2014). The most common ones are the gamma and log-normal distributions. Even though the gamma models do not have closed forms expressions for survival and hazard functions, from a computational view, it fits well to frailty and it is easy to derive the closed form expressions for unconditional survival and hazard functions (Zike et al., 2018). In this thesis we used the gamma frailty as the distribution for the random effect with baseline hazard estimated using the penalised likelihood approach as proposed by Rondeau and Gonzalez (2005). The popularity of a gamma distribution is based on mathematical and computational aspects that it has a simple Laplace transform and that makes inference less complicated (Goethals et al., 2008).

Suppose that T is a gamma distributed random variable with scale and shape

parameters b and α, respectively. The probability density function is given by:

$$f(t) = \frac{b^{\alpha} t^{\alpha-1} e^{-bt}}{\Gamma(\alpha)}, \tag{3.4}$$

where $\Gamma(\alpha)$ is the gamma function given by:

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} dt.$$

Parameters b and α are greater than 0. The expected value and variance of the gamma distribution are as follows:

$$E(T) = \frac{\alpha}{b}$$

and

$$Var(T) = \frac{\alpha}{b^2}$$

respectively.

The survival and hazard functions are given by:

$$S(t) = 1 - I_{\alpha}(bt)$$

and

$$h(t) = \frac{b^{\alpha} t^{\alpha-1} e^{-bt}}{1 - I_{\alpha}(bt).\Gamma(\alpha)}$$

respectively.
$I_{\alpha}(x)$ is the incomplete gamma function defined as

$$I_{\alpha}(x) = \frac{\int_0^x b^{\alpha-1} e^{-x} dx}{\Gamma(\alpha)}.$$

In a gamma frailty model, the restriction $\alpha = b$ is used which resulted to ex-

pectation of 1. The variance of the frailty variable is then $1/b$. Assume that the frailty term Z is distributed as gamma with $E(Z) = 1$ and $Var(Z) = \theta$. Then $b = \alpha = 1/\theta$. The probability density function is given by

$$f(z) = \frac{z^{(1/\theta-1)}exp(-z/\theta)}{\Gamma(1/\theta)\theta^{1/\theta}} \tag{3.5}$$

and the Laplace transform is given by

$$\mathcal{L}(s) = (1 + \theta s)^{-1/\theta}. \tag{3.6}$$

One way to express dependence in a frailty model is to use Kendall's tau to quantify dependence. Kendall's tau measures the dependence between any two event times from the same cluster. For a gamma distribution, Kendall's tau is expressed as:

$$\tau = \frac{\theta}{\theta + 2} \, \epsilon \, (0, 1).$$

3.3 Estimation methods for the shared frailty models

Depending on the assumptions made about the baseline hazard and the distribution of the random effect, various likelihood-based procedures have been proposed to estimate the variance of the random effect (Legrand et al., 2006). In this section the estimation methods for the shared frailty model are discussed. Drawbacks of common methods used previously by other authors are discussed as well as the Penalised likelihood estimation method used in this thesis. There are several methods that can be used to estimate unknown parameters in a frailty model. These include maximum likelihood methods such as Expectation-Maximisation (EM) algorithm and penalised likelihood method.

The maximum likelihood estimation in the gamma frailty model is straightforward as we can easily integrate the frailties out in the likelihood function and obtain the parameter estimates using classical maximum likelihood techniques (Abdulkarimova, 2013). In this thesis, the penalised full likelihood method was used.

3.3.1 Expectation-Maximisation (EM) algorithm

The EM algorithm is a method to find maximum likelihood estimates of parameters in a statistical model where the model depends on unobserved latent variables. The approach consists of two steps namely: an Expectation step (E-step) and a maximisation step (M-step). In the E-step, the expected values of the unobserved frailties conditional on the unobserved information and current estimates are obtained. In the M-step, these expected values of the parameters of interest are obtained by maximisation of the likelihood given the expected values (Duchateau and Janssen, 2007). The drawbacks of this approach were highlighted by Therneau and Grambsch (2013), i.e. the algorithm is slow and proper variance estimates need further computation. Rondeau et al. (2012) noted that no implementation has appeared in any of the more widely available packages.

3.3.2 The (partial) penalised likelihood method

An alternative estimation method to the EM algorithm is the penalised likelihood approach where the random effect is treated as a penalty term. The penalised likelihood approach is favoured over the EM algorithm because it is faster and is implemented in most standard software (Gachau, 2014). The penalised partial likelihood method proposed by Therneau and Grambsch (2013) has some drawbacks, i.e, the convergence can be slow, direct estimate of the variance of the frailty term is not provided and that the method cannot be used

to estimate a hazard function (Rondeau and Gonzalez, 2005).

Due to the drawbacks of the two approaches mentioned above, i.e., EM algorithm and penalised partial likelihood, we decided to use an alternative method, a non-parametric penalised full likelihood approach. The use a non-parametric penalised likelihood and the smooth estimation of the baseline hazard function is provided by using an approximation by splines. Penalised technique is a useful estimation tool (Therneau and Grambsch, 2013). The difference between partial penalised likelihood by Rondeau et al. (2012) and full penalised likelihood approach is that the baseline hazard function is penalised in full penalised approach and the frailties are penalised in partial penalised approach.

The full log-likelihood for left-truncated and right censored data in the shared gamma-frailty models are given by:

$$\begin{aligned} l(h_0(.),\beta,\theta) = \sum_{i=1}^{k}\{\sum_{j=1}^{n}\delta_{ij}\{\beta^{'}X_{ij} + ln(h_0(Y_{ij}))\} \\ - (1/\theta + m_i)ln[1+\theta\sum_{j=1}^{n}H_0(Y_{ij})exp(\beta^{'}X_{ij})] \\ + 1/\theta ln(1+\theta\sum_{j=1}^{n}H_0(L_{ij})exp(\beta^{'}X_{ij})) \\ + I_{m_i\neq 0}\sum_{p=1}^{mi}[ln(1+\theta(m_i-p))]\}, \end{aligned} \quad (3.7)$$

where $H_0(.)$ is the cumulative baseline and $m_i = \sum_{j=1}^{Ji} I_{(\delta_{ij=1})}$ is the total number of observed events in the i^{th} group. The penalised log-likelihood is then defined as:

$$pl(h_0(.), \beta, \theta) = \iota(h_0(.), \theta) - \kappa \int_0^\infty h_0^{''2}(t)dt, \tag{3.8}$$

where $l(h_0(.), \beta, \theta)$ is the full log-likelihood defined in 3.7. $\kappa \geq 0$ is a positive smoothing parameter. Maximum penalised likelihood estimators $\hat{h}_0(t), \hat{\beta}, \hat{\theta}$ can be found by maximising equation 3.8. $\hat{H}^{-1}$ can be used as a variance estimator of the parameters, where $\hat{H}$ is the converged Hessian of the penalised log-likelihood.

3.3.2.1 Approximation with splines

A baseline hazard function $h_0(t)$ is estimated using splines after estimating the covariate coefficients in the shared frailty model. The estimator of the baseline hazard $h(.)$ can be approximated on the basis of splines with Q knots: $\tilde{h}_0(.) = \sum_{i=1}^{m} \eta_i M_i(.)$, where $m = Q + 2$, Q is the number of knots, $\eta_i^{'}s$ are control points of increasing knots and M_i represents the cubic M-splines (Rondeau et al., 2012). In frailtypack, we set a knot on the first and last data values and the other knots are put in such a way that the distance is equal between them. It is advisable to start with a small number of knots and then gradually increase the number until the graph of baseline hazard function remains unchanged. The more the number of knots used, the longer the time of running. According to Rondeau et al. (2012), the suggested number of knots should be between 4 and 20. Initial values are needed in most of the algorithms in the frailtypack programs. It is important to choose good initial values to maximise the penalised likelihood because the convergence is faster if the initial value is closer to the true value. In our case, the splines, regression coefficients and variance of the frailty term were initialised to 0.1.

3.3.2.2 Estimation of smoothing parameter κ

The maximisation of a likelihood cross-validation criterion is used for estimating the smoothing parameter κ (Tang, 2014). To find the smoothing parameter

κ, we minimise the function

$$\bar{V}(k) = \frac{1}{Q}[tr(\hat{H}_{pl}^{-1}\hat{H}_l - pl(\hat{\Phi}_k))], \tag{3.9}$$

where Q is the number of knots, $\hat{H}_{pl}$ is the converged Hessian matrix, $\hat{H}_l$ is the converged Hessian matrix of the log-likelihood and $pl(\hat{\Phi}_k)$ is the penalised log-likelihood.

3.4 Data analysis and results

In this section, the data set introduced in Section 1.7.4.2, excluding clusters of size one, is analysed.

3.4.1 Descriptive statistics

A total of 250260 children and 123110 mothers (clusters) were included in the analysis. A summary of descriptive statistics is presented in Table 3.1.

Table 3.1: Descriptive and summary statistics of the data

Covariate	Level	Total (%)	Death N (%)	Censored N (%)	Left truncated N (%)
Gender	Female	124901(49.9%)	2482(2.0%)	122419 (98.0%)	15282(50%)
	Male	125359 (50.1%)	2720(2.2%)	122639(97.8%)	15274(50%)
Province	Limpopo	29779(11.9%)	824(2.8%)	28955(97.2%)	3189(10.4%)
	Eastern Cape	33146(13.2%)	598(1.8%)	32548(98.2%)	4761(15.6%)
	Free State	11299(4.5%)	402(3.6%)	10897(96.4%)	997(3.3%)
	Gauteng	51328(20.5%)	1047(2.0%)	50281(98.0%)	4812(15.7%)
	KwaZulu Natal	59876(23.9%)	766(1.3%)	59110(98.7%)	9425 (30.8%)
	Mpumalanga	21073(8.4%)	458(2.2%)	20615(97.8%)	2705(8.9%)
	North West	15530(6.2%)	517(3.3%)	15013(96.7%)	2431(8.0%)
	Northern Cape	5782(2.3%)	249(4.3%)	5533(95.7%)	1703(5.6%)
	Western Cape	22447(9.0%)	341(1.5%)	22106(98.5%)	533(1.7%)
Year	2010	6887(2.8%)	11(0.2%)	6876 (99.8%)	6887 (22.5%)
	2011	8676 (3.5%)	24(0.3%)	8652(99.7%)	8676(28.4%)
	2012	14993 (6.0%)	189(1.3%)	14804(98.7%)	14993(49.1%)
	2013	103811 (41.5%)	3624(3.5%)	100187(96.5%)	0(0%)
	2014	12694 (5.1%)	191(1.5%)	12503(98.5%)	0(0%)
	2015	103199 (41.2%)	1163(1.1%)	102036(98.9%)	0(0%)
Twin	No	197956(79.1%)	3775(1.9%)	194181 (98.1%)	28781 (94.2%)
	Yes	52304 (20.9%)	1427(2.7%)	50877(97.3%)	1775(5.8%)
Order	0	147274(58.8%)	4073(2.8%)	143201 (97.2%)	27525 (90.1%)
	1	100424 (40.1%)	1098(1.1%)	99326(98.9%)	3000(9.8%)
	2	2504 (1.0%)	29(1.2%)	2475(98.8%)	31(0.1%)
	3	55 (0.0%)	2(3.6%)	53(96.4%)	0(0%)
	4	3 (0.0%)	0(0.0%)	3(100.0%)	0(0%)

Out of the total number of 250260 children, 5202 (2.1%) were dead and 245058 (97.9%) were still alive at the date of the survey. The mortality rates of children under the age of five varied from one South African province to another. The highest percentage of deaths was observed in Northern Cape (4.3%), followed by Free State (3.6%), while the lowest percentage of deaths was recorded in KwaZulu Natal (1.3%), followed by Western Cape (1.5%).

With regard to gender of these children, a higher percentage was observed among male children (2.2%) compared to female children (2.0%). Across birth year, the highest death rate (3.3%) was recorded in 2013 and the lowest (0.2%) was recorded in 2010.

Children born as a result of multiple births (twins) recorded the highest percentage of death compared to those who were not part of twins (singletons).

About 2.7% of the children in multiple births (twins) had died before reaching the age of five, compared to those born out as singletons, which recorded 1.9%. The majority of children included in the data set were censored. The death rate also varied by the number of previous children that a mother had (birth order). The lowest percentage of the death was when mothers had four previous living children (0.0%) than when they had no previous living children. This shows that the more children born previously to a mother, the more experience the mother had, and the lower the risk of dying.

With regard to left truncated individuals, 30556 (12.2%) were left truncated. Half of them were females and half were males. The highest percentage of left truncated individuals were observed in KwaZulu Natal province (30.8%) followed by Gauteng (15.7%) and closely by Eastern Cape (15.6%), and the lowest was in Western Cape (1.7%). Across birth year, by far the highest percentage of left truncated individuals (49.1%) was recorded in 2012. There were no left truncated individuals recorded between 2013 and 2015. This confirms the information given in Chapter 1, Section 1.7.4.2. We had more left truncated singletons (94.2%) in our data set as compared to twins (5.8%). With regard to previous number of living children that mothers had (birth order), it was found that the highest percentage of left truncated individuals (90.1%) was when mothers had no previous living children.

3.4.2 Results

The Cox Proportional model and shared gamma frailty models were analysed using a penalised likelihood on the hazard function. A p-value < 0.05 was considered statistically significant. The parameter estimates for the two models are presented in Table 3.2.

Table 3.2: Cox PH and shared frailty models with parameters estimated by penalised likelihood maximisation

Factors	Levels	Cox PH model				Shared frailty model			
		Coef	Hazard ratio	S_e	p-value	Coef	Hazard ratio	S_e	p-value
Gender	Female	Ref							
	Male	0.095	1.099	0.028	0.0007	0.096	1.100	0.029	0.0010
Province	Limpopo	Ref							
	Eastern Cape	-0.390	0.677	0.055	< 0.0000	-0.404	0.667	0.062	< 0.0000
	Free State	0.251	1.286	0.062	< 0.0000	0.268	1.308	0.070	0.0001
	Gauteng	-0.331	0.718	0.047	< 0.0000	-0.336	0.714	0.054	< 0.0000
	KwaZulu	-0.737	0.478	0.052	< 0.0000	-0.758	0.468	0.055	< 0.0000
	Mpumalanga	-0.221	0.802	0.058	0.0002	-0.226	0.798	0.066	0.0007
	North West	0.260	1.296	0.057	< 0.0000	0.270	1.310	0.066	< 0.0000
	Northern Cape	0.478	1.613	0.073	< 0.0000	0.517	1.678	0.083	< 0.0000
	Western Cape	-0.632	0.532	0.065	< 0.0000	-0.646	0.524	0.071	< 0.0000
Year	2010	Ref							
	2011	0.821	2.273	0.293	0.0051	0.824	2.279	0.497	0.0972
	2012	1.283	3.609	0.324	< 0.0000	1.288	3.626	0.192	< 0.0000
	2013	1.835	6.264	0.342	< 0.0000	1.883	6.577	0.175	< 0.0000
	2014	1.422	4.146	0.325	< 0.0000	1.446	4.247	0.197	< 0.0000
	2015	1.687	5.406	0.346	< 0.0000	1.723	5.603	0.180	< 0.0000
Twin		0.167	1.182	0.035	< 0.0000	0.180	1.197	0.037	< 0.0000
Order		-0.367	0.693	0.047	< 0.0000	-0.372	0.690	0.049	< 0.0000
θ (***P*-value**)			-					2.342 ($p < 0.0000$)	
Penalised marginal log-likelihood			-55833.07					-55742.73	
LCV			0.2232					0.2228	

Table 3.2 shows the potential risk factors contributing to high rate of under-five child mortality in South Africa. The Cox PH model and the shared frailty model parameters were fitted by penalisation approach suggested by (Rondeau et al., 2012). Factors that were expected to affect the survival of children were gender of a child, birth province, birth year, twin and birth order. A hazard rate of 1.1 was obtained for male children compared to female children. The chances of male children to die was found to be very high compared to those for female children. The hazard rates of Eastern Cape, Gauteng, KwaZulu Natal, Mpumalanga and Western Cape provinces were all less than 1, indicating that children residing in those provinces were less likely to die compared to those in Limpopo. The results also show that the hazard of death for children born between 2011 and 2015 were higher compared to the hazard of death for those born in 2010. This also confirms the descriptive statistics in Table 3.1.

We also compared the hazard rates of children who were part of twin with those who were singletons. Both models show that a child born as part of twin turns to have a probability of dying which is more than 1.1 times the probability of dying for a singleton. This means that children who were part of twins were more likely to die than those who were singletons. The results from the output

show that birth order was another significant factor of under five child mortality. The hazard ratio of children who were born from a mother with other children is 0.690. This means that the more children born previously to the mother, the lower the hazard of dying.

To test the importance of frailty effect in our model, we evaluated the value of theta shown in Table 3.2. The variance of the frailty term $\theta = 2.342$ with a p-value < 0 shows evidence of frailty, that heterogeneity was found between mothers of children and strong connection between children from the same mother (siblings). The results of the frailty term also show that there are other factors except the ones included in the model that affect high mortality of children under the age of five at mother's level. The shared frailty model showed a lower likelihood cross validation value ($LCV = 0.2228$) compared to the Cox Proportional hazard model ($LCV = 0.2232$). This shows that the shared frailty model was the most efficient model to describe the data set.

3.5 Discussion

The results of this study show that the death of a child under the age of five is contributed by gender, birth province, birth year, birth order and whether a child is part of twin or not. Male children were more likely to die than female children. These results agreed with other studies done in other countries such as in Bangladesh by Khan and Awan (2017), in Uganda by Nasejje (2013), in Ethiopia by Zike et al. (2018) and in Turkey by Seçkin (2009). It has been indicated by Khan and Awan (2017) that female children have a biological advantage against many causes of mortality than male children and that might be the reason of higher risk of male children deaths. Twins were more likely to die than singletons. These findings are similar to many previous researchers

such as (Zike et al., 2018). First born children from mothers were likely to die than second, third, fourth and fifth born children. This might be due to the experience of handling children mothers had. It can also be concluded from the results that the risk of dying is lower for children born in provinces like Eastern Cape, Gauteng, KwaZulu Natal, Mpumalanga and Western Cape compared to those born in Limpopo.

The value of θ shows evidence of the existence of unobserved heterogeneity at mother's level. This shows the presence of other factors contributing to the death of children which were not described by those factors included in the model. Although there is little variation of the parameter estimates in the two models, the regression estimates showed an increase in the shared frailty model where the frailty term was included. This was expected because the shared frailty model accounts for extra variance associated with risks not measured. We notice that the effects of covariates included in the model is biased downward when the effects of frailty are not considered.

Stone (1974) showed that LCV was asymptotically identical to AIC, but more flexible because it can be applied to penalised likelihood estimators. The LCV value of the shared frailty model was found to be minimum as compared to the Cox PH model, this indicates that the shared frailty model is the most efficient model to describe the under five child mortality data set. Although Cox PH model can describe the association between survival probabilities and covariates in the model, it does not take into consideration the unmeasured variability among individuals beyond that of measured covariates. Adding a frailty term made a significant contribution when we compare the penalised log-likelihoods for the Cox PH and the shared frailty models. Considering correlation among survival times of individuals in the same cluster can improve the efficiency in the estimation of regression coefficients.

3.6 Summary of the chapter

In this chapter, we considered the penalised likelihood estimation technique proposed by Rondeau and Gonzalez (2005) on the hazard function for both Cox proportional hazard and shared frailty models. This technique was more flexible with the use of penalised splines for the baseline hazard. The proposed technique was applied to a data set with 250260 children and 123110 mothers (clusters). We used an R package called *frailtypack* to estimate model parameters in both Cox PH and gamma shared frailty models. The drawbacks of the EM algorithm and the partial likelihood methods commonly used for estimation were highlighted and the reason for choosing the full penalised likelihood method.

The estimator of the baseline hazard function was approximated on the basis of splines with 7 knots. It is advisable not to use large number of knots to avoid running problems. The maximum number of knots should be limited to 20 and the minimum should be limited to 4. We used 7 knots as the recommended number to start with until one sees that the graph of the baseline hazard function is stable.

Results showed that gender, birth province, birth year, twin and birth order were significant contributors in the survival of the under five children in South Africa at 5% level. There was heterogeneity between mothers and a strong association between survival times of children from the same mother. The gamma shared frailty estimates were quite similar to the Cox proportional model without frailty term. As expected, the hazard of dying was found to be higher for twins as compared to singletons. The hazard of dying was found to be higher for boys than girls and higher for children born between 2011 and 2015 than those born in 2010. Children born in Eastern Cape, Gauteng, KwaZulu Natal and Western Cape had a lower risk of dying compared to those born in Limpopo.

The results also revealed that the hazard of dying is low in cases where mothers had children before due to experience of handling children that those mothers previously had. We compared the penalised log-likelihoods for the two models and noticed the importance of adding a frailty term. We finally used LCV criterion to compare the fit of the Cox PH and the shared gamma frailty models. It can be concluded that the use of shared frailty model that consider the correlation in the data was necessary because of the positive correlation which existed in the data set.

Chapter 4

Copula model for clustered data

In this chapter, we describe a copula model for clustered survival data by considering Archimedean copulas, a class of copulas with completely monotone generator. Clusters in our data set are large and vary in size and this class of copulas can handle that very well.

We illustrate the methodology discussed in the chapter on the under five child mortality data set described in Chapter 1 Section 1.7.4.2, but excluded children with left truncated survival times and clusters of size 1. We used *timereg* R package to analyse the data as it can handle large clusters of different sizes. The main emphasis in this chapter is on Archimedean copula models, in particular the Clayton copula, and the focus is on survival copulas. The primary interest is on the association between failure times of children from the same mother (siblings).

This chapter begins by outlining general introduction to copula model, which includes basics of joint distributions that are used more often in copula modelling. In Section 4.2, we give a short literature review on copulas. In Section

4.3, we discuss copula in a survival analysis setting paying more attention on Archimedean copulas. A sample splitting technique used to partition our large data set is given in Section 4.4. Results of the analysis and discussions are given in Section 4.5. Finally, in Section 4.6, we concluded the chapter by presenting a summary of the whole chapter.

4.1 General introduction to the copula model

4.1.1 Basics of joint distributions

The joint distribution of random variables $T_1, ..., T_n$ is defined by Trivedi et al. (2007) as:
$F(t_1, ..., t_n) = P(T_1 \leq t_1, ..., T_n \leq t_n)$ and the **survival function** corresponding to $S(t_1, ..., t_n)$ is given by

$$\begin{aligned} S(t_1, ..., t_n) &= P(T_1 > t_1, ..., T_n > t_n) \\ &= 1 - F_1(t_1) \text{ for } n = 1 \\ &= 1 - F_1(t_1) - F_2(t_2) + F_1(t_1)F_2(t_2) \text{ for } n = 2 \\ &= 1 - F_1(t_1) - F_2(t_2) - F_3(t_3) + F_{12}(t_1, t_2) + F_{13}(t_1, t_3) + F_{23}(t_2, t_3) - F(t_1, t_2, t_3) \\ &\text{for } n = 3. \end{aligned}$$

Survival function expressions for any given *n* are also available.

4.1.1.1 The Frechet-Hoeffding bounds for joint distribution function

Consider any joint *cdf* $F(t_1, ..., t_n)$ with univariate marginals cumulative distribution functions given by $F_1, F_2, ..., F_n$. Each marginal distribution can take any value between 0 and 1. The joint *cdf* is bounded below and above by the Frechet's Hoeffding lower (W) and upper (M) bounds defined by Trivedi et al. (2007) as: $W = \max[\sum_{j=1}^{n} F_j - n + 1, 0]$
and

$M = \min[F_1, ..., F_n]$

so that

$W \leq F(t_1, ..., t_n) \leq M$.

The upper bound (M) is always a *cdf*.

4.1.2 The Copula model

The copula model is used to join the marginal survival functions and the joint survival function and gives the type of association (Goethals et al., 2008). It is also multivariate distribution function with margins that are uniform on the interval $[0, 1]$.

In copula modelling, a copula function is introduced to model the association between members belonging to a cluster. This copula function links the marginal survival functions of different individuals in a cluster and then generates the joint survival function. The concept of copula was introduced by Sklar (1959) and since then it has been recognised as a powerful tool for modelling association between random variables (Munyamahoro, 2016). Statisticians are interested in copulas because it is a way of studying measures of dependence and also a point where one can start to construct families of bivariate distributions (Nelsen, 2007). Another reason is that copula approach to model correlated variables is very useful because a copula can give dependence structures regardless of the form of the margins (Trivedi et al., 2007). The dependence structure and the margins can be modelled and estimated separately (Tran et al., 2020). What makes copula functions desirable is that the marginal distributions may come from different families (Trivedi et al., 2007). There are many copulas that are proposed in the literature, but the most commonly applied copula families are Archimedean, Gaussian and *t*-copulas. The most suitable copula for a particular case is the one that best captures dependence feature of the data.

4.1.2.1 Concepts of copula modelling

In this section we review some of the concepts of copulas, general definitions, known properties and theorems that will be used in the sections that follow. We will focus on main topics, but a full explanation about copulas can be found in Georges et al. (2001), Djehiche and Hult (2004), Tibaldi (2004) and Nelsen (2007). Firstly, we introduce the definition of the multi-dimensional copula function.

Let $T_1, T_2, ..., T_n$ be random variables with marginal cumulative distribution functions $F_i(t_i) = P(T_i \leq t_i)$ for $i = 1, 2, ..., n$. If we apply the probability intergral transform to each component, $U_i = F_i(t_i)$ has a uniform distributed marginal. The copula of $(T_1, T_2, ..., T_n)$ is defined as the joint cumulative distribution function of $(U_1, U_2, ..., U_n)$ namely:

$$C(u_1, u_2, ..., u_n) = P(U_1 \leq u_1, U_2 \leq u_2, ..., U_n \leq u_n). \tag{4.1}$$

The copula C contains all information on the dependence structure between the components of $(T_1, T_2, ..., T_n)$ whereas $F_i(t_i)$ contains all the information on the marginal distributions (Chen, 2014).

Equation 4.1 can be rewritten as the expression of the inverse function $F_i^{-1}(t_i)$ for $i = 1, 2, ..., n$ as

$$C(u_1, u_2, ..., u_n) = P(T_1 \leq F_1^{-1}(u_1), T_1 \leq F_2^{-1}(u_2), ..., T_n \leq F_n^{-1}(u_n)). \tag{4.2}$$

The Frechet's-Hoeffding bounds discussed in Section 4.1.1.1 also apply to copulas because copulas are multivariate distribution functions. If we denote the upper bound as $C_U(u_1, u_2, ..., u_n)$ and the lower bound as $C_L(u_1, u_2, ..., u_n)$, then the Frechet bounds for copulas are as follows:

$$C_L(u_1, u_2, ..., u_n) \leq C(u_1, u_2, ..., u_n) \leq C_U(u_1, u_2, ..., u_n).$$

We now state the Sklar's theorem which links the marginals with the dependence structure.

4.1.2.2 Sklar's theorem

Copula models are popular because of Sklar's theorem which says that any joint distribution of random variables can be described by the marginal distributions and a copula.

Let $T_1, T_2, ..., T_n$ be random variables with $F_i(t_i) = P(T_i \leq t_i)$, the distribution function of T_i. Let $F(t_1, t_2, ..., t_n) = P(T_1 \leq t_1, T_2 \leq t_2, ..., T_n \leq t_n)$ be a joint density function. Then there exists a copula C such that for all $t_1, t_2, ..., t_n$ in $\mathcal{R}^2$

$$F(t_1, t_2, ..., t_n) = C(F_1(t_1), F_2(t_2), ..., F_n(t_n)). \tag{4.3}$$

Conversely, if C is a copula and $F_1, ..., F_n$ are distribution functions, then $F(t_1, t_2, ..., t_n)$ defined by (4.3) is a joint distribution function with marginal distribution functions $F_1, ..., F_n$ (Djehiche and Hult, 2004).

4.1.2.3 Bivariate copulas

A bivariate copula C is a function from $[0, 1]$x$[0, 1]$ into $[0, 1]$ with the following properties:

- For every u_1, u_2 in $[0, 1]$, $C(u_1, u_2)$ is grounded, i.e., $C(u_1, u_2) = 0$ if at least one of the coordinates is zero.
- $C(u_1, u_2)$ is two-increasing, i.e., for every a_1 and a_2 in $[0, 1]$ such that $a_1 < a_2$, the C-volume $V_C([a_1, a_2])$ of the box $[a_1, a_2]$ is positive.
- $C(u_1, 1) = u_1$ and $C(u_2, 1) = u_2$ for every $(u_1, u_2) \in [0, 1]$x$[0, 1]$.

It follows that a copula is a bivariate distribution function with uniform margins. When the margins are independent, the following product copula is ob-

tained: $C_p(u_1, u_2) = u_1 u_2$.

4.1.2.4 Multivariate copulas

In this section, the results of the bivariate case are extended to the multivariate case. A multivariate copula is a continuous multivariate distribution function with uniform margins on the unit interval (Tibaldi, 2004).
An *n*-copula is a function C from $[0,1]^n$ into $[1,0]$ satisfying the following conditions in order to be a distribution function with standard uniform marginal distributions (Trivedi et al., 2007):

- $C(1, ..., 1, u_i, 1, ..., 1) = u_i$ for every $i \leq n$ and all u_i in $[0,1]$.
  This property says that if the joint probability of the i outcomes is the same as the probability of the remaining uncertain outcome, then the realisations of $i-1$ variables are known with marginal probability.

- $C(u_1, ..., u_n) = 0$ if $u_i = 0$ for any $i \leq n$.
  This property says that the joint probability of all outcomes is zero if the marginal probability of any outcome is zero. This property is also referred to as the grounded property of a copula.

- $C(u_1, ..., u_n)$ is *n*-increasing.
  This property says that the C-volume of any n-dimensional interval is positive.

An *n*-copula is an *n*-dimensional distribution function with all n univariate margins being $U(0,1)$. A natural extension of the bivariate copula to a multivariate case is by considering $F_1(u_1), ..., F_n(u_n)$. Then the function $C(F_1(u_1), ..., F_n(u_n)) = F(u_1, ..., u_n)$, denotes a multivariate distribution function evaluated at $u_1, ..., u_n$ with marginal distributions $F_1, ..., F_n$.

4.1.2.5 Measures of dependence

In this section, different types of dependence measures are discussed, namely: the Pearson linear correlation, the rank correlation and the tail dependence.

Pearson's linear correlation

This is the most commonly used type of dependence measure which measures the direction and the degree to which variables are linearly related to one another. It is developed with an intention of measuring correlation and addresses only linear dependence (Munyamahoro, 2016).

Let T_1 and T_2 be two vectors of random variables with finite and non-zero variances, then the Pearson's linear correlation coefficient is given by:

$$\rho(T_1, T_2) = \frac{Cov(T_1, T_2)}{Var(T_1)Var(T_2)},$$

where $Cov(T_1, T_2)$ is the covariance between T_1 and T_2 and $Var(T_1)$ and $Var(T_2)$ are the two variances of two random variables T_1 and T_2, respectively. The Pearson correlation coefficient always lies between -1 and 1. When $\rho(T_1, T_2) = 1$, then T_1 and T_2 are said to be perfectly dependent by an increasing relationship. When $\rho(T_1, T_2) = -1$, then T_1 and T_2 are said to be perfectly dependent by a decreasing relationship and when $\rho(T_1, T_2) = 0$, then T_1 and T_2 are independent (Mahfoud and Michael, 2012). Inference about $\rho(T_1, T_2)$ for small samples is dependent on the assumption of normality of the data. When these assumptions are not met, non-parametric methods may be applied.

Rank correlation

Rank correlation is the correlation between two variables whose values are ranks. In this section, two rank correlation measures are discussed, namely: Spearman's rank and Kendall's rank correlations. Both measures are based on

concordance concept which says that large values of one variable are associated with large values of another variable. The disconcordance says that large values of one variable are associated with small values of another variable.

a. Spearman's rank

This is a non-parametric correlation defined by Cherubini et al. (2004) as:

$$\rho_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)},$$

where n is the number of paired ranks and d_i is the difference between two ranks of the i^{th} observation. The ranking of variables is done by assigning the highest rank to the highest value. The advantages of Spearman's rank are that, it is easier to calculate by hand and that it can be used for any data that can be ranked including quantitative data. Another advantage of the Spearman's rank is its ability to capture the non-linear dependence between the two variables (Mahfoud and Michael, 2012).

b. Kendall's rank correlation

Kendall's tau τ measures the strength of the relationship between two failure times and it ranges between -1 and 1 (Hsieh, 2010).

Suppose that T_1 and T_2 are correlated random variables. Suppose that (T_{1i}, T_{2i}) and (T_{1j}, T_{2j}), where $i \neq j$ are independent realisations from T_1 and T_2. The pairs (T_{1i}, T_{2i}) and (T_{1j}, T_{2j}) are concordant if $(T_{1i} - T_{1j})(T_{2i} - T_{2j}) > 0$ and discordant if $(T_{1i} - T_{1j})(T_{2i} - T_{2j}) < 0$. The population Kendall's tau is now given by:

$\tau = P[(T_{1i} - T_{1j})(T_{2i} - T_{2j}) > 0] - P[(T_{1i} - T_{1j})(T_{2i} - T_{2j}) < 0]$.

In a situation where censoring is not present, the sample value of Kendall's tau can be estimated as

$$\hat{\tau} = \frac{1}{\binom{n}{2}} \sum_{1 \leq i \leq j \leq n} g_{ij}$$

where

$$g_{ij} = \begin{cases} -1, & \text{if } (T_{1i} - T_{1j})(T_{2i} - T_{2j}) < 0 \\ 1, & \text{if } (T_{1i} - T_{1j})(T_{2i} - T_{2j}) > 0 \end{cases}$$

In cases where censoring is present in T_2, we observe data of the form (T_{1i}, Z_i, δ_i), where $Z_i = \min(T_{1i}, C_i)$ and $\delta_i = 1(T_{1i} \leq C_i)$. C_i is the censoring variable.

The Kendall's tau in the presence of censoring can be estimated based on Brown Jr et al. (1973) as follows:

$$\hat{\tau} = \frac{\sum_{i=1}^{n} \sum_{i=1}^{n} m_{ij} p_{ij}}{\sqrt{(\sum_{i=1}^{n} \sum_{i=1}^{n} m_{ij}^2)(\sum_{i=1}^{n} \sum_{i=1}^{n} p_{ij}^2)}}, \tag{4.4}$$

where $m_{ij} = 1(T_{1i} > T_{1j}) - 1(T_{1i} < T_{1j})$
and $p_{ij} = P(T_{2i} > T_{2j} | Z_i, Z_j, \delta_i, \delta_j; \bar{G}_n) - P(T_{2i} < T_{2j} | Z_i, Z_j, \delta_i, \delta_j; \bar{G}_n)$.

$\bar{G}_n(t_2)$ represents a Kaplan-Meier estimate of $\bar{G}(t_2) = P(T_2 > t_2)$.

In 4.4, p_{ij} reduces to $1(Z_i > Z_j) - 1(Z_i < Z_j)$ whenever $\min(Z_i, Z_j)$ is observed, that is when $\delta_i = \delta_j = 1$.

Tail dependence

Another useful concept in copula modelling is the upper and lower tail dependence. They measure the dependence between the two variables in the upper and lower tail values of the two variables. Tail dependence looks at the concordance between extreme values of random variables. It is the most appropriate when interested in the probability that one variable exceeds or fall below some

given threshold (Mahfoud and Michael, 2012). According to Nelsen (2007), the upper tail dependence parameter λ_U is the limit (if it exists) of the conditional probability that the variable takes a very high value given that the other also takes a very high value. Similarly, λ_L is the conditional probability in the limit that one variable takes a very low value given that the other also takes a very low value.

To define the upper and the lower tail dependence parameters:
Let T_1 and T_2 be two continuous random variables with copula C and marginal functions $F_1(.)$ and $F_2(.)$ for a quartile threshold t, the upper and the lower tail dependence are defined as

$$\lambda_U = \lim_{t \to 1^-} P(F_2(t_2) \geq t | (F_2(t_1) \geq t = \lim_{t \to 1^-} \frac{1 - 2t + C(t,t)}{1-t}$$

$$\lambda_L = \lim_{t \to 0^+} P(F_2(t_2) \leq t | P(F_1(t_1) \leq t = \lim_{t \to 0^+} \frac{C(t,t)}{t},$$

provided that $\lambda_U \in [0,1]$ and $\lambda_L \in [0,1]$ exist.

If $\lambda_U \in [0,1]$, T_1 and T_2 are asymptotically dependent in the upper tail; if $\lambda_U = 0$, T_1 and T_2 are asymptotically independent in the upper tail. Similarly, if $\lambda_L \in [0,1]$, T_1 and T_2 are asymptotically dependent in the lower tail; if $\lambda_U = 0$, T_1 and T_2 are asymptotically independent in the lower tail. The association is positive if the copula attains the upper Frechet Hoeffding bound and negative if it attains the lower Frechet Hoeffding bound (Trivedi et al., 2007).

4.2 Review of related literature

Copula models were first applied by Clayton (1978) who studied the life tables of fathers and sons in a bivariate survival data. He was then followed by other researchers such as Hougaard (1986) and Genest and MacKay (1986). Copula models are popular in modelling dependence between random variables, espe-

cially in the fields of actuarial science, biostatistics and finance. It is often a good idea to study the joint mortality pattern of groups of individuals instead of a single individual because of the strong confirmation that supports the association of mortality on pairs of individuals (Emamverdi et al., 2014). This group could be patients in a hospital, married couple or children from the same mother (siblings) as it applies in our study. It has been pointed out by Frees and Valdez (1998) that pairs of individuals show association in mortality because they share common risk factors which might be genetic in the case of siblings or environmental in the case of married couples.

Studies involving joint mortality pattern of individuals were studied by different researchers including Parkes et al. (1969), Ward (1976), Frees et al. (1996), Emamverdi et al. (2014) and King et al. (2017). A study done by Parkes et al. (1969) to investigate mortality pattern among widowers showed a higher rate of mortality among widowers during the first six months of bereavement after the death of their spouses. The highest percentages of causes of death were related to coronary thrombosis and heart diseases. In a study conducted by Ward (1976), a death pattern of widowers and widows was followed for two years after the death of their spouses. The results showed more widowers' deaths during the first six months of bereavement and that confirmed the study by (Parkes et al., 1969).

The work done by Frees et al. (1996), which investigated mortality of individuals surviving spouses entitled to collect annuity investments using a copula, showed a reduction in annuity values when accounting for dependency in mortality than the standard models that assume independence. A strong positive association between joint lives was shown. A study done by Emamverdi et al. (2014) to investigate joint life policy's premium using copula revealed a lower joint life insurance premium than when the sum of two policies of the spouse

were bought separately. King et al. (2017) studied men and women who died and had been living with other people as husband and wife, but not legally married to them. The results of their study showed a higher risk of death in the first three months of bereavement.

(Dufresne et al., 2018) conducted a study to model the association between survival times within married couples. In their paper, age difference and gender of the elder partner were introduced as an argument of the association parameter of the copula. The results revealed that correlation decreases with age difference and that the association between survival times were higher when the husband is older than the wife. The scholars further suggested that survival times dependence factors should be taken into consideration when evaluating annuity products involving couples. Studies discussed above showed the importance of considering association between observations of related individuals or items.

4.3 Copula in survival analysis

The notations of copulas as introduced before will be extended to survival copulas.

4.3.1 Sklar theorem in survival functions

The Sklar's theorem is also applicable to survival functions.

Let $X = (X_1, X_2, ..., X_p)$ be a random variable with joint survival $\bar{F}$ and univariate survival marginals $\bar{F}_i$, $i = 1, 2, ..., p$, then
$\bar{F}(x_1, x_2, ..., x_p) = \tilde{C}(\bar{F}_1(x_1), \bar{F}_2(x_2), ..., \bar{F}_p(x_p))$.
$\tilde{C}$ is the survival copula of X.

In particular, let C be the copula of X and $U = (U_1, U_2, ..., U_p)$ be a vector such that $U \backsim C$. This means that U follows a copula function C (Durante and Sempi, 2010).

Then,

$\tilde{C}(u) = \bar{C}(1 - u_1, 1 - u_2, ..., 1 - u_p)$, where

$\bar{C}(u) = P(U_1 > u_1, U_2 > u_2, ..., U_p > u_p)$ is the survival function associated with C.

4.3.2 Archimedean copulas

Archimedean copulas are a popular class of copulas because of the nice properties they possess which makes it easy to construct them. They are also popular because they can capture wide ranges of dependence (Trivedi et al., 2007). There are three families of Archimedean copulas that are commonly used: Clayton-Oakes, Gumbel and Frank, but the focus in this thesis is on Clayton-Oakes.

Let C_θ be an Archimedean copula with a generator function $p_\theta(.)$. Then

- C_θ is symmetric, i.e., $C_\theta(u_1, u_2) = C_\theta(u_2, u_1)$ for all u_1, u_2 in I.
- C_θ is associative, i.e., $C_\theta(C_\theta(u_1, u_2), u_3) = C_\theta(u_1, C_\theta(u_2, u_3))$ for all u_1, u_2, u_3 in I.

θ denotes the dependence parameter of the copula measuring the association between the marginals. A copula C_θ is called Archimedean if it takes the form:

$$C_\theta(u_1, u_2, ...u_n) = p_\theta(p_\theta^{-1}(u_1) + ... + p_\theta^{-1}(u_n)). \tag{4.5}$$

Here $p_\theta(.)$ is a strictly decreasing generator function of the Archimedean copula with $p_\theta(0) = 1$ and $p_\theta(\infty) = 0$. The inverse function of $p_\theta(.)$ is denoted by p_θ^{-1}.

Three common families of Archimedean copulas with their generators and Kendall's are summarised in Table 4.1.

Table 4.1: Archimedean copulas with their generators and Kendall's tau

Copula family	Range of θ	Generator $p_\theta(t)$	$C_\theta(u_1, u_2)$	Kendall's tau
Frank	$\theta\ \epsilon[-\infty, \infty]$	$-\log\frac{e^{-\theta t}-1}{e^{-\theta}-1}$	$\frac{-1}{\theta}\log\left[1+\frac{(e^{-\theta u_1}-1)(e^{-\theta u_2}-1)}{e^{-\theta}-1}\right]$	$1-\frac{4}{\theta}[1-D^\star(\theta)]$
Gumbel-Hougaard	$\theta\ \epsilon[1, \infty]$	$(-\log t)^\theta$	$\exp\left[-\{(-\log u_1)^\theta+(-\log u_2)^\theta\}^{\frac{1}{\theta}}\right]$	$\frac{\theta-1}{\theta}$
Clayton	$\theta\ \epsilon[0, \infty]$	$\frac{t^{-\theta}-1}{\theta}$	$\max\{(u_1^{-\theta}+u_2^{-\theta}-1)^{\frac{-1}{\theta}}, 0\}$	$\frac{\theta}{\theta+2}$

$D^\star(x) = x^{-1}\int_0^x \frac{t}{e^t-1}dt$

Generators of these families have only one parameter θ which is used to measure the degree of dependence between two random variables u_1 and u_2.
The three families of Archimedean copulas are described briefly in subsequent subsections.

4.3.2.1 Frank copula

The Frank copula takes the form:
$C_\theta(u_1, u_2) = \frac{-1}{\theta}\log\left[1+\frac{(e^{-\theta u_1}-1)(e^{-\theta u_2}-1)}{e^{-\theta}-1}\right]$.
The dependence parameter θ may assume any value between $-\infty$ and ∞. A value of $-\infty$ corresponds to the Frechet lower bound, a value of zero shows independence and ∞ corresponds to the Frechet upper bound. The Frank copula is popular because negative dependence between the marginals is allowed and that the dependence is symmetric in both tails. We can use this model to model outcomes with both negative and positive association and is the right model for data that show weak tail dependence.

4.3.2.2 Gumbel-Hougaard copula

The Gumbel copula takes the form:

$$C_\theta(u_1, u_2) = \exp\left[-\{(-\log u_1)^\theta+(-\log u_2)^\theta\}^{\frac{1}{\theta}}\right]. \tag{4.6}$$

The dependent parameter θ assumes any value between 1 and ∞. A value of 1 corresponds to independence and a value of ∞ corresponds to the Frechet lower bound. Gumbel copula is the most suitable in cases where outcomes are strongly associated at high values but less associated at low values.

4.3.2.3 Clayton-Oakes copula

Clayton copula is an asymmetric Archimedean copula that exhibits greater dependence in the negative tail than in the positive tail. When correlation between two events is strongest in the left tail of the joint distribution, Clayton copula is the right model to be used (Trivedi et al., 2007). Similarities between Clayton-Oakes and shared gamma frailty model have been pointed out in literatures such as Goethals et al. (2008).

The Clayton Oakes copula is given by:
$C_\theta(u_1, u_2) = \max\{(u_1^{-\theta} + u_2^{-\theta} - 1)^{\frac{-1}{\theta}}, 0\}$,
where θ is the copula parameter restricted on the interval $(0, \infty)$. The marginals become independent as the value of θ approaches zero. As the value approaches infinity, the marginals are dependent, and the copula takes the Frechet upper bound. No value takes the Frechet lower bound (Trivedi et al., 2007).
Its generator is given by:
$p_\theta(t) = \frac{1}{\theta}(t^{-\theta} - 1)$.

Coefficients of the upper and the lower tail dependence

The coefficients of the upper tail dependence are $\lambda_u = 0$ and the lower tail dependence is
$\lambda_l = \lim_{u \to 0} \frac{C(u,u)}{u} = \frac{(2u^{-\theta}-1)^{\frac{-1}{\theta}}}{u} = 2^{\frac{-1}{\theta}}$ for $\theta > 0$.

Relationship between Kendall's tau τ and θ

The relationship between Kendall's τ and θ is given by:

$\tau = \frac{\theta}{\theta+2}$ and $\theta = \frac{2\tau}{1-\tau}$.

The values of τ range between -1 and 1. A value of 1 corresponds to a perfect correlation and -1 a perfect inverse correlation. If $\tau = 0$, then the survival times are independent.

4.3.2.4 Description of the Archimedean model

Suppose that we have a total of k clusters $(i = 1, ..., k)$. Let T_{ij} denote the survival times for different individuals ($j = 1, ..., n$) in each cluster. Here n is the total number of members in cluster i. For each member in a cluster, we assume that there is an independent random censoring variable C_{ij}.

The observed quantities under the right censoring scheme are $Y_{ij} = min(T_{ij}, C_{ij})$ and the censoring indicators $\delta_{ij} = I_{T_{ij} \leq C_{ij}}$. The risk of death may also depend on a set of covariates x_{ij}. The joint survival function for the survival time of different members within cluster i is given by:

$$\begin{aligned} S(t_{i1}, ..., t_{in}|X_{i1}, ..., X_{in}) &= P(T_{i1} > t_{i1}, ..., T_{in} > t_{in}|X_{i1}, ..., X_{in}) \\ &= p_\theta[p_\theta^{-1}(S(t_{i1}|X_{i1})) + ... + p_\theta^{-1}(S(t_{in}|X_{in}))]. \end{aligned} \tag{4.7}$$

$S(t_{ij}|X_{ij})$ is the survival function for the j^{th} univariate marginal given covariate X_{ij}. The generator, p_θ, is a continuous strictly decreasing function with $p_\theta(0) = 1$ and $p_\theta(\infty) = 0$ and its inverse is p_θ^{-1}. The Archimedean copula has to be defined in such a way that it accommodates all clusters of different sizes. Therefore, this generator is assumed to be completely monotonic, i.e., all the derivatives exist and have alternating signs:$(-1)^r \frac{d^r}{dt^r} p_\theta(s) \geqslant 0$, for all $s > 0$ and $r = 0, 1, 2, ...$ (Prenen et al., 2017).

4.3.3 Estimation in copula models

In this section, the estimation method usually used in copula models is a two-stage method (Martinussen and Scheike, 2007). Nonparametric approaches do not achieve maximum productivity and when there is censoring they can be inconsistent.

Copula models combine the marginal approach with a model for the association within individuals. The joint survival function is modelled through the marginal survival functions and an association parameter. Due to the way copula models divide the estimation process into two stages, a two-stage estimation procedure is naturally suggested by first estimating the marginal survival functions in the first stage and then replacing the marginal function in the likelihood function by their estimates obtained in the first stage so that the association parameter can be estimated in the second stage. When marginal survival functions are estimated in the first stage, clustering is not taken into consideration and therefore, the event times of members are taken as independent of each other even though they belong to the same cluster.

A two-stage estimation procedure was first suggested by Hougaard (1986) and then studied by Genest et al. (1995), Shih and Louis (1995), Glidden (2000), Andersen (2005) and Othus and Li (2010).
In a study done by Hougaard (1986), a Nelson-Aalen type estimator was used for the margins. The two-stage procedure was investigated using a data on 50 litters of female rats each containing one drug treated and two control animals. Genest et al. (1995) used this procedure in a bivariate setting without including covariates. In their paper, comparisons with other procedures in Clayton's family were made using simulated data. In a study conducted by Shih and Louis (1995), each margin was modelled separately. Both two-stage parametric and semiparametric estimation procedures were investigated for the association

parameter in a bivariate setting without covariates included. The two estimation procedures were applied to AIDS data set. The work done by Genest et al. (1995) and Shih and Louis (1995) were limited to bivariate settings excluding covariates.

Glidden (2000) studied two-stage estimation using Clayton-Oakes copula, and the model for association parameter was based on gamma model. His work was an extension to the work done by Shih and Louis (1995) to allow the marginal hazard for survival of individuals to follow a stratified Cox model. The approach by Glidden (2000) considered multiple survival times per cluster with covariates modelled by a marginal Cox model. The two-stage estimation was applied to a data set containing sets of twins to analyse the association between monozygotic and dizygotic twins in various diseases.

A study done by Andersen (2005) was also an extension to the work done by (Shih and Louis, 1995). In this study, groups of siblings were followed until disease occurrence, death or follow-up period. The disease times were then said to be correlated within siblings if there was a familial clustering of disease. A contribution of this study was in the adjustment for confounders, while estimating the association parameter and the variation of the estimates. Although the method was found to be very efficient in both parametric and semi-parametric models, the author also concentrated only on paired survival times, and not consider families of different sizes.

Massonnet et al. (2009) extended the above-mentioned studies to include clusters of size four to model the time until infection in the four different quarters of a cow udder. A two-stage estimation approach was used, and a new bootstrap algorithm was proposed so that the standard errors of the parameter estimates can be obtained. Othus and Li (2010) applied a two-stage estimation

approach using a Gaussian copula to a Children's Oncology data set. The aims of the study were to test if there was an association between the chemotherapy schedule and improved survival and also to test whether correlation exists, while controlling for known factors or characteristics that will define the natural history of the disease. In most of the studies cited above, researchers used clusters of small and equal sizes because it is difficult to obtain an expression for the likelihood function when we have large cluster sizes. As an example, if the cluster size is 2, then there are only 4 contributions $(2^2 = 4)$ to the likelihood for the observed outcomes within the cluster and that depends on whether none, the first, second or both of individuals are censored. In general, when we have clusters of size n, we will have 2^n possible combinations, i.e., a likelihood function will contain different 2^n possible terms and to find each term we need to take the derivatives of the joint survival function over the uncensored cases, and that becomes difficult.

Prenen et al. (2017) considered large and varying clusters to model time to insemination in cows clustered in herds using both one-stage and two-stage estimation procedures. They further showed consistency and asymptotic normality of estimators produced by both one-stage and two-stage estimation procedures. Although Glidden (2000), Othus and Li (2010) and Prenen et al. (2017) gave results for unbalanced and balanced designs for different copula models,in general, copula models have not been used when cluster sizes are more than four and differ over clusters to model mortality of under five children in South Africa using *timereg* package.

In this thesis, a two-stage semi-parametric estimation procedure is used, and the marginal survival functions are estimated using a marginal Cox model under the independence working assumption.

4.3.3.1 Two stage semi-parametric estimation procedure

Two stage parametric estimation can be illustrated as follows:

First stage: estimation of the marginal survival functions

In this stage, we estimate the marginal survival functions using the marginal Cox model with covariate vector x_{ij} given by:

$$h_{ij}(t) = h_0(t)\mathbf{exp}(\beta_0^{'} X_{ij}), \tag{4.8}$$

where $h_0(t)$ is the baseline hazard at time t and β_0 is the fixed effect parameter. To estimate β_0 in 4.8, we ignore clustering and we act as if the event times of individuals are independent of each other. This is called the independence working assumption. We then obtain the estimator $\hat{\beta}$ for the true parameter β_0 by solving the score equation $U_\beta(\beta)$ given by:
$U_\beta(\beta) = \sum_{i=1}^{k}\sum_{j=1}^{n} \delta_{ij}\frac{\partial log(f(y_{ij}|X_{ij}))}{\partial\beta} + (1-\delta_{ij})\frac{\partial log(f(S_{ij}|X_{ij}))}{\partial\beta} = 0.$
The marginal survival functions are then estimated as follows:

$$\hat{S_{ij}}(t) = \mathbf{exp}[-\hat{H}_0(t,\hat{\beta})\mathbf{exp}(\hat{\beta}^{'} X_{ij}]),$$

where $\hat{H}_0$ is the estimator for the cumulative baseline hazard H_0.

Second stage: estimation of the marginal survival functions

The association parameter θ is estimated by plugging in the estimates for the margins into the likelihood expression which is then maximised to solve for $\hat{\theta}$. The two-stage estimator for θ is the solution to
$U_\theta(\hat{\beta}) = \frac{\partial log(L(\hat{\beta},\hat{\theta}))}{\partial\theta} = 0.$

4.4 Sample splitting technique to partition large data sets

Data analysts always come across data sets that are difficult to analyse at once. Molenberghs et al. (2011) proposed pseudo-likelihood methodology which can split such data sets into sub-samples, each of which can be analysed separately and then combine all estimates to become one. Their method was found to be efficient in a situation where the sub-samples are independent. We adopted this sample splitting technique and applied it in our survival analysis setting because it was difficult to include all observations at once to fit a Clayton copula. The data set was properly split-up in such a way that every child was a member of one and only one sub-sample. We did the splitting of the data base based on clusters (mothers) mainly in order to make sure that those children with the same mother will be in the same data set for further analysis. The sub-samples varied in sizes. The main goal was to find the overall association parameter and the overall standard error. The overall association parameter and the overall standard error were found using the sample splitting technique as follows:

The data set was split into G sub-samples each of size n_i.

Let
N denote the total number of individuals in the study,
θ_i be the association parameter obtained in sub-sample i,
S_{e_i} be the standard error obtained in sub-sample i.
The overall association parameter $\hat{\theta}$ was obtained as follows:

$$\hat{\theta} = \sum_{i=1}^{G} \frac{n_i}{N} \theta_i \quad i = 1, 2, ..., G, \tag{4.9}$$

and the overall standard error as:

$$\hat{S}_e = \sqrt{\sum_{i=1}^{G} \left(\frac{n_i}{N}\right)^2 (S_{e_i})^2} \tag{4.10}$$

4.5 Data analysis and results

In this section, the Clayton model and the two-stage estimation procedure discussed previously are illustrated using the data set introduced in Chapter 1 Section 1.7.4.2. We excluded children with left truncated survival times and all children belonging to clusters of size 1. The total number of clusters in this data set was 120033 with 219704 children. Clusters varied in size between 2 and 5 children. Out of 219704 children, 4978 (2.3%) were indicated as died and 214726 (97.7%) as alive. The number of clusters in this data set was too big to be analysed at once. Due to this reason, the data set was divided into three sub-samples of equal number of clusters selected randomly. There were 40011 clusters in each sub-sample. The total number of children in each sub-sample were different. There were 75481 children in sub-sample 1, 74271 children in sub-sample 2 and 69952 children in sub-sample 3 and that makes it to be an unbalanced design. The three sub-samples were analysed separately. The marginal parameters estimated in the first stage of a two-stage estimation procedure for the three sub-samples are given in Table 4.2.

Table 4.2: Marginal parameter estimates from the Clayton-Oakes copula model

Factors	Sub-sample 1				Sub-sample 2				Sub-sample 3			
	Coef	S_e	Robust S_e	p-value	Coef	S_e	Robust S_e	p-value	Coef	S_e	Robust S_e	p-value
Gender	0.0665	0.0469	0.0470	< 0.0000	0.0928	0.0487	0.0487	< 0.0000	0.0863	0.0523	0.0515	< 0.0000
Province	-0.0183	0.0107	0.0111	< 0.0000	-0.0361	0.0112	0.0118	< 0.0000	-0.0165	0.0124	0.0129	< 0.0000
Year	-0.0702	0.0387	0.0402	< 0.0000	-0.0984	0.0422	0.0431	< 0.0000	-0.0254	0.0468	0.0479	< 0.0000
Twin	0.0617	0.0542	0.0559	< 0.0000	0.1340	0.0626	0.0650	< 0.0000	0.2170	0.0783	0.0831	< 0.0000
Order	-0.4590	0.0860	0.0882	< 0.0000	-0.4590	0.0874	0.0848	< 0.0000	-0.5060	0.0878	0.0909	< 0.0000

The marginal parameters in Table 4.2 are estimated under the working independence assumption in which the cluster structure is not considered when

estimating the effects of the covariates and the cluster structure is only used when deriving valid estimates of standard errors (robust standard errors) to ensure correct inference. It can be seen from the output that robust standard errors in all three sub-samples are higher than their naive counterparts. These robust standard errors are derived by considering that survival times of siblings cannot be taken as independent.

From Table 4.2 one can see that in all three sub-samples, there is an increased risk of death in male children as compared to female children. The results also show the decreased risk of death in children born in 2013 as compared to those born in 2014 and 2015. We also observe that being part of a twin reduces the chance of survival. The results show that a child who is born second, third, fourth or fifth has a significantly higher chance of survival than a first born child.

The association parameters for the three sub-samples were estimated in the second stage by using Clayton-Oakes model. The association parameters (θ) together with their standard errors, p-values and number of children in each sub-sample are given in Table 4.3.

Table 4.3: Association parameters with standard errors for three sub-samples

Sub-sample	No of children	Association parameter θ	S_e	p-value
1	75481	0.0515	0.0271	0.0576
2	74271	0.0664	0.0304	0.0289
3	69952	0.0364	0.0315	0.248
Overall	2190704	0.0517	0.0171	0.0025

The overall values of the association parameter and the standard error were found by using the sample splitting technique described in Section 4.4. The overall estimate of θ was found to be 0.0517 with an estimated standard error of 0.0171 and a p-value of 0.00250. We can see that the p-value was highly

significant. Based on these results, we can conclude that there is a positive correlation between survival times of siblings at 5% level.

4.5.1 Discussion of results

The results in all three sub-samples showed that female children were more likely to survive than male children. The results showed the decreased risk of death in situations where mothers had previous living children. This might be due to the fact that mothers are getting better ways to improve survival chances with each additional child born. It has also been reported in a paper by Masset and White (2003) that the mortality is higher among first birth due to the fact that some mothers have their first children before being ready physically and with no reproductive maturity. Sullivan et al. (1994) and Pebley and Stupp (1987) noticed that high birth order increased mortality in ages between 1 and 4 outside Sub Saharan Africa. Pebley and Stupp (1987) suggested that medical factors played a role in high birth order deaths and the chance of spreading infectious diseases because most high birth order come from larger families.

We also observe that being part of a twin reduces chance of survival. These findings are also in line with results from previous studies like Alam et al. (2007) and (Pebley and Stupp, 1987). A study by Alam et al. (2007) showed that infant mortality among multiple births was more than five times higher among singletons.

The results of this study also suggest that province of birth has a significant effect on under five child mortality and that children born in Limpopo province had better survival probabilities than children in other provinces. With regard to birth year, the results of this study showed that children born in 2010 had better survival probabilities than children born in other years. Based on the

output, we can conclude that all factors included in the copula model significantly contributed to the under five child mortality in South Africa and that there is a positive correlation between survival times of siblings.

In general, the findings based on a Clayton-Oakes copula are similar to the findings from a shared frailty model in the previous chapter.

4.6 Summary of the chapter and concluding remarks

The aim of this chapter was to model time to death of children clustered in mothers using Clayton-Oakes copula. We used *two.stage* function available in *timereg* R package to fit a Clayton-Oakes model.

We investigated the two-stage semiparametric estimation procedure in which the marginal survival functions were estimated using the Cox PH model and the association structure was modelled by a Clayton copula. We focused on clusters with at least 2 individuals. Due to the big size of the data set, we used the splitting technique to break the data into three sub-samples of equal number of clusters and analyse each sub-sample separately. In the first stage the marginal parameters were obtained under the working independence assumption and in the second stage the estimate of the association parameter was obtained. The results from the Clayton-Oakes model showed that gender, birth province, birth year, twin and birth order had an effect on survival of under five children. The results further revealed a poorer survival associated with male children and also with being part of a twin. It was found that the more children born previously to a mother, the lower the chance of dying due to the experience of handling children of the mother. Children born in Limpopo showed a lower risk of death compared to other children. Those who were born

in 2010 showed lower risk of death than those who were born in other years. The conclusion based on the results showed association between survival times of children from the same mother. We can conclude that the use of the Clayton Oakes model was necessary in this particular setting.

Chapter 5

Comparison between shared frailty and copula models

In this chapter, shared frailty and copula models are compared to assess how the two models handle association within clusters. Another reason is to check if the equivalence between the two models really exists as claimed in some literature such as Andersen (2005). To make a good comparison between the two models, we applied the same data set to both models. This data set contains all clusters with at least two individuals, but excluded left truncated observations. Due to the number of clusters in this data set, the data set was divided into three sub-samples of equal number of clusters and each sub-sample was analysed separately.

The shared frailty model analysed in Chapter 3 included all left truncated observations. In this chapter, the shared frailty model was revisited without including left truncated observations to establish how it handles association in this new condition. Another reason for eliminating those observations was because it was difficult to include them in a copula model since most clusters

contained one single individual, hence splitting clusters into sub-samples led to some sub-samples having only independent single individual cluster. It was not going to be possible for us to compare two models using two different data sets.

5.1 Similarities and differences between copula and frailty models

5.1.1 The copula and the frailty models compared

Let T be the survival time of an individual and Z be the frailty variable in a univariate frailty model without considering clustering. The conditional hazard function for a given frailty variable $Z = z$ at time $t > 0$ is given by
$h(t|z) = zh_0(t)\mathbf{exp}(X\beta)$,
where $h_0(t)$ is the baseline hazard function and β is the column vector of regression coefficients. The conditional survival function for a given frailty at time $t > 0$ is given by:

$$\begin{aligned} S(t|z) &= \mathbf{exp}\left[\int_0^t h(x|z)dx\right] \\ &= \mathbf{exp}[-zH_0(t)\mathbf{exp}(X\beta)]. \end{aligned} \tag{5.1}$$

The marginal survival function can be obtained by integrating over the range of frailty variable z as follows:

$$\begin{aligned} S(t) &= \int_0^\infty S(t|z)f(z)dz \\ &= \int_0^\infty \mathbf{exp}[-zH_0(t)\mathbf{exp}(X\beta)]f(z)dz \\ &= L[H_0(t)\mathbf{exp}(X\beta)], \end{aligned} \tag{5.2}$$

where $L(.)$ is the Laplace transformation of the distribution of Z.
Now consider the clustered survival times $(T_1, ..., T_n)$ for a cluster of size n

and let $S_{1,c}(t_1), ...S_{n,c}(t_n)$ be the marginal survival functions for individuals (children) in a cluster (mother). These marginal survival functions are given by $S_c(t_j) = P(T_j > t_j)$ and are obtained from the marginal approach without taking clustering into consideration. The joint survival function of the *n*-dimensional survival copula model is given by:

$$\begin{aligned} S_c(t_1, t_2, ..., t_n) &= P(T_1 > t_1, ..., T_n > t_n) \\ &= C_\theta(S_{1,c}(t_1), ...S_{n,c}(t_n)), \end{aligned} \tag{5.3}$$

with C_θ a copula function with parameter vector θ. The subindex *c* is added to denote that the joint survival function is obtained from the copula presentation. C_θ is a copula function on the unit square. This is a copula function in the interval $[0, 1]$ x $[0, 1]$. A copula is a function that assigns any point in the unit square $[0, 1]$ x $[0, 1]$ to a number in the interval $[0, 1]$ i.e $C_\theta : [0, 1]^2 \rightarrow [0, 1]$ (Goethals et al., 2008).

The joint survival function for the copula model is given by:

$$\begin{aligned} S_c(t_1, t_2, ..., t_n) &= C_\theta(u_1, u_2, ...u_n) \\ &= \mathcal{L}[\mathcal{L}^{-1}(S_{1,c}(t_1)) + ... + \mathcal{L}^{-1}(S_{n,c}(t_n))]. \end{aligned} \tag{5.4}$$

On the other hand, the hazard for the j^{th} individual from cluster i in a frailty model is given by:

$$h_{ij}(t) = z_i h_o(t) \text{exp}(X_{ij}^{'}\beta), \tag{5.5}$$

where $h_{ij}(t)$ is the hazard at time t in cluster i, z_i is the frailty term, $h_o(t)$ is the baseline hazard and X_{ij} is a set of covariates for individuals j in cluster i.
Let us assume that the frailty term z_i is distributed as gamma. Then the probability density function is given by:
$f_Z(z_i) = \frac{z_i^{(1/\theta - 1)} exp(-z_i/\theta)}{\Gamma(1/\theta)\theta^{1/\theta}}$.

Equation 5.5 can also be written as

$$h_{ij}(t) = z_i h_{j,z}(t), \tag{5.6}$$

where $h_{j,z}(t)$ is the conditional hazard at time t for a cluster with frailty equal to one and for individual *j* and z_i is the frailty term.
To make a good comparison between the copula and frailty models, we consider the family of Archimedean copulas discussed in Chapter 4 which take the form:

$$C_\theta(u_1, u_2, ...u_n) = p_\theta(p_\theta^{-1}(u_1) + ... + p_\theta^{-1}(u_n)). \tag{5.7}$$

To find a link between the copula and the frailty models, we need to consider functions $p_\theta(.)$ that are Laplace transform of frailty densities $f_Z(.)$
$p_\theta(s) = \mathcal{L}(s) = E[\exp(-Zs)] = \int_0^\infty \exp(-zs) f_Z(z) dz$.
Replacing p_θ by $\mathcal{L}$ in equation 5.7, we get

$$\begin{aligned} C_\theta(u_1, u_2, ...u_n) &= p_\theta(p_\theta^{-1}(u_1) + ... + p_\theta^{-1}(u_n)) \\ &= \mathcal{L}[\mathcal{L}^{-1}(u_1) + ... + \mathcal{L}^{-1}(u_n)]. \end{aligned} \tag{5.8}$$

The joint conditional survival function for any cluster in a frailty model is:

$$S(t_1, ..., t_n) = \exp[-z(H_{1,z}(t_1) + ... + H_{n,z}(t_n))], \tag{5.9}$$

where $H_{j,z}(t) = \int_0^t h_{j,z}(s) ds$ is the cumulative baseline hazard for individual j.
Integrating out the frailties with respect to the frailty density, we get the following joint survival function according to Goethals et al. (2008):

$$\begin{aligned} S_f(t_1, ..., t_n) &= \int_0^\infty S(t_1, ..., t_n) f_Z(z) dz \\ &= \int_0^\infty \exp[-z(H_{1,z}(t_1) + ... + H_{n,z}(t_n))] f_Z(z) dz \\ &= E[\exp(-Z(H_{1,z}(t_1) + ... + H_{n,z}(t_n))]. \end{aligned} \tag{5.10}$$

The subindex *f* in the joint survival function of a frailty model is added to de-

note that the joint survival function is obtained from the conditional frailty model.

The joint survival function derived from the frailty model in equation 5.10 and the joint survival function for the copula model in equation 5.4 are two different ways to model $P(T_1 > t_1, ..., T_n > t_n)$.
Expression in equation 5.10 is the Laplace transform of the frailty distribution when $s = H_{1,z}(t_1) + ... + H_{n,z}(t_n)$ so that

$$\begin{aligned} S_f(t_1, ..., t_n) &= \mathcal{L}(s) \\ &= \mathcal{L}(H_{1,z}(t_1) + ... + H_{n,z}(t_n)). \end{aligned} \tag{5.11}$$

According to Goethals et al. (2008), the marginal survival function for each child in the cluster can be obtained by putting the survival times for other children in the same cluster to zero in equation 5.11 and thus
$S_{j,f}(t) = \mathcal{L}(H_{j,z}(t))$ and it follows that

$$H_{j,z}(t) = \mathcal{L}^{-1}(S_{j,f}(t)). \tag{5.12}$$

By applying this relationship, equation 5.11 can be written as

$$S_f(t_1), ..., t_n) = \mathcal{L}[\mathcal{L}^{-1}(S_{1,f}(t_1) + ... + \mathcal{L}^{-1}(S_{n,f}(t_n))]. \tag{5.13}$$

The correlation structure used to obtain the joint survival function from the marginal survival functions in equation 5.4 and equation 5.13 is the same. However, the arguments of the correlation structure and the marginal survival functions are not the same. We can clearly see that the two models indicated in equation 5.4 and equation 5.13 are different in nature. This is the most important distinction between frailty and copula models. To show clearly the difference between the marginal survival functions for the two models, we consider the Clayton-Oakes copula and the gamma shared frailty models in the

next section.

5.1.2 Gamma shared frailty model versus Clayton-Oakes copula model

It has been reported by different authors that some copula models can be deduced from shared frailty models by choosing the appropriate distribution for the frailty term. In this section we show that Clayton-Oakes copula and gamma shared frailty model are only equivalent with respect to the copula function used.

To model the correlation, we use the joint survival function in equation 5.4 with $\mathcal{L}(s) = (1 + \theta s)^{-1/\theta}$. According to Clayton (1978), the corresponding copula $C_\theta(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$ is a Clayton-Oakes copula.
The joint survival function becomes:

$$S_c(t_1, t_2, ..t_n) = [S_1(t_1)^{-\theta} + ... + S_n(t_n)^{-\theta} - 1]^{-1/\theta}. \tag{5.14}$$

The joint conditional survival function of a frailty model is $S_i(t_1, ..., t_n) = exp[-z_i(H_{1,z}(t_1) + ... + H_{n,z}(t_n))]$ and the Laplace transform for the gamma frailty model is given by $\mathcal{L}(s) = E[-sZ] = (1 + \theta s)^{-1/\theta}$.
The joint conditional survival function can be obtained by integrating out the frailties using the frailty distribution as
$S_f(t_1, ..., t_n) = \int_0^\infty \exp[-z(H_{1,z}(t_1) + ... + H_{n,z}(t_n))] f_Z(z) dz.$
This integral can be solved analytically for the gamma distribution resulting in

$$S_f(t_1, ..., t_n) = [1 + \theta(H_{1,z}(t_1) + ... + H_{n,z}(t_n))]^{-1/\theta}. \tag{5.15}$$

On the other hand, the marginal survival function can be obtained from the conditional frailty model using a similar derivation as for the joint survival function

$$\begin{aligned} S_{j,f}(t_j) &= \int_0^\infty \exp[-z(H_{j,z}(t_j)f_Z(z)]dz \\ &= [1+\theta H_{j,z}(t_j)]^{-1/\theta}. \end{aligned} \tag{5.16}$$

It follows that

$$H_{j,z}(t_j) = \frac{(S_{1,f}(t_j))^{-\theta}-1}{\theta}. \tag{5.17}$$

It is clear from equation 5.16 that the marginal survival functions arising from the conditional frailty model are also functions of parameter θ, which is not the case for the Clayton model in 5.14. In the frailty model, the frailty parameter appears in the marginal survival functions. The parameter θ related to the copula function appears in the expression for the marginal survival function in the frailty model, but it does not show up in the expression for the copula model. The marginal survival functions in the frailty model are obtained by integrating out the frailty from the conditional survival functions.
If we substitute 5.17 in 5.15, we get

$$S_f(t_1,...,t_n) = [(S_{1,f}(t_1))^{-\theta} + ... + (S_{n,f}(t_n))^{-\theta} - 1]^{-1/\theta}. \tag{5.18}$$

The association parameters for the two models are defined in a different way. In a shared frailty model, θ is the frailty variance which indicates unobserved heterogeneity between clusters which is also a measure of association. In other words, θ is also influenced by the marginal setting as we have seen in equation 5.16. On the other hand, in a Clayton-Oakes copula model θ is only a measure of association. The Kendall's tau τ for both gamma shared frailty and Clayton-Oakes models is given by $\tau = \frac{\theta}{\theta+2}$. The Kendall's tau is used to measure the dependence within the clusters.

The similarities and differences between the two models can be summarised in the next section.

5.1.3 Summary of the similarities and differences

5.1.3.1 Parameter estimates

Parameter estimates in the copula model are found by modelling separately the marginal survival functions in stage one and the copula function in stage two. In frailty model, the frailty parameter appears in the marginal survival functions, making separate estimation possible (Goethals, 2011).

5.1.3.2 Interpretation of covariate effects

Interpretation of covariate effects are interpreted at the conditional level in frailty model, while, in the copula the interpretation is at the marginal level (Goethals, 2011).

5.1.3.3 Association between observations

In copula model, the parameter θ represents association only while in frailty model it is a measure for both association and heterogeneity. The dependence structure in the shared frailty model is implied by the choice of a probability density function, whereas in the Archimedean copula the dependence structure is specified through generator functions. Furthermore, the correlation between survival times in the frailty model is modelled through the frailties, whereas in the copula model the association is modelled through the survival times themselves (Geerdens et al., 2016). The copula function allows us to model the dependence between survival times separately from their marginals.

5.1.3.4 Copula functions and marginal survival functions

The copula functions for Clayton-Oakes and shared frailty models used for the joint functions are similar, but the marginal survival functions are modelled differently. The marginal survival functions in a frailty model contain the association parameter θ, which is not the case in the copula model. The difference

in the parameters for the two models was also illustrated using the data set in the next section.

5.1.4 Data analysis

The data set contains 219704 individuals with 120033 clusters. The data set was divided into three sub-samples of equal number of clusters. Each sub-sample was analysed separately. There were 40011 clusters in each sub-sample and the number of individuals in each sub-sample varied.

5.1.4.1 Parameter estimates from shared frailty and Clayton-Oakes copula models

Table 5.1 shows the parameter estimates for the shared frailty model. The results based on the three sub-samples show that male children died at a higher rate than female children. Based on the year of birth, children born in 2014 and 2015 had a lower risk of death compared to those who were born in 2013. All three sub-samples showed that children who were part of twin were dying at a faster rate than singletons. With regard to birth order, children who were not the first borns had a lower risk of death compared to first borns.

Table 5.2 shows the marginal parameter estimates from Clayton-Oakes copula which were estimated in the first stage of a two-stage estimation procedure. These estimates were estimated under the working independence assumption in which clustering is not taken into consideration. From Table 5.2 one can see that in all three sub-samples, there is an increased risk of death in male children as compared to female children. We observe relative to children born in Eastern Cape that children from other provinces are less likely to survive. The results also show the decreased risk of death in children born in 2014 and 2015 as compared to those born in 2013. We also observe that being part of a twin reduces the chance of survival. The results show that a child who is born second,

third, fourth or fifth has a significantly higher chance of survival than a first-born child. This might be due to the fact that women became more educated and experienced about child birth and caring after their first experiences.

Table 5.1: Parameter estimates from the shared frailty model

Factors	Levels	Sub-sample 1				Sub-sample 2				Sub-sample 3			
		Coef	Hazard ratio	S_e	p-value	Coef	Hazard ratio	S_e	p-value	Coef	Hazard ratio	S_e	p-value
Gender	Female	Ref											
	Male	0.0662	1.0684	0.0487	< 0.0000	0.0976	1.1025	0.0514	< 0.0000	0.0850	1.0887	0.0535	< 0.0000
Province	Eastern Cape	Ref											
	Free State	0.5488	1.7312	0.1222	< 0.0000	0.6843	1.9825	0.1259	< 0.0000	0.7573	2.1326	0.1186	< 0.0000
	Gauteng	-0.0868	0.9169	0.0916	< 0.0000	0.0567	1.0585	0.0988	< 0.0000	0.2507	1.2849	0.0982	< 0.0000
	Kwazulu	-0.4366	0.6462	0.1013	< 0.0000	-0.2947	0.7447	0.1036	< 0.0000	-0.3544	0.7016	0.0997	< 0.0000
	Limpopo	0.3173	1.3734	0.0982	< 0.0000	0.4990	1.6471	0.1034	< 0.0000	0.3986	1.4897	0.1009	< 0.0000
	Mpumalanga	0.1736	1.1896	0.1143	< 0.0000	0.2588	1.2954	0.1189	< 0.0000	0.0961	1.1009	0.01174	< 0.0000
	Northern Cape	1.0673	2.9075	0.1391	< 0.0000	0.8710	2.3893	0.1557	< 0.0000	0.7907	2.2049	0.1449	< 0.0000
	North West	0.6253	1.8688	0.1121	< 0.0000	0.6906	1.9950	0.1199	< 0.0000	0.5600	1.7506	0.1169	< 0.0000
	Western Cape	-0.3355	0.7150	0.1168	< 0.0000	-0.3285	0.7200	0.1293	< 0.0000	0.0082	1.0082	0.1287	< 0.0000
Year	2013	Ref											
	2014	-0.1966	0.8215	0.1302	< 0.0000	-0.6113	0.5426	0.1586	< 0.0000	-0.5576	0.5726	0.1466	< 0.0000
	2015	-0.1704	0.8434	0.0778	< 0.0000	-0.2447	0.7830	0.0912	< 0.0000	-0.0830	0.9203	0.0958	< 0.0000
Twin		0.0974	1.1023	0.0572	< 0.0000	0.2014	1.2231	0.0690	< 0.0000	0.2845	1.3291	0.0815	< 0.0000
Order		-0.3547	0.7014	0.0826	< 0.0000	-0.4038	0.6678	0.0918	< 0.0000	-0.4072	0.6655	0.091	< 0.0000

Table 5.2: Marginal parameter estimates from the Clayton-Oakes copula model

Factors	Sub-sample 1				Sub-sample 2				Sub-sample 3			
	Coef	S_e	Robust S_e	p-value	Coef	S_e	Robust S_e	p-value	Coef	S_e	Robust S_e	p-value
Gender	0.0665	0.0469	0.0470	< 0.0000	0.0928	0.0487	0.0487	< 0.0000	0.0863	0.0523	0.0515	< 0.0000
Province	-0.0183	0.0107	0.0111	< 0.0000	-0.0361	0.0112	0.0118	< 0.0000	-0.0165	0.0124	0.0129	< 0.0000
Year	-0.0702	0.0387	0.0402	< 0.0000	-0.0984	0.0422	0.0431	< 0.0000	-0.0254	0.0468	0.0479	< 0.0000
Twin	0.0617	0.0542	0.0559	< 0.0000	0.1340	0.0626	0.0650	< 0.0000	0.2170	0.0783	0.0831	< 0.0000
Order	-0.4590	0.0860	0.0882	< 0.0000	-0.4590	0.0874	0.0848	< 0.0000	-0.5060	0.0878	0.0909	< 0.0000

5.1.4.2 Association measures for Clayton-Oakes and shared frailty model

Table 5.3: Association measures for Clayton-Oakes and shared frailty model

Sample	Number of observations	Clayton-Oakes model				Shared frailty model			
		θ	S_e	p-value	τ	θ	S_e	p-value	τ
1	75481	0.0515	0.0271	0.0576	0.0251	2.2406	0.3869	0.0000	0.5284
2	74271	0.0664	0.0304	0.0289	0.0321	3.2067	0.5134	0.0000	0.6159
3	69952	0.0364	0.0315	0.2480	0.0179	1.3670	0.4237	0.006	0.4060
Overall estimates									
$\hat{\theta}$		0.05173				2.2890			
SE		0.0171				0.2569			
p-value		0.0289				0.0000			
$\hat{\tau}$		0.0252				0.5190			

Table 5.3 shows the association measures for the Clayton and shared frailty models. The overall parameter estimates in the copula model are $\hat{\theta} = 0.0517$ with $\hat{\tau} = 0.0252$. In the shared frailty model the overall estimates are given by $\hat{\theta} = 2.2890$ with $\hat{\tau} = 0.5190$. We can clearly see that the overall estimates for the two models are quite different. The overall estimates for the frailty are much higher as compared to the estimates for the copula model. Based on the estimated value of Kendall's tau ($\hat{\tau}$), the two models show the presence of positive association between survival times of children in the same cluster. The difference between the overall estimates for the two models can be clearly seen using a bar chart in Figure 5.1.

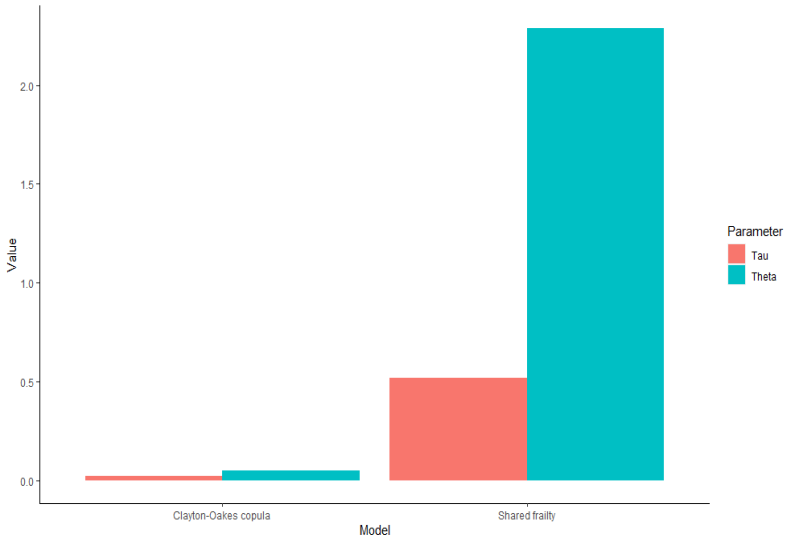

Figure 5.1: Bar chart showing association measures for the shared frailty and Clayton-Oakes copula model

We can clearly see from the bar chart that the two models are quite different. This is because the marginals are estimated separately in the copula model, while in the frailty model the marginals are included in the association parameter.

5.2 Discussion

It was revealed in our study that the two models are different. Tran et al. (2020) agreed that the copula and the frailty models are quite different in nature. Although the copula functions used in the two models are the same, their marginal functions are quite different. Our findings agree with the study conducted by Goethals et al. (2008). We have shown that the equivalence usually mentioned by a number of investigators does not hold because the modelling of the marginal survival function for the two models is done differently, which makes their joint survival functions to be different. The marginal survival functions for the frailty model contained θ, and this not the case with the copula model. The frailty model can be written in such a way that the joint survival function is expressed in terms of the marginal survival functions which takes a form of an Archimedean copula and that might be the reason why it is often stated that the shared frailty and the Clayton copula models are similar whereas they are not (Goethals et al., 2008).

We have further identified that the estimated association parameters $\hat{\theta}$ are quite different and that using $\hat{\theta}$ directly is not the right way to go about. It has been pointed out by Geerdens et al. (2016) that the association on the survival times of individuals in the same cluster are modelled indirectly through the use of frailties for the frailty model, while the associations in the copula model are modelled through the survival times themselves. That might also be the reason why their association parameters are quite different. This is also due

to the fact that the frailty parameter $\hat{\theta}$ in the frailty model is also influenced by the marginal setting and does not only depend on the association. On the other hand, the copula parameter estimate $\hat{\theta}$ in the copula model depends only on the association.

Furthermore, both models showed that gender, birth province, birth year, twin status and birth order were significant factors affecting under five child mortality in South Africa. Both models showed that the hazard of death is higher for males than for females and also higher for multiple births than singletons. The results from the two models also showed that the hazard of dying is lower when a mother had children previously than when a mother had no previous living children at all.

5.3 Conclusion

In this chapter, similarities and differences between the copula model and the frailty model were discussed. The main focus was on Clayton-Oakes copula and the shared frailty model. The two models showed that gender, birth province, birth year, twin and birth order were significant factors affecting under five child mortality in South Africa. It can be concluded from the shared frailty model as well as the copula model that there is positive correlation between survival times of children from the same mother. The copula functions for the shared frailty and Clayton-Oakes models used for the joint functions are similar, but the marginal survival functions for the two models are obtained in a different way. In the copula model, the parameter estimates are found by modelling the survival functions separately in the first stage and the copula function in the second stage, whereas the frailty parameter in a frailty model appears in the marginal survival functions. The results from our study showed that the estimates for theta ($\hat{\theta}$) and Kendall's tau ($\hat{\tau}$) for the two models were

quite different. We therefore conclude that the two models cannot be comparable as they are quite different in nature.

Chapter 6

General discussion and conclusion

6.1 Discussion

Under five mortality data sets have been analysed, presented and discussed in various chapters of this thesis. The study assessed survival patterns of under five children and examined determinants of under five mortality in South Africa by considering clustered survival models. We started with univariate survival models without considering clustering. Different techniques such as Cox Proportional hazard and logistic regression model were considered to determine factors contributing to child mortality and also those influencing probability of stillbirths.

We then formed clusters by using mother's identity and applied clustered survival models, specifically, Clayton-Oakes copula and shared frailty models. In Chapter 3, the shared frailty model was explored considering left truncation,

but excluding clusters of size one. Clusters of size one were excluded because the main focus was on links between siblings. In Chapter 4, our initial plan was to apply different Archimedean copula families to the whole data set, but that could not happen due to the size of our data set and lack of computer program. The data was too big to analyse at once and a lot of clusters contained only one individual. The only option left for us was to apply Clayton copula by splitting the data into pieces and then analysing each piece separately. We then combined estimates from each sub-sample to get the overall estimates such as association parameter, standard error and p-value. We used the two-stage estimation procedure in which the marginal parameters are estimated in the first stage under the working independence assumption and the association parameter estimated in the second stage.

In Chapter 5, we compared the shared frailty and Clayton copula models to determine if the two are equivalent, i.e., to identify similarities and differences between the two models. Our plan was to use all data points at once for a copula model in the first stage procedure. We wanted to estimate the fitted survival probability for each individual based on the marginal model from stage 1 and then plug-in at the second stage likelihood of the sub-group. In the second stage, the plan was to apply the splitting techniques to the data set that contains only clusters with at least two individuals in order to find the association in each sub-group. That could not happen again due to programming. We decided to analyse the copula model as we did in Chapter 4, i.e., we split up the data set containing at least two individuals into sub-samples and then apply two stage estimation procedure in each sub-sample. The estimates from each sub-sample were combined to obtain the overall estimates. In order to make a good comparison, we applied both shared frailty and Clayton copula to the same data set containing clusters with at least two individuals.

6.2 Thesis summary and concluding remarks

The thesis was organised into chapters in order to address the objectives indicated in Chapter 1, Section 1.5.2.

Chapter 1 dealt with historical background of the study area (South Africa) and gave a brief literature review about the importance of considering association when analysing clustered survival data. Data sets used to illustrate methodologies developed in various chapters of this thesis were also described. The purpose of the study including all variables used in the study were fully described in Chapter 1. The outline of the subsequent chapters in the rest of the thesis were also given.

The thesis has answered the objectives as stipulated in Chapter 1. In responding to the first and the second objectives, i.e., *to compare survival curves using non-parametric tests and to analyse the under five child mortality data set using marginal survival model*, Chapter 2 investigated the univariate survival models without taking clustering into consideration. The data set described in Chapter 1, Section 1.7.4.1 was used in all analyses included in this chapter. Models and procedures applied were Cox Proportional Hazard model, logistic regression model and non-parametric measures such as log rank test and KM estimator. The Cox PH model was included in order to determine the significant variables associated with child mortality. The results showed that gender, province and year significantly affected under five child mortality. The logistic regression model was applied to identify factors influencing probability of stillbirths. It was found that gender, province and year were the factors affecting stillbirth cases. To check if there were significant differences among survival experiences of covariates, KM estimator and log-rank test were considered. The KM plots showed that female children survived longer than male children and that children born in Eastern Cape lived longer than children

born in other provinces of South Africa. The KM results further revealed that children born in 2014 survived longer than children born in other years. The Logrank test statistically confirmed significant differences in the occurrence of death of different categories of gender, birth province and birth year.

To address the third and the fourth objectives, i.e., *to compare Cox proportional and shared frailty models and to model the association of individuals within a cluster using frailty models that also consider left truncation*, Chapter 3 investigated the importance of ignoring frailty. Using the data set described in Chapter 1, Section 1.7.4.2, the Cox PH model and the shared frailty model were compared to see their performance. It turned out that among those factors considered in the analysis, gender, province, year, order and twin status were significant contributors of under five child mortality in the study area. Heterogeneity between mothers and strong association between children from the same mother in the shared frailty model have improved the final results of the study. Furthermore, comparing these two models, it was found that the shared frailty model with the lowest LCV value, have provided a better fit for the study than the Cox PH model. Estimation methods for the shared frailty model were discussed and the drawbacks of the commonly used estimation methods were highlighted. The full penalised likelihood estimation method to estimate model parameters was used and the analysis was done using the R package called *frailtypack*.

In responding to the fifth and the sixth objectives, i.e., *to explore association within a cluster by using copula models and to apply sample splitting techniques in a survival analysis setting*, Chapter 4 investigated the two stage semi-parametric estimation approach with Cox marginals. We used the *two.stage* function available in *timereg* R package to fit the Clayton-Oakes copula model. In the first stage, the parameters of the marginal survival functions were es-

timated and then inserted in the copula function. In the second stage, the parameters of the copula functions together with association parameters, were estimated. The Clayton-Oakes model was illustrated using the data set described in Chapter 1, Section 1.7.4.2, but excluded clusters of size 1 and left truncated observations. Due to the size of the data set, Clayton model could not be analysed at once. We partitioned the data set into three sub-samples of equal number of clusters, but with different sample sizes and then analysed each sub-sample separately. We applied a sample splitting technique to combine the association parameters, standard errors and p-values for the three sub-samples in order to obtain the overall association parameter, overall standard error and overall p- value. We used the overall estimates to make the final conclusion. The conclusion reached was that all covariates included in the Clayton copula model had an effect on survival of the children and that there is a link between individuals in the same cluster.

To address the seventh objective, i.e., *to compare Clayton-Oakes copula and shared frailty models with respect to how they handle association within a cluster*, we made a comparison between the shared frailty model and the Clayton-Oakes model in Chapter 5. To make a good comparison between the two models, we used the same data set without clusters of size 1 and left truncated observations. The data set was also partitioned into three sub-samples as in Chapter 4 and the three sub-samples were analysed separately. We have seen that the copula functions for the joint survival functions were similar for the two models, but their marginal survival functions were modelled in a different way. Our study also proved that the parameter estimates for the two models were quite different and we reached a conclusion that the two models cannot be compared as they were different in nature.

The eighth objective, i.e., *to assess if there are unobserved genetic and envi-*

ronmental factors that aggravate under five child mortality was addressed in Chapters 2, 3 and 4. Gender, province, year, birth order and twin status were found to be factors aggravating under five child mortality.

6.3 Contributions of the study

It has been pointed out in the problem statement that the most of previous studies on child mortality used logistic regression and Cox models. In the same section some of the weaknesses of the two methods were also highlighted, i.e., logistic regression does not consider the time to event variable and Cox proportional hazards model assumes that survival times of individuals are independent. The major contribution of this thesis is in the application of shared frailty and copula survival models that take care of clustering and left truncation in modelling the under five child mortality in RSA. The specific contributions are:

1. Introduction of sample splitting techniques in a survival analysis setting.

2. Modelling association of siblings using a non-parametric penalised likelihood estimation approach.

3. Improvement of the existing copula models to allow clusters of large and unequal sizes using an R package called *timereg*.

4. Extension and improvement of the existing models in the literature such as Clayton copula which was applied to the data set with large and different cluster sizes.

6.4 Summary of the key findings

The main focus of this thesis was on methodological aspects of clustered survival models. We believe that the results of this study are the true reflection of under five mortality in South Africa and could be applicable anywhere with

similar setting.

Kaplan Meier curves show that survival times of male children were shorter than their female counterparts. Gender, province, year and twin status were found to be highly associated with the increased risk of death in children under the age of five. The majority of stillbirths were males and most of them were found in the Free State province. This might be due to the fact that Free State province is situated in the rural part of South Africa with limited health facilities and health practitioners. Twins and first-born children were more likely to die. This shows a need to train women on child caring before they could even become pregnant.

Due to the fact that a positive correlation existed in the data, using clustered survival models was the right way to analyse the data set than using the Cox model which does not consider clustering. The estimates for theta $\hat{\theta}$ and tau $\hat{\tau}$ for the shared frailty and the Clayton-Oakes copula were quite different and therefore the two models cannot be comparable as they are quite different in nature.

6.5 Limitations of the thesis

The data set used to analyse clustered survival models could not represent the random sample of the entire population in South Africa due to the number of cases with missing mother's identity number and other missing information necessary for clustered survival models. There was missing death information in the data set for children who died between 2010 and 2012 and that resulted in left truncation. Individuals with left truncated survival times were not considered in the copula model because the software could not allow us to capture them.

Some important factors that might affect under five mortality could not be addressed due to unavailability of information in the data sets obtained from Stats SA. Factors affecting under five child mortality are vital for policy makers to plan and draft policy to reduce child mortality.

Due to unavailability of software and volume of our data set, we could not analyse all data points at once in Chapter 4 and Chapter 5. Our initial plan was to use all data points at once and at the same time when calculating the marginal parameters in stage 1 of the Clayton copula model and then use the sampling splitting techniques on all clusters with at least two individuals to find association parameter in each sub-sample. We could not find a way of plugging-in the fitted survival probability for each individual in the second stage likelihood. Our future plan is to develop a programme that can allow us to plug-in the fitted survival probabilities of individuals in the likelihood of stage 2 of the estimation process.

6.6 Future research directions

The following possible research directions are suggested:

- In this thesis, the gamma distribution has been used as a frailty distribution. In future, it might be of interest to consider other frailty distributions and compare their performance.
- Future research might also consider other methods of parameter estimation in frailty modelling. In our case, full penalised likelihood estimation method was considered.
- In future studies, factors associated with child mortality risk other than those considered in this study can be included to the model because there

might as well be other factors that are associated with under five mortality.

- The shared frailty model presented in Chapter 3 of this thesis is limited to one level of clustering. That is, children are clustered within their mother. It would be a good idea in future to incorporate two levels of clustering in a shared frailty model.

- In copula modelling, individuals with truncated survival times were not considered. It might also be a good idea to consider other software that can allow truncated survival times to be captured in the model.

- The Bayesian survival analysis techniques which considers prior information when estimating parameter estimates can also be experimented to model under five child mortality in a situation of left truncation in future researches.

In general, researches for under five mortality should be considered as an ongoing process to reduce child mortality rate in South Africa. Factors contributing to high under five mortality should be given special attention and improvement of quality care for pregnant women is needed.

List of References

Abdulkarimova, U. (2013), Frailty Models for Modelling Heterogeneity, PhD thesis, McMaster University.

Alam, N., Van Ginneken, J. K. and Bosch, A. M. (2007), 'Infant mortality among twins and triplets in rural bangladesh in 1975–2002', *Tropical Medicine & International Health* **12**(12), 1506–1514.

Aminu, M., Unkels, R., Mdegela, M., Utz, B., Adaji, S. and Van Den Broek, N. (2014), 'Causes of and factors associated with stillbirth in low-and middle-income countries: a systematic literature review', *BJOG: An International Journal of Obstetrics & Gynaecology* **121**, 141–153.

Andersen, E. W. (2005), 'Two-stage estimation in copula models used in family studies', *Lifetime Data Analysis* **11**(3), 333–350.

Balan, T. A. and Putter, H. (2020), 'A tutorial on frailty models', *Statistical Methods in Medical Research* p. 0962280220921889.

Bamford, L., McKerrow, N., Barron, P. and Aung, Y. (2018), 'Child mortality in south africa: Fewer deaths, but better data are needed', *South African Medical Journal* **108**(3), 25–32.

Basar, E. (2017), 'Aalens additive, cox proportional hazards and the cox-aalen model: Application to kidney transplant data', *Sains Malaysiana* **46**(3), 469–476.

Bhattacharyya, R. and Pal, A. (2012), 'Stillbirths in a referral medical college hospital, west bengal, india: A ten-year review', *Journal of Obstetrics and Gynaecology Research* **38**(1), 266–271.

Bouwmeester, W., Twisk, J. W., Kappen, T. H., van Klei, W. A., Moons, K. G. and Vergouwe, Y. (2013), 'Prediction models for clustered data: comparison of a random intercept and standard regression model', *BMC medical research methodology* **13**(1), 19.

Brown Jr, B. W., Hollander, M. and Korwar, R. M. (1973), Nonparametric tests of independence for censored data with application to heart transplant studies, Technical report, Florida State University Tallahassee Department of Statistics.

Bryce, J., Terreri, N., Victora, C. G., Mason, E., Daelmans, B., Bhutta, Z. A., Bustreo, F., Songane, F., Salama, P. and Wardlaw, T. (2006), 'Countdown to 2015: tracking intervention coverage for child survival', *The Lancet* **368**(9541), 1067–1076.

Cesar, C. C., Palloni, A. and Rafalimanana, H. (1997), *Analysis of child mortality with clustered data: a review of alternative models and procedures*, Center for Demography and Ecology, University of Wisconsin–Madison.

Chen, Z. (2014), 'A flexible copula model for bivariate survival data'.

Cherubini, U., Luciano, E. and Vecchiato, W. (2004), *Copula methods in finance*, John Wiley & Sons.

Clayton, D. G. (1978), 'A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence', *Biometrika* **65**(1), 141–151.

Djehiche, B. and Hult, H. (2004), 'An introduction to copulas with applications', *Svenka Akturiforeningen, Stockholm* .

Duchateau, L. and Janssen, P. (2007), *The frailty model*, Springer Science & Business Media.

Dufresne, F., Hashorva, E., Ratovomirija, G. and Toukourou, Y. (2018), 'On age difference in joint lifetime modelling with life insurance annuity applications', *Annals of Actuarial Science* **12**(2), 350–371.

Durante, F. and Sempi, C. (2010), Copula theory: an introduction, *in* 'Copula theory and its applications', Springer, pp. 3–31.

Emamverdi, G., Karimi, M. S., Firouzi, M. and Emdadi, F. (2014), 'An investigation about joint life policy's premium using copula; the case study of an insurance company in iran', *Asian Journal of Research in Business Economics and Management* **4**(10), 222–242.

Ezeh, O. K., Agho, K. E., Dibley, M. J., Hall, J. J. and Page, A. N. (2015), 'Risk factors for postneonatal, infant, child and under-5 mortality in nigeria: a pooled cross-sectional analysis', *BMJ open* **5**(3), e006779.

Feresu, S. A., Harlow, S. D., Welch, K. and Gillespie, B. W. (2004), 'Incidence of and socio-demographic risk factors for stillbirth, preterm birth and low birthweight among zimbabwean women', *Paediatric and perinatal epidemiology* **18**(2), 154–163.

Frees, E. W., Carriere, J. and Valdez, E. (1996), 'Annuity valuation with dependent mortality', *Journal of risk and insurance* pp. 229–261.

Frees, E. W. and Valdez, E. A. (1998), 'Understanding relationships using copulas', *North American actuarial journal* **2**(1), 1–25.

Gachau, W. S. (2014), Frailty models with applications in medical research: observed and simulated data, PhD thesis, University of Nairobi.

Geerdens, C., Claeskens, G. and Janssen, P. (2016), 'Copula based flexible modeling of associations between clustered event times', *Lifetime data analysis* **22**(3), 363–381.

Genest, C., Ghoudi, K. and Rivest, L.-P. (1995), 'A semiparametric estimation procedure of dependence parameters in multivariate families of distributions', *Biometrika* **82**(3), 543–552.

Genest, C. and MacKay, R. (1986), 'Archimedean copulas and bivariate families with continuous marginals', *Canadian Journal of Statistics* **14**, 145–159.

Georges, P., Lamy, A.-G., Nicolas, E., Quibel, G. and Roncalli, T. (2001), 'Multivariate survival modelling: a unified approach with copulas'.

Glidden, D. V. (2000), 'A two-stage estimator of the dependence parameter for the clayton-oakes model', *Lifetime Data Analysis* **6**(2), 141–156.

Glidden, D. V. and Vittinghoff, E. (2004), 'Modelling clustered survival data from multicentre clinical trials', *Statistics in medicine* **23**(3), 369–388.

Goethals, K. (2011), Multivariate survival models for interval-censored udder quarter infection times, PhD thesis, Ghent University.

Goethals, K., Janssen, P. and Duchateau, L. (2008), 'Frailty models and copulas: similarities and differences', *Journal of Applied Statistics* **35**(9), 1071–1079.

Graner, S., Klingberg-Allvin, M., Phuc, H. D., Krantz, G. and Mogren, I. (2009), 'The panorama and outcomes of pregnancies within a well-defined population in rural vietnam 1999–2004', *International journal of behavioral medicine* **16**(3), 269–277.

Guo, G. and Rodriguez, G. (1992), 'Estimating a multivariate proportional hazards model for clustered data using the em algorithm, with an application to

child survival in guatemala', *Journal of the American Statistical Association* **87**(420), 969–976.

Hanagal, D. D. (2011), *Modeling survival data using frailty models*, Chapman and Hall/CRC.

Hougaard, P. (1986), 'A class of multivanate failure time distributions', *Biometrika* **73**(3), 671–678.

Hougaard, P. (2012), *Analysis of multivariate survival data*, Springer Science & Business Media.

Hsieh, J.-J. (2010), 'Estimation of kendalls tau from censored data', *Computational statistics & data analysis* **54**(6), 1613–1621.

Islam, S., Islam, M. A. and Padmadas, S. S. (2010), 'High fertility regions in bangladesh: a marriage cohort analysis', *Journal of biosocial science* **42**(6), 705–719.

Jensen, H., Brookmeyer, R., Aaby, P. and Andersen, P. K. (2004), *Shared frailty model for left-truncated multivariate survival data*, University of Copenhagen. Department of Biostatistics.

Khan, J. R. and Awan, N. (2017), 'A comprehensive analysis on child mortality and its determinants in bangladesh using frailty models', *Archives of Public Health* **75**(1), 58.

King, M., Lodwick, R., Jones, R., Whitaker, H. and Petersen, I. (2017), 'Death following partner bereavement: A self-controlled case series analysis', *PloS one* **12**(3), e0173870.

Kleinbaum, D. G. (1998), 'Survival analysis, a self-learning text', *Biometrical Journal: Journal of Mathematical Methods in Biosciences* **40**(1), 107–108.

Knox, K. L., Bajorska, A., Feng, C., Tang, W., Wu, P. and Tu, X. M. (2013), 'Survival analysis for observational and clustered data: an application for assessing individual and environmental risk factors for suicide', *Shanghai archives of psychiatry* **25**(3), 183.

Kristiansen, L. C. (2012), Statistical analysis of familial aggregation of adverse outcomes, Master's thesis, Technical University of Denmark.

Legrand, C., Duchateau, L., Sylvester, R., Janssen, P., van der Hage, J. A., Van de Velde, C. J. and Therasse, P. (2006), 'Heterogeneity in disease free survival between centers: lessons learned from an eortc breast cancer trial', *Clinical trials* **3**(1), 10–18.

Li, G. and Wu, Z. (2018), An introduction to frailty models for multivariate survival data with application to martial dissolution and retirement, *in* 'PAA 2018 Annual Meeting', PAA.

Mabin, A. S., Gordon, D. F., Hall, M., Vigne, R., Bundy, C. J., Thompson, L. M., Cobbing, J. R., Lowe, C. C. and Nel, A. (2021), 'South africa', Encyclopedia Britannica. Accessed 4 May 2021.
URL: *https://www.britannica.com/place/South-Africa*

Madhi, S. A., Briner, C., Maswime, S., Mose, S., Mlandu, P., Chawana, R., Wadula, J., Adam, Y., Izu, A. and Cutland, C. L. (2019), 'Causes of stillbirths among women from south africa: a prospective, observational study', *The Lancet Global Health* **7**(4), e503–e512.

Mahfoud, M. and Michael, M. (2012), 'Bivariate archimedean copulas: an application to two stock market indices', *BMI Paper* .

Mahmood, S., Zainab, B. and Latif, A. M. (2013), 'Frailty modeling for clustered survival data: an application to birth interval in bangladesh', *Journal of Applied Statistics* **40**(12), 2670–2680.

Martinussen, T. and Scheike, T. H. (2007), *Dynamic regression models for survival data*, Springer Science & Business Media.

Masset, E. and White, H. (2003), 'Infant and child mortality in andhra pradesh: Analysing changes over time and between states'.

Massonnet, G., Janssen, P. and Duchateau, L. (2009), 'Modelling udder infection data using copula models for quadruples', *Journal of Statistical Planning and Inference* **139**(11), 3865–3877.

Mauguen, A. (2014), Prognosis of cancer patients: input of standard and joint frailty models, PhD thesis, Bordeaux.

Mills, M. (2011), *Introducing survival and event history analysis*, Sage Publications.

Moerbeek, M., van Breukelen, G. J. and Berger, M. P. (2003), 'A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies', *Journal of clinical epidemiology* **56**(4), 341–350.

Molenberghs, G., Verbeke, G. and Iddi, S. (2011), 'Pseudo-likelihood methodology for partitioned large and complex samples', *Statistics & probability letters* **81**(7), 892–901.

Munyamahoro, F. (2016), 'Copula-based dependence measures for under-five mortality rate in rwanda', *Archivos De Medicina* **2**(4), 34.

Nasejje, J. (2013), Application of Survival Analysis Methods to Study Under-five Child Mortality in Uganda, PhD thesis, University of KwaZulu-Natal, Pietermaritzburg.

Nelsen, R. B. (2007), *An introduction to copulas*, Springer Science & Business Media.

Nguti, R. (2003), Random effects survival models applied to animal breeding data, PhD thesis, UHasselt Diepenbeek.

Niragire, F., Wangombe, A. and Achia, T. N. (2011), 'Use of the shared frailty model to identify the determinants of child mortality in rwanda', *Rwanda Journal* **20**(1), 90–107.

Othus, M. and Li, Y. (2010), 'A gaussian copula model for multivariate survival data', *Statistics in biosciences* **2**(2), 154–179.

Park, H. (2013), 'An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain', *Journal of Korean Academy of Nursing* **43**(2), 154–164.

Parkes, C. M., Benjamin, B. and Fitzgerald, R. G. (1969), 'Broken heart: a statistical study of increased mortality among widowers', *Br med J* **1**(5646), 740–743.

Pebley, A. R. and Stupp, P. W. (1987), 'Reproductive patterns and child mortality in guatemala', *Demography* **24**(1), 43–60.

Prenen, L., Braekers, R. and Duchateau, L. (2017), 'Extending the archimedean copula methodology to model multivariate survival data grouped in clusters of variable size', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**(2), 483–505.

Rondeau, V. and Gonzalez, J. R. (2005), 'Frailtypack: a computer program for the analysis of correlated failure time data using penalized likelihood estimation', *Computer methods and programs in biomedicine* **80**(2), 154–164.

Rondeau, V., Mazroui, Y. and Gonzalez, J. R. (2012), 'frailtypack: an r package for the analysis of correlated survival data with frailty models using penalized likelihood estimation or parametrical estimation', *J Stat Softw* **47**(4), 1–28.

SA, S. (2019), *Statistical release P0302: Mid-year population estimates*, Statistics South Africa.

Sainani, K. (2010), 'The importance of accounting for correlated observations', *PM&R* **2**(9), 858–861.

Sastry, N. (1997), 'A nested frailty model for survival data, with an application to the study of child survival in northeast brazil', *Journal of the American Statistical Association* **92**(438), 426–435.

Schober, P. and Vetter, T. R. (2018), 'Survival analysis and interpretation of time-to-event data: the tortoise and the hare', *Anesthesia and analgesia* **127**(3), 792.

Seçkin, N. (2009), Determinants of infant mortality in Turkey, PhD thesis, MA Thesis]. Middle East Technical University.

Shih, J. H. and Louis, T. A. (1995), 'Inferences on the association parameter in copula models for bivariate survival data', *Biometrics* pp. 1384–1399.

Sklar, M. (1959), 'Fonctions de repartition an dimensions et leurs marges', *Publ. inst. statist. univ. Paris* **8**, 229–231.

Stone, M. (1974), 'Cross-validatory choice and assessment of statistical predictions', *Journal of the Royal Statistical Society: Series B (Methodological)* **36**(2), 111–133.

Sullivan, J. M., Rutstein, S. O. and Bicego, G. T. (1994), *Infant and child mortality*, number 15, Macro International Calverton, Maryland.

Tang, C. (2014), 'Analyses of 2002-2013 chinas stock market using the shared frailty model'.

Therneau, T. M. and Grambsch, P. M. (2013), *Modeling survival data: extending the Cox model*, Springer Science & Business Media.

Tibaldi, F. (2004), Modeling of correlated data and multivariate survival data, PhD thesis, UHasselt Diepenbeek.

Tran, T. M., Abrams, S. and Braekers, R. (2020), 'A general frailty model to accommodate individual heterogeneity in the acquisition of multiple infections: An application to bivariate current status data', *Statistics in medicine* **39**(12), 1695–1714.

Trivedi, P. K., Zimmer, D. M. et al. (2007), 'Copula modeling: an introduction for practitioners', *Foundations and Trends® in Econometrics* **1**(1), 1–111.

UNDP (2019), *Sustanable development goals -Good health and wellbeing*.

Vaupel, J. W., Manton, K. G. and Stallard, E. (1979), 'The impact of heterogeneity in individual frailty on the dynamics of mortality', *Demography* **16**(3), 439–454.

Ward, A. (1976), 'Mortality of bereavement.', *Br Med J* **1**(6011), 700–702.

Worku, Z. (2009), Factors that affect under-five mortality among south african children: analysis of the south african demographic and health survey data set, *in* 'proceedings of the world congress on engineering and computer science', Vol. 2, pp. 1–3.

Zhenzhen, Z. (2000), 'Social–demographic influence on first birth interval in china, 1980–1992', *Journal of biosocial science* **32**(3), 315–327.

Zike, D. T., Fenta, H. M., Workie, D. L. and Swain, P. K. (2018), 'Determinants of under-five mortality in ethiopia: An application of cox proportional hazard and frailty models.', *Turkiye Klinikleri Journal of Biostatistics* **10**(2).

Chapter 7

Appendices

SOME SELECTED R CODES

7.1 R code for Chapter 2

In this section, R codes used to analyse data in Chapter 2 are given.

7.1.1 Code for KM curve

```
#selecting non-stillbirths
group2 <- subset(dataset1,dataset1$Stillbirth==0)

# The overall K-M curve excluding stillbirths
kaplan1a<- survfit (Surv(group2$Time,Status)~1,data=group2)
summary(kaplan1a)
plot(kaplan1a,xlab="Time until death (in days)",
ylab="Survival probability (%)", ylim=c(0.95,1))
title("KM survival plot for all individuals excluding still births")
```

7.1.2 Code for Log-rank test

```
attach(group2)
```

```
survdiff(formula=Surv(Time,Status) ~ Province, data =group2)
survdiff(formula=Surv(Time,Status) ~ Gender, data =group2)
survdiff(formula=Surv(Time,Status) ~ Year, data =group2)
```

7.1.3 Code for Cox PH model

```
#selecting non-stillbirths
group2 <- subset(dataset1,dataset1$Stillbirth==0)

#selecting baseline reference groups
group2$Gender<-relevel(group2$Gender,ref = "Female")
group2$Province<-relevel(group2$Province,ref = "Limpopo")

#COX PH model
coxph(formula=Surv(Time,Status) ~ Gender + Province+Year+
Gender* Province+Gender*Year+Province*Year, data =group2)
```

7.1.4 Code for logistic regression model

```
attach(dataset1)

#selecting baseline reference groups
dataset1$ Gender<-relevel(dataset1$ Gender,ref = "Female")
dataset1$Province<-relevel(dataset1$Province,ref = "Limpopo")

#Logistic regression model
logistic18 <- glm(Stillbirth ~ Gender + Province+Year
+Gender*Province+Province*Year,
family=binomial(link='logit'),data=dataset1)
summary(logistic18 )
```

7.2 R code for Chapter 3

In this section, R codes used to analyse data in Chapter 3 are given.

7.2.1 Code to create truncation time

```
#Creating truncation time#
attach(data2)
library(lubridate)
data2$TrunYear[is.na(data2$TrunYear)] <- 2013
data2$TrunMonth[is.na(data2$TrunMonth)] <- 01
data2$TrunDay[is.na(data2$TrunDay)] <- 01

data2$DateTrun <- paste(data2$TrunYear,
data2$TrunMonth, data2$TrunDay,sep = "-")

data2$DateBirth <- paste(data2$Year,
data2$BirthMonth, data2$BirthDay,sep = "-")

data2$DateTrun <- as.Date(strptime(data2$DateTrun,"%Y-%m-%d" ))
data2$DateBirth <- as.Date(strptime(data2$DateBirth, "%Y-%m-%d"))
TrunTime <- difftime(data2$DateTrun, data2$DateBirth, units = "days" )
```

7.2.2 Code for Penalized Cox model

```
# Creating clusters#
data2$clusterid <- data2 %>% group_indices( Mother_ID)

# Eliminating clusters of size 1
tt <- table(data2$clusterid)
data2a <- subset(data2, clusterid %in% names(tt[tt > 1]))

#Penalized Cox model#
```

```
Coxph(Surv(TrunTime, Time, Status) ~ Gender + Province +
    factor(Year) + Twin + order, data = data2a)
```

7.2.3 Code for shared frailty model

```
#Penalized shared frailty model#

frailtyPenal(formula = Surv(TrunTime, Time, Status) ~ Gender +
    Province + factor(Year) + Twin + order, data = data2a
    n.knots = 7, kappa = 10000)
```

7.3 R code for Chapter 4

In this section, R codes used to analyse data in Chapter 4 are given.

7.3.1 Code for copula model with Cox proportional hazards model as marginal

```
## Copula model with Cox proportional hazards model as
marginal ##

# Step 1: Marginal model
result1b <- cox.aalen(Surv(Time,Status)~prop(Gender)+
prop(Province)+prop(Year)+prop(order)+prop(Twin)+
cluster(clusterid),data=data2a,resample.iid=1)

# Step 2: Estimation of association parameter
result1bb<-two.stage(result1b,data=data2a,step=0.1,theta=0.5,Nit=500)
summary(result1bb)
```