UNIVERSITY OF GHANA

DEPARTMENT OF STATISTICS

MODELLING THE RISK FACTORS OF NEONATAL MORTALITY USING SURVIVAL ANALYSIS

BY

DANIEL ABOTWE DZIMAH

10507613

THIS THESIS IS SUBMITTED TO THE UNIVERSITY OF GHANA, LEGON IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE AWARD OF A MPHIL IN STATISTICS DEGREE

JUNE 2016

DECLARATION

CANDIDATE'S DECLARATION

This is to certify that this thesis is the result of my own research work and that no part of it has been presented for another degree in this university or elsewhere.

Signed: ..... Date: .....
Daniel Abotwe Dzimah (10507613)
(Candidate)

SUPERVISORS' DECLARATION

We hereby certify that this thesis was prepared from the candidate's own work and supervised in accordance with guidelines on supervision of thesis laid down by the University of Ghana.

Signed: ..... Date: .....
Dr. Kwabena Doku-Amponsah
(Principal Supervisor)

Signed: ..... Date: .....
Dr. Samuel Iddi
(Co - Supervisor)

ABSTRACT

Several strategies have been put in place in an attempt to reduce childhood mortality in Ghana, however the proportions of death among neonates are still quite high. The study therefore seeks to model neonatal mortality using survival analysis approach. The data used for the study was obtained from the neonate's folders at St. Jude Hospital in Obuasi in the Ashanti Region between January 1, 2012 and December 31, 2015. Data on maternal characteristics was also obtained. Neonates who were born before the 28^{th} day and those who have experienced the event (death) were considered for the study. The study employed the Kaplan Meier (K-M) and Log rank test for the descriptive analysis.

The Cox PH and Parametric models (Exponential, Weibull, Gompertz, Log-logistic and Log-normal) were fitted to the neonatal data and their results were compared using the AIC to determine the best model to explain survival of neonates. A semi parametric shared frailty model was also fitted to the data to examine whether there are unobserved heterogeneity among neonates at the community level. The Proportional Hazards assumption was checked using both graphical methods and the PH assumption test based on the Schoenfeld residual and was observed that the PH assumption was not violated. Results from the study showed that hazard ratios for the PH models (Cox, Exponential, Weibull and Gompertz) were similar, however comparison of the PH models using the AIC showed that the Gompertz PH model best fit the data.

A comparison of AFT models (Weibull, Exponential, Lognormal, Gompertz, and Log logistic) also showed that the Lognormal AFT fit the data best. A comparison of the best PH (Gompertz PH) and AFT (Lognormal AFT) model using the AIC showed that the Gompertz PH is the best model in predicting neonatal survival. Parity, Apgar score 1, birth weight and

place of residence were significantly related with neonatal mortality. A comparison of the shared frailty models (Cox, Exponential, Weibull, Gompertz, Lognormal and Log-logistics) using AIC revealed that exponential distribution with Gamma frailty is the best model for checking the unobserved heterogeneity in the data. Unobserved heterogeneity in categories of neonates based on place of residence was found.

ACKNOWLEDGEMENTS

My sincere thanks to Dr. Kwabena Doku-Amponsah and Dr. Samuel Iddi for being very supportive supervisors. Their tireless guidance and support made this work possible.

Special thanks to Mr. Bernard Bansah of St. Jude Hospital Hospital for his immense support during the data gathering. Finally, I appreciate all those who in one way or the other contributed to this project.

DEDICATION

I dedicate this work God Almighty for the divine knowledge, wisdom and strength given me to go through this research successfully.

Also to Mrs. Veronica Naa Lamiley Dzimah and my children Barack Daniel Anfo Dzimah and Michelle Akua-Leetso Dzimah.

Table of Contents

List of Tables

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIC	Akaike's Information Criterion
AFT	Accelerated Failure Time
BIC	Bayesian Information Criterion
CPR	Cardio-Pulmonary Resuscitation
CS	Caesarean Section
EM	Expectation Maximization
GDHS	Ghana Demographic Health Survey
HR	Hazard Ratio
KM	Kaplan-Meier
LBW	Low Birth Weight
MCEM	Markov Chain Expectation Maximization
MCMC	Markov Chain Monte Carlo
MDG	Millennium Development Goal
NDHS	Nigeria Demographic Health Survey
NICU	Neonatal Intensive Care Unit
PH	Proportional Hazard
PPL	Penalized Partial Likelihood
PTB	Pulmonary Tuberculosis Bacteria
RDHS	Rwanda Demographic Health Survey
SVD	Spontaneous Vaginal Delivery
UDHS	Uganda Demographic Health Survey

CHAPTER ONE

1.1 Introduction

Neonatal mortality as defined by Lawn et al. (2005) is the probability of a baby dying between the first day of birth and 28 days of life. Put differently, all deaths taking place between the 28th days of life is considered as neonatal mortality (Araújo et al. 2005). Neonatal mortality can either be considered as early or late. Early neonatal mortality period means that the death of the neonate occurred before the seventh day of life while late neonatal period means that the death of the neonate occurred after the seventh day of life. It is important to mention that the subdivision of neonatal mortality period is very necessary as the causes of the early and late are distinct. The neonatal period (from birth to 28th day of life) is normally considered to be the most vulnerable and high-risk time in the neonate life because of the highest mortality and morbidity incidence in human life during that period. According to Yinger & Ransom (2003) during this period the neonate risk of death is almost 15 times more than any other time before the first birthday.

Neonatal mortality continues to be thought of one amongst the most important public health issues worldwide as a result of it accounts for over 60% of newly born deaths before their initial birthday (UNICEF., 2008). Kassar et al. (2013) posited that post-neonatal deaths have seen a major decline, whereas baby deaths have had associate accelerated increase and these represent the most part of child mortality within the world. About two-thirds of infant mortality death is noted to occur within the initial month of lifetime of that over common fraction die within the initial week and even among them common fraction die within 24 hours (Lawn, McCarthy, & Ross, 2013). Rajaratnam et al. (2010) stipulated that of world's 7.7 million deaths among infants below five years, 3.1 million happened between birth and

first month of life (neonatal death), 2.1 million occurred between one month to one year (post neonatal death) and 2.3 million occurred between the ages of one to four years (childhood death).

Lawn et al. (2005) also noted that mortality rate is highest during early neonatal period (25-40 percent), however it decreases as the age of the neonate advances in number of days. A report by Peristat (2008) suggested however that this trend has shifted from early to late neonatal period among western countries. Globally, developing countries are considered to have high rate of neonatal mortality. For instance Parlato et al. (2004) indicated that the average neonatal mortality rate among developing countries is over eight times (33/1,000 live births) more than what exists in developed countries (4/1000 live births). Reports by the Safer (2007) revealed that Africa had the highest risk of neonatal mortality (41/1,000 live births) with sub-Saharan African countries (Central, Western and Eastern) recording between 42 and 49/1,000 live births while south-central countries recorded about 43/1,000 live births.

Though Ghana has made significant effort in reducing neonatal mortality, the current rate is still unacceptable. The Ghana Statistical Service Report (2015) on Ghana Health Survey revealed that the neonatal mortality rate for Ghana increased from 29 per 1,000 live births in 2003 to 33 per 1,000 live births in 2008. However, in 2014 the neonatal mortality rate reduced to 30 per 1,000 live births. They postulated that about 68 percent of all deaths occurring among children under 5 years take place before a child's first birthday with 48 percent taking place during the first month of life.

Many factors have been linked to the high neonatal mortality rate and these include complications of pre-term birth, birth asphyxia, sepsis and pneumonia (Black, Morris, &

Bryce, 2003). Other factors include poor maternal health, complications during pregnancy and delivery, first critical hours after birth, inadequate care during pregnancy, poor hygiene and lack of new born care. The World Organization Health (2003) posited that nutritional status, early childbearing, status of mother's in society, harmful practices and too many closely spaced pregnancies are some of the factors contributing to the high neonatal mortality.

The Millennium Development Goal four (MDG 4) indicates the need to reduce infant mortality by two-thirds by the year 2015. Several questions however have been raised regarding the attainment of this goal. Yinger & Ransom (2003) are of the opinion that unless neonatal mortality is reduced to half, it will be impossible to meet the MDG 4. World Health Organization (2012) hinted that the post 2015 agenda of MDG's is likely to include goals which aim at reducing neonatal mortality. The causes of neonatal deaths normally differ from region to region due to the differences in availability of medical resources, cultural practices, and other social issues (Kassar et al., 2013). Without a full disclosure of the causes of neonatal deaths, it would be very difficult for government and health authorities to plan and provide appropriate interventions to reduce neonatal deaths in Ghana. The study therefore examines the factors responsible for neonatal mortality in Ghana.

1.2 Problem Statement

The survival of new born babies is of great interest to the world and especially to a developing country like Ghana. Though Government has made frantic effort in reducing infant mortality, neonatal mortality has however not seen a major improvement. Results from the six GDHS surveys conducted between 1988 and 2014 has shown an increasing and decreasing trend over the past 15 years. The 2014 GHDS reported a neonatal mortality rate of 29/live births for 2003, 33/1,000 live births for 2008 and 30/1,000 live births for 2014. The

2014 GHDS document revealed that sixty-eight percent of all deaths among infants below the age of five in Ghana occurs before their first birthday, with 48 percent taking place during the first month of life. With this rate of progress it would be impossible to achieve the MDG of reducing child mortality to one-third. It is against this backdrop that the study sought to examine the risk factors related to neonatal deaths.

1.3 Objectives of the Study

1.3.1 General Objective

The main objective of the study is to model the risk factors associated with neonatal mortality using survival analysis approach.

1.3.2 Specific Objectives

1. To identify the risk factors associated with neonatal mortality.
2. To compare the results of the semi-parametric cox proportional hazards model and parametric models (Exponential, Weibull, Gompertz, Lognormal and Log-logistics).
3. To investigate the existence of unobserved heterogeneity among neonates with emphasize on place of residence using shared gamma frailty models.

1.4 Significance of the Study

The results obtained from this study would be of interest to practitioners in the health fraternity and policy makers such as the government, World Health Organization, researchers and the general public in the following ways: First of all, the study would help in identifying the risk factors of neonatal mortality in the Ghana. This will be very useful to government when planning on how to reduce or avoid neonatal deaths in the country. The general public

would also be aware of the major determinant of neonatal deaths. Secondly, the results will reveal the real impact of these risk factors on neonatal mortality. Thus it will find out the extent at which these factors are more life threatening than the others. Finally, the results obtained will inform academia and research by adding up to existing knowledge on the determinant of neonatal mortality in Ghana. It will also serve as a basis for further research for both academic researchers and health practitioners.

1.5 Scope and Methodology

This study was conducted using data from St. Jude Hospital at Obuasi in the Ashanti region of Ghana. It is one of the hospitals in Obuasi that has a well-equipped neonatal intensive care unit. Patients from both the urban centres and the rural areas of Obuasi visit this facility. The study involves neonates who were either born or died at the St. Jude Hospital. A secondary data was obtained from the hospitals Biostatistics Department. The researcher reviewed folders of all children aged less than a month between 2012 and 2015 and used a simply sampling method in gathering the data. The outcome variable for this study was the time until a neonate death is reached as reported by the hospital. The predictor or confounding variables used for the study were mode of delivery, neonatal sex, Apgar score, birth weight of neonates, gestational age, parity, maternal age, complications during delivery and place of residence.

The study employed semi-parametric cox proportional hazards model, PH and AFT parametric models and the semi-parametric frailty models. The Akaike's Information Criterion (AIC) was used to compare the efficiency of semi-parametric and parametric models. The analysis was carried out using R software and STATA version 13.

1.6 Organisation of the study

The study has five main chapters. Chapter one focused on the introduction, problem statement, objectives, significance of the study, scope and methodology and organization of the study. Chapter two also gave an in-depth knowledge about the subject matter as well as a review of related literature. Chapter three presents the detailed methodology used for the study whiles chapter four covered the data analysis, presentation and discussion of the results. The final chapter (chapter five) presents the summary, conclusion and recommendations of the study.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter reviews the relevant literature on the overview of neonatal mortality as well as related works conducted by other researchers on the subject area. Thus the literature review will address the various issues under investigation, providing a clearer background to the problem studied.

2.2 Overview

Though children serve as the human resource base of every nation, they are however more prone to diseases and other health related risks. Neonatal death therefore deprives a nation of its economic labour force thus leading the country into human resource crisis and slows development. According to World Health Organization (2012) under-five mortality seems to be decreasing in almost all the countries, however neonatal mortality which is a component of under-five mortality continues to increase causing countries to pay more attention on it. Every country aspires to minimize or completely eliminate neonatal mortality because a decline in mortality would result in accelerated growth.

Neonatal mortality has received more attention from researchers due to the role it plays in an economy. Research has shown that the study of mortality levels, trends and patterns helps to provide information on the state of health of a country, which serves as a measure of living standards in the country. It also brings to light the social differences that exist within a particular community or society and finally it provides information on the population's future growth potential. Knowledge about a country's mortality situation is therefore relevant for

effective development planning. The need for the evaluation of country's neonatal mortality has therefore become necessary.

A framework for determining infant survival was provided by Mosley & Chen (1984) for developing countries and they were grouped it into five groups namely: maternal factor (birth interval, age and parity,); environmental contamination (food, air, water etc.); personal illness control (preventive measures and treatments); nutrient deficiency and injury (accidental, intentional). Gandotra & Das (1990) further re-grouped the determinant factors into five broad categories and they are medical factors, sanitation and hygienic factors, socio-economic factors, environmental factors and demographic factors. These factors will be discussed in the next subsection.

2.3 Risk factors of Neonatal Mortality

2.3.1 Maternal Characteristics

According to Victora, Black, & Bryce (2007) maternal factors like marital status, maternal age during childbirth, birth order, parity, birth interval, birth type (singleton or multiple) and accessibility of healthcare plays a critical role in determining the survival of a neonate. Other notable research have also shown the effects of mother's age and birth order on neonatal mortality are almost similar. Thus first born neonates and those with high birth order were seen to have significant higher mortality than the intermediary births. Similarly mothers who give birth below 20 years and above 30 years could are more probable of losing their offspring than mothers between the ages of 20-30 years. Powell (1988) posited that marital status have significant influence of the survival of a neonate. According to Powell (1988) marital status is sometimes used as a proxy to assess overall economic status, housing and nutrition.

Van den Broek & Graham (2009) also noted that there is an association between neonatal mortality and increased level of maternal education. To Pickett & Hanlon (1990) mothers who are educated have a better chance of being employed or even access income. It is important to state therefore that neonates whose mothers are educated are likely to have high chance of survival than neonates whose mothers are not educated. This is because mothers with high level of education have access to better health care than those without education.

The knowledge aspect of being educated always makes the difference between two mothers. Educated mothers usually have the knowledge on providing nutritious diet, using a piped water source, washing of hands before eating, using better toilet facility, taking care of the sick and making appropriate use of health care services. Other studies have also shown that household factors such as place of residence, number of rooms, number of people living in the house and the source of drinking water have effect on neonatal mortality. Though these factors indirectly causes neonatal death, they however use other mediums like nutrition, access to health services, environment, etc. which lead to neonatal mortality.

2.3.2 Biological and Environmental Characteristics

According to Lawn et al. (2005) biological factors such as preterm birth, sepsis/pneumonia, asphyxia and tetanus are the major causes of neonatal mortality. Liu et al. (2012) also posited that preterm, intrapartum related complications and sepsis/meningitis/tetanus are main biological factors which cause neonatal mortality. Goldenberg et al. (2008) added that inflammation of the uterus, bleeding or ischemia of the utero-placenta, over-distension caused by multiple pregnancies and immunological factors lead to neonatal death.

Mesike & Mojekwu (2012) conducted a study which concluded that that household environmental characteristics do have significant effect on mortality. A study by Folasade (2000) on the effect of environmental factors on women and infant mortality also showed that the main predictor of infant mortality is domestic environmental conditions. Studies have also shown that ecological factors impacts on infant survival especially in rural areas. Thus unfavourable climate conditions result in a sharp decline of food production thereby resulting in malnutrition. Another study by Santos & Henry (2008) on rainfall variations and infant survival showed that infant survival is dependent on rainfall situations, however the various rainfall patterns impacts on child mortality differently.

2.3.3 Demographic Characteristics

A study by Matthews et al. (2010) in Zambia revealed that demographic characteristics such as short birth interval and small size at birth have strong influence on neonatal mortality. Another study was conducted by Manda (1999) in Malawi on the relationship between child mortality and birth interval, maternal age at birth and birth order. He found that the effect of birth interval and maternal age neonates are usually experienced at the period of infancy. He also found that breastfeeding status does not have significant effect on neonatal mortality. Worku (2009) studied the key predictors of mortality among children under five using logistic regression and cox regression. He found that marital status, duration of breastfeeding and ownership of flush toilet have significant impact on child mortality. According to Pandey et al. (1998) child's sex, child's age, maternal age during childbirth, residence, maternal literacy, type of cooking fuel, accessibility of toilet facility, have effect infant mortality.

2.4 Review of Related Works

Infant and child survival has received lots of attention from researchers in recent times. Thus various research have been conducted on the risk factors of neonatal mortality both locally and international. This is particularly so because neonatal mortality rate is seen as one of the key health indicators of a country. Some of the related works conducted by various researchers both locally and internationally are reviewed below.

A study was conducted by Lawn et al. (2005) on why, when and where 4 million babies die each year. Their findings revealed the direct risk factors of neonatal death: preterm birth (28%), asphyxia (23%), severe infections (26%) and neonatal tetanus (7%). Their findings also revealed the indirect causes of neonatal mortality and they are low birth weight and maternal complications in labour. Their findings concluded that poverty is strongly related with neonatal mortality.

Mercer et al. (2006) studied the risk factors connected to neonatal mortality in Bangladesh. The study adopted a case-control design to collect data from mothers who gave birth in the year 2003. They employed crude and adjusted odds ratio in calculating the relative risk for neonatal death. Their findings showed that complications during delivery, prematurity and care for a sick new-born infant from unaccredited traditional healer are the main factors for neonate deaths among the singleton babies. Their study concluded that about 40.25 of the neonate deaths occurs within hours of delivery.

Ezeh et al. (2014) conducted a study on the factors of neonatal mortality in Nigeria: using 2008 NDHS data. The 2008 NDHS sampled 36,298 households and of these households 27147 singleton live-born survival information was captured with 996 neonatal death cases.

The study employed Cox regression in analysing the data. Their findings revealed that higher birth order of new born with both shorter and longer birth interval were significantly linked to neonatal death. They also found that infants born to mothers less than 20 years, neonates born to mothers living in villages, male neonates, small neonate's body size, birth through caesarean section affected neonatal deaths significantly.

Chukwu & Okonkwo (2015) studied the factors of under-five Mortality in Nigeria. The study compared the cox proportional hazard and Cox frailty models and employed the 2008 NDHS data in investigating the frailty among individuals. The Traditional Cox Proportional hazard model and Cox Frailty model were used in analysing the data. Their findings revealed the presence of frailty. Also mother's age and education were found to be significant determinants of under-five mortality.

Dahiru (2015) investigated the determinants of early neonatal mortality in Nigeria. The study used data from 2013 NDHS. The 2013 NDHS survey sampled 38,948 women aged 15-49 years and 17,359 men aged 15-59 years living in 38,904 households. The study adopted statistical model adopted the cox PH regression model in investigating the factors influencing early neonatal mortality. The study found that being a large baby, residing in rural area and mother experiencing pregnancy complication significantly increase the risk of early neonatal.

Li et al. (2015) examined the predictors of neonatal death in the villages of Shaanxi Province of Northwestern China. The study adopted a cross-sectional approach and used multivariate logistic regression analysis in analysing the data. The study sampled 4750 women who gave birth in the past three years and interviewed them. The result showed that multiparous mothers and mothers who did not visit antenatal in their first trimester of pregnancy had

higher odds of neonatal death. It also revealed that women who delivered in the hospital as well as those with some form of educational level were less likely to experience neonatal death.

A study was conducted by Yehuala, Ayalew, & Teka (2015) on survival analysis of premature infants in Northwest Ethiopia using Neonatal Intensive Care Unit (NICU) as a case study. A retrospective study approach was used in the collection of data from patients chart from December 29, 2011 to April 6, 2014. They compared the results of cox proportional hazard model and the semi-parametric gamma frailty model. The results from the study showed that a frailty effect was present and also the values of frailty were distributed inducing greater heterogeneity in the infant hazards. Both models also identified Prenatal Asphyxia, respiratory distress syndrome, antenatal care visit, HIV status of mother, anaemia, gravidity of (6-10) and breastfeed initiated as the main risk factors associated with death of neonates admitted to NICU. The study concluded that semi-parametric gamma frailty model best fit the data.

Nasejje, Mwambi, & Achia (2015) worked on under-five child mortality in Uganda using survival analysis methods. The study estimated the unobserved heterogeneity at the household and community level using the frequentist and Bayesian survival analysis. Uganda Demographic Health Survey (UDHS) data 2011 was employed for the study. The results from the study showed that there is a frailty effect at the household level whiles on the contrary no unobserved heterogeneity existed at the community level. It was also found that sex of the child, household head sex and parity in the past one year were main determinants of under-five mortality.

Niragire et. al. (2011) investigated the determinants of child mortality in Rwanda. The study employed 2000 and 2005 RDHS data and used the shared frailty model in analysing the data. The results from the study showed that frailty effects were significant in childhood. It was also found that child deaths were mostly determined by socioeconomic and demographic factors such as household socioeconomic status.

Lanfranchi, Viola, & Nascimento (2011) worked on the determinants of neonatal death in a private NICU using Cox regression. The study design was a longitudinal epidemiological study and the data for the study was gathered from folders of all new born admitted to a private NICU between January 2005 to December 2007. The study employed chi-square test, the Kaplan-Meier survival plot and the Cox regression analysis. The results from the study showed neonatal death is significantly associated with mechanical ventilation, birth weight, 5th minute Apgar score and previous stillbirth.

Mekonnen et al. (2013) worked on the trend and determinants of neonatal death in Ethiopia. The study made use of the 2000, 2005 and 2011 Ethiopia DHS and employed multivariate Cox proportional hazards regression model using a hierarchical approach to examine the related risk factors. The study showed that male sex, neonates born to mothers whose age are less than 18 years and those born within 2 years of the prior birth was associated with neonatal death. It was also revealed that neonates born during winter had a higher risk of death than neonates born during spring. Mothers who were injected with Tetanus Toxiod before childbirth had a lower neonatal mortality risk and neonates born to mothers who had secondary or higher education had lower risk of death than those with no education.

Afzal & Alam (2013) examined the causes of under-five mortality between rural and urban area in Bangladesh. The study used Bangladesh DHS 2007 data adopted the Kaplan-Meier, Cox PH and Accelerated Failure Time (AFT) Regression model in analysing the data. The results from the study revealed that for both rural and urban, the survival times for infants whose mothers have higher education is very high. Also it was revealed that children from the urban areas with poor economic status had a very high failure rate. The Cox-PH analysis showed that the hazard risk for infants whose mothers were mature and highly educated were lower than those with no education from the rural areas. From urban areas, infants from rich homes and infants who are either 2nd or 3rd born have lower risk of death compared to poor and first born. The Weibull distribution under the AFT models was found fit the data best.

Haghighi et al. (2013) studied the survival predictors of preterm babies in Iran from 2010-2011. The study was a hospital based study and it included all preterm (26-37 weeks) babies born alive in the hospital, during one year period. The study adopted the survival analysis approach in identifying the risk factors of neonatal mortality. The revealed that the risk factors among preterm babies were extremely low birth weight, low birth weight, Apgar score less than 7, multiple pregnancy, presence of anomalies, non-cephalic presentation, history of previous dead neonate, very early PTB, early PTB, LBW, need to cardio-pulmonary resuscitation (CPR), need for neonatal intensive care unit admission and postnatal administration of surfactant.

Other studies on infant and child mortality in Ghana include one by Kojo (2012) which focused on assessing and analysing the risk factors of neonatal deaths in Ghana. The study used the 2008 GDHS data gathered by the Ghana Statistical Service and the binary logistic regression was use for the presentation and analysis. The study developed three models which are child level factors, mother level factors and environmental level factors. The results from

the study showed that at the mother level factors, maternal age and the wealth index were the main determinants. Also at the child level factors, size of child, sex of child and whether the child was a twin or not were found not to be significant and for the environmental level factors only region (site of delivery) was significant.

Siakwa et al. (2014) also conducted a study on the trend of neonatal mortality in a Teaching Hospital in Ghana. The study retrieved data from the Okomfo Anokye Teaching Hospital from 2008 to 2012 and analysed the data using descriptive analysis. The results revealed that the neonatal mortality rates declined steadily from 32/1000 live births in 2008 to 14/1000 live birth in 2011, but increased drastically to 39/1000 live births in 2012. They also found that birth asphyxia, new-born jaundice, prematurity, neonatal sepsis, and respiratory distress were the main reasons for new-born death within the period of study.

Another study was conducted by Osei-Kwakye et al. (2010) on the risk factors of under-five mortality in the East Region of Ghana. The study employed a case-control study in collecting data from mothers. The study employed Bivariate conditional logistic regression in identifying the risk factors associated with under-five mortality. The findings from the study revealed that infants whose mothers have had previous child deaths were 8 times more probable to death. Also infants who were not given vitamin A supplement were about 10 times more probable to death. It was however found that educational level of mothers was not significant determinant of infant death. Finally, the study found the following protective risk factors: exclusive breastfeeding, use of an insecticide-treated bed net, immunization and number of live babies a mother had.

Kayode et al. (2014) studied the individual and community determinants of neonatal mortality in Ghana. The study used a combined data of 2003 and 2008 Ghana DHS and adopted hierarchical modelling for the data analysis. Thus the study applied a two-level multivariable multilevel logistic regression on the data. The result showed that both individual and community characteristics were linked to neonatal death. The main risk factors identified were infants of multiple-gestation, neonates with inadequate birth spacing, low birth weight, infants of grand multiparous mothers, non-breastfed infants and living in an area with high socioeconomic deprivation. It was however found that adequate utilization of antenatal, delivery and postnatal health services reduced the odds of neonatal mortality.

Engmann et al. (2012) worked on stillbirths and early neonatal mortality in rural Northern Ghana. The data used for the study was collected from the Navrongo Health and Demographic Surveillance System over a 7-year period. Thus 24,097 pregnant women were considered for the study. The study used logistics regression in analysing the data. The results showed that perinatal mortality rate was 39 deaths / 1000 deliveries and stillbirth rate was 23 / 1000 deliveries. Also it was revealed that factors such as first-time delivery, multiple gestation and prematurity increased the odds of perinatal death. Finally it revealed that birth asphyxia and injury, infections and prematurity were the most common causes of death.

CHAPTER THREE

METHODOLOGY

3.1 Introduction

This chapter deals with the methodology for the study. It present the area of study, types and source of data, statistical techniques employed, Data analysis and ethical consideration.

3.2 Study Area

The study was conducted in St. Jude Hospital located at Obuasi in the Ashanti Region. The hospital has a well-equipped neonatal intensive care unit and it provides many specialist care services. The study was restricted to only neonates born at the hospital and their mothers.

3.3 Type and Source of Data

The study relied solely on secondary data from the Biostatistics department of the hospital. In this study, the researcher reviewed folders of all neonates admitted to the hospital between January 1, 2012 and December 31, 2015. Neonates whose folders were missing important information were excluded from the sample. Information collected from these folders using a researcher's developed checklist were the sex of the child, birth weight, Apgar score, gestational age, mode of delivery, maternal age, parity, place of residence, complications of delivery, survival time and survival status.

3.4 Basic Concept of Survival Analysis

Survival analysis involves the analysis of time to event of interest data. Stevenson & EpiCentre (2009) defined survival analysis as a statistical technique used in describing and

quantifying time to an event data. The term survival analysis arose because the most common event of interest was death. Time variable in survival analysis is normally referred to as survival time because it specifies the time an individual 'survives' over a period. Time could be days, weeks, months or years of an individual until an event occurs. An event is an outcome on an individual unit that is of scientific interest. It is typically referred to as failure because they are commonly associated with death, disease incidence, recovery or negative individual experiences. According to Wintrebert et al. (2004) the event in all cases could be seen as a transition from one state to another. Time to event data includes time from patient admitted to being discharged from the hospital, time from marriage to divorce, time from disease diagnosis to death, time from release from prison to re-arrest and so on.

Survival data requires unique statistical methods of analysis because the data is always incomplete. Aalen, Borgan, & Gjessing (2008) noted that incomplete dataset arises because of censoring and truncation. Censoring is said to be present when we have some information about a subject's event time, but do not know the exact event time. Censoring can be classified into three main forms and these are: right censoring, left censoring and interval censoring. Right censoring is said to be present when an individual experience the event of interest after the given time t. There are three types of right censoring and they are type 1 and 2 and random censoring. The fixed type 1 censoring is a type of censoring where a censored subject does not experience the event of interest because the experiment is set to end after "C" years of follow-up. Hence any subject who does not experience the event before the end of the experiment and follow-up time elapses is censored. The fixed type 2 censoring is type of censoring where the experimental design is such that, there are a pre-specified number of events of interest. Hence any subject that does not experience the event before the required number is obtained is censored. Random censoring occurs when subjects have different

censoring time even though the experimental design has a fixed study time. Left censoring is said to be present when an individual experience the event of interest even before the start of the study. Interval censoring is where the only information is that the event occurs within some interval of time. The study employed the fixed type 1 in censoring the subjects. Survival analysis has two basic functions and they are the survival and hazard function. This will be discussed in the next subsection.

3.4.1 Survival Function

Survival Function is the probability of an individual surviving or been event-free beyond time t. It is denoted by $S(t)$ and is expressed mathematically as:

$$S(t) = \Pr(T > t) = \int_{t}^{\infty} f(t)dt = 1 - F(t) \qquad 3.1$$

Where $f(t)$ and $F(t)$ are the probability density and cumulative density function respectively of any given distribution. The survival function is a non-increasing function implying that as time increases, the survival chance of an individual decreases. Thus $S(0) = 1,\ S(t) \rightarrow 0\ as\ t \rightarrow \infty$. However it is possible that as $t \rightarrow \infty, S(\infty) > 0$. The survival function is normally useful when comparing the survival between two or more groups.

3.4.2 Hazard Function

The hazard function $h(t)$ is the instantaneous rate at which an event occurs given that the event has not occurred already. It is also defined as the probability of an individual in in the risk set experiencing the event in the small time interval $[t,\ t + \Delta t]$. Mathematically the hazard function is given as:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \qquad 3.2$$

The hazard function takes on any shape of a non-negative function and it varies depending on the type of survival data. For a continuous random variable, the hazard function is given as:

$$h(t) = \frac{f(t)}{S(t)} = \frac{-dIn[s(t)]}{dt} \tag{3.3}$$

The cumulative hazard function which is the sum of the hazards to a time t, is also given as:

$$H(t) = \int_0^t h(u)du = -In[S(t)] \tag{3.4}$$

Hence we can deduce the survival function from equation (3.4) thus from the cumulative hazard function as:

$$S(t) = exp[-H(t)] = exp\left[-\int_0^t h(u)du\right] \tag{3.5}$$

The hazard function is noted to provide a mored detailed information about the risk of failure at any time point.

3.4.3 Relationship between Survival function and hazard function

There is a clearly defined relationship between $S(t)$ and $h(t)$. From equation 3.2,

$$h(t) = H'(t) = \frac{1}{S(t)} \lim_{\Delta t \to 0} \frac{S(t) - S(t + \Delta t)}{\Delta t} = \frac{S'(t)}{S(t)}$$

$$H(t) = -\frac{S'(t)}{S(t)} = -In[S(t)] \tag{3.6}$$

where $H(t) = \int_0^t h(u)du$ is the cumulative hazard function and $H'(t)$ denotes the first derivative of the cumulative hazard function.

3.5 Estimation of Survival Data

The most common approaches used in estimating survival data are the parametric, semi parametric and non-parametric approach. The parametric model includes exponential, Weibull, Gamma, Lognormal and Log logistic distribution. Semi parametric model is an example of the Cox PH model while the Kaplan-Meier (KM) estimation is also an example of the non-parametric model. These approaches are discussed in the subsection below.

3.5.1 Non-parametric Model

Non-parametric methods summarize survival data through the estimation of the hazard and survival function (Kaplan & Meier 1958). The main objective of non-parametric estimation of the survival function is to produce a graphical summaries of the survival times for a given group of individuals considered in the study. The Kaplan-Meier (KM) also known as the Product Limit estimator is the standard nonparametric approach often used in estimating the survival function. The use of K-M estimator gives a simple and quick estimate of the survival function especially when right censoring is present. The K-M estimator of $S(t)$ is defined as:

$$\hat{S}(t) = \begin{cases} 1 & if\ t < t_i \\ \prod_{t_1 \le t} \left[1 - \frac{d_i}{Y_i}\right] & if\ t < t_i \end{cases} \quad 3.7$$

Where t_i denotes the first observed failure time, d_i represents the number of individuals who experience the event of interest at t and Y_i indicates the number of individuals who have not experienced the event of interest and have also not been censored by time t. It can be noticed from equation 3.6 that before the first failure occurs, the survival probability is always 1. The K-M estimator of the survival function is expected to decrease as the failures occur. After obtaining the survival functions using the K-M estimator, a graph of the survival function against time is plotted to enhance understanding of the function. A step function with jumps

at the observed event times is obtained after plotting the survival function. The jumps on the survival curve depend not only on the number of events observed at each event time, but also on the pattern of the censored observations before the event time. It is important to note that in the absence of censoring and truncation, the Kaplan Meier curve is equivalent to the empirical distribution function.

The survival curve of two groups can also be compared by plotting them on the same graph and then observe to see if there differences exist between the curves (whether one group has a higher chance of survival than the other group). The difficulty with this procedure is that there might be a difference between the two groups but may not be significant. According to Fleming & Harrington (2011) a more formal procedure like the Log rank test can be used in comparing the differences between two groups. Other test such as Wilcoxon test weights, Tarone-Ware weights and Peto-Prentice weights can also be used to compare survival among independent groups. This study used the log rank test to compare the survival among different covariates because of the ease in interpreting the results. The chi-square procedure is used in calculating for the log rank test statistic. The observed survival is compared with the expected survival at each time point. Mathematically the test statistic for the log rank test for a two category covariate is given as:

$$\text{Log rank statistic} = \frac{(O_i - E_i)^2}{Var(O_i - E_i)} \tag{3.8}$$

Where O is the observed survival count and E is the expected survival count

We approximate the test statistic in other to apply it in comparing two or more category covariates. The approximated test statistic is therefore given as:

$$\text{Log rank statistic} \approx \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{3.9}$$

The general hypothesis for the test is stated as:

H_o = There are no differences between the survival curves

H_1 = There are differences between the survival curves

3.5.2 Semi-parametric Model (Cox-Proportional Hazard Model)

Cox proportional hazard model is one of the most commonly used regression technique for survival outcomes. This model is called a semi-parametric model because it makes no assumption about the baseline hazard function but assumes a parametric form for the effect of the prognostic variables on the hazard. Hosmer, Lemeshow, and May (2008) posited that the Cox model is not classified as a full parametric model because though its survival time assumes a parametric regression structure, it reliance on time is left unspecified. The Cox model was proposed by Cox (1972) and it is given as:

$$h(t|X) = h_0(t)\exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right) = h_0(t)\exp(\beta' x) \qquad 3.10$$

where $h_0(t)$ represent the baseline hazard function, $x = (x_1, x_2, \ldots, x_p)'$ represent the values of the vector of explanatory variables for a particular individual and $\beta' = (\beta_1, \beta_2, \ldots \beta_p)$ represent the vector of regression coefficients.

The cox proportional hazard model assumes that the survival times between individuals in the sample are independent, multiplicative relationship between the predictors and the hazard and a constant hazard ratio over time. A unique feature about the cox model is that though the baseline hazard is unknown, the regression coefficient, hazard ratio and adjusted hazard curves can still be estimated. The hazard ratio (HR) measures the effect of the given predictors on the survival times and it is obtained without estimating the hazard function. The hazard ratio of two individuals with distinct covariates say x and x^* is given as:

$$HR = \frac{\hat{h}_0(t)exp\left(\sum_{i=1}^{k}\beta_i\, x_i^*\right)}{\hat{h}_0(t)exp\left(\sum_{i=1}^{k}\beta_i\, x_i\right)} = exp\left[\sum_{i=1}^{k}\beta_i(X_i^* - X_i)\right] \qquad 3.11$$

When the hazard ratio (HR)=1, it denote that individuals in both categories are at the same risk of attaining the event. Also when HR>1, it denote that individuals in the first group (x^*) are at a high risk of obtaining the event and finally when HR<1, it implies that the individuals in the second group (x) are at a high risk of getting the event.

In other to estimate the parameters in equation 3.10 $[h_0(t)$ and $\beta]$, Cox (1972) proposed the use of partial likelihood function which takes into account censored data. The partial likelihood function for the Cox model is given by:

$$L(\beta) = \prod_{j=i}^{r}\frac{\exp(\beta' x_i(t_{(j)})}{\sum_{k\in R(t_{(j)})}\exp(\beta' x_k(t_{(j)})} \qquad 3.12$$

Where $x_i(t_{(j)})$ denote the vector of covariate values for the i individual who experience the event at $t_{(j)}$ and $R(t_{(j)})$ is the risk set at timet_i. It could be noticed that only event times contribute their factor to the numerator but the denominator contains both the censored and uncensored observations where the sum over the risk set comprises of all individuals who are still at risk just before time t_i. Taking the log of the partial likelihood we obtain:

$$l(\beta) = In\big(L(\beta)\big) = \sum_{j=1}^{r}\left\{X_i\beta - In\left(\sum_{k\in R(t_{(j)}}\exp(X_k\beta)\right)\right\} \qquad 3.13$$

It is important to note that the partial likelihood is only valid when there are no tied event times in the dataset. Thus two subjects do not experience the event simultaneously. In the case of ties, approximations to the partial likelihood such as those proposed by Breslow (1974) and Efron (1977) will be needed.

3.5.3 Parametric Survival Model

Though the Cox proportional hazard model is the most common regression technique used in analysing survival data, it is however not suitable when the proportional hazard assumption is not met hence the need to introduce the parametric models. Unlike the Cox model, the parametric survival model is assumed to follow a known probability distribution for the survival times or outcome (Hosmer et al. 2008). Distributions such as the Weibull, exponential (a special case of the Weibull), log-logistic, Gompertz, log-normal and generalized gamma are mostly used for survival time. According to Collett (2015), under the parametric survival models time to event is assumed to follow a specific distribution whose probability density function $f(t)$ can be expressed in terms of unknown parameters. Once a distribution is specified for survival time, the corresponding survival and hazard functions can be determined. The simplicity and completeness of the parametric survival model makes statistical tests more powerful. Some of the distributions applied on survival times data are discussed briefly.

Exponential Distribution

The exponential distribution has one parameter (λ) and a constant hazard rate. This means that the probability of an individual dying within a small interval of time Δt given that the individual has survived up to time t is constant for any time period. The hazard function $h\,(t)$ and the cumulative hazard function $H(t)$ is given as:

$$h(t) = \frac{f(t)}{S(t)} = \lambda\,, \qquad H(t) = \lambda t$$

The cumulative distribution function of the exponential distribution is also given by:

$$F(t) = 1 - \exp(-\lambda t)\,, \quad 0 \leq t < \infty \qquad 3.14$$

From equation 3.14, we can derive the survival function $S\,(t)$ as follows:

$$S(t) = P(t) = 1 - F(t) = \exp(-\lambda t) \qquad 3.15$$

The probability density function of the exponential distribution is therefore given by:

$$f(t) = \lambda \exp(-\lambda t), \quad t > 0 \quad \lambda > 0 \qquad 3.16$$

The exponential distribution has an important property called the loss of memory property. This property means that the occurrence of a future event does not depend on an event that has already occurred. The exponential distribution is not applicable in some of lifetime data like biological and social process because of its memory loss property.

Weibull Distribution

The Weibull distribution is a more flexible and general distribution. It has two main parameters that is the scale parameter (λ) and the shape parameter (α). The shape parameter (α) shows the failure rate behaviour. When $\alpha = 1$, the failure rate remains constant which is a special case of the exponential distribution. If $\alpha < 1$, the failure rate decreases with time and if $\alpha > 1$, the failure rate increases with time. Rinne (2008) posited that a change in the parameter λ has the same effect on the distribution as a change of the abscissa scale. Thus an increase in the value of λ, holding α constant stretches out the pdf. If λ increases, the distribution stretches to the right with the height decreasing while maintaining its shape. On the other hand if λ decreases, the distribution is pushed towards left with the height increasing. The hazard rate $h(t)$ and the cumulative hazard rate $H(t)$ is given as:

$$h(t) = \lambda\alpha(\lambda t)^{\alpha-1}, \qquad H(t) = (\lambda t)^{\alpha}$$

The survival function $S(t)$ is also given as:

$$S(t) = e^{-(\lambda t)^{\alpha}}$$

The probability density function of the Weibull distribution is therefore given by:

$$f(t) = \alpha\lambda^{\alpha}t^{\alpha-1}e^{-(\lambda t)^{\alpha}} \qquad 0 < t < \infty \qquad 3.17$$

Gompertz Distribution

The Gompertz distribution has two parameters just like the Weibull Distribution and it is mostly employed in modelling mortality curves. A random variable follows a Gompertz distribution with parameters $\alpha > 0$ and > 0 $(T \sim Gompertz\ (\alpha, \beta))$. The hazard rate of the model is given as:

$$h(t) = \alpha e^{\beta t}$$

The survival function $S(t)$ is also given as:

$$S(t) = exp\left[\frac{\alpha}{\beta}(1 - e^{\beta t})\right]$$

The probability density function of the Gompertz distribution is therefore given by:

$$f(t) = \alpha e^{\beta t} exp\left[\frac{\alpha}{\beta}(1 - e^{\beta t})\right] \qquad \text{where } \alpha, \beta > 0 \text{ and } t \geq 0 \qquad 3.18$$

The hazard function increases from α at time zero to ∞ at time ∞. It must be noted that just by adding a constant to the hazard, the Gompertz generalizes to Gompertz-Makeham. This is given as:

$$h(t) = \alpha e^{\beta t} + c \qquad 3.19$$

Log-Normal Distribution

The Log-normal distribution is defined as the distribution of a variable whose logarithm follows the normal distribution. Hanagal (2011) noted that the lognormal distribution is a flexible distribution just like the Weibull distribution hence can fit many types of failure data. Consider the survival time T such that $log_e T$ is normally distributed with mean μ and variance σ^2. T is said to follow a lognormal distribution that is $\wedge(\mu, \sigma^2)$. The lognormal distribution is popular because it's cumulative values of $y = log_e t$ can be obtained from the standard normal distribution table and the corresponding values of t are obtained by taking

antilog. That is the percentiles of the lognormal distribution are easy find. The probability density function of the lognormal distribution is given by:

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} exp\left[-\frac{1}{2\sigma^2}(log_e t - \mu)^2\right] \qquad t > 0, \quad \sigma > 0 \qquad 3.20$$

and survival function is also given as:

$$S(t) = \frac{1}{\sigma\sqrt{2\pi}} \int_t^{\infty} \frac{1}{x} exp\left[-\frac{1}{2\sigma^2}(log_e x - \mu)^2\right] dx \qquad 3.21$$

Let $a = \exp(-\mu)$, then $-\mu = log_e a$. Equation 3.20 and 3.21 can therefore be rewritten as

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} exp\left[-\frac{1}{2\sigma^2}(log_e ax)^2\right] \qquad 3.22$$

$$S(t) = \frac{1}{\sigma\sqrt{2\pi}} \int_t^{\infty} exp\left[-\frac{1}{2\sigma^2}(log_e ax)^2\right] \frac{dx}{x}$$

$$= 1 - G\left(log_e \frac{ax}{\sigma}\right) \qquad 3.23$$

where $G(y)$ is the cumulative distribution function of the standard normal variable

$$G(y) = \frac{1}{\sqrt{2\pi}} \int_0^{y} e^{-\frac{\mu^2}{2}} du$$

The lognormal distribution is completely specified by two parameters scale (μ) and shape (σ^2) where $\mu = \exp\left(\mu + \frac{1}{2}\sigma^2\right)$ and $[\exp(\sigma^2) - 1]\exp(2\mu + \sigma^2)$ unlike the normal distribution whose parameters are location and scale. The two-parameter lognormal distribution can also be generalized to a three-parameter distribution by replacing t with $t - G$ in equation 3.20. Thus the survival time $log_e(T - G)$ follows the normal distribution with mean μ and variance σ^2. The lognormal distribution is noted to be positively skewed and thus

the greater value of σ^2, the greater the skewness. From equation 3.22 and 3.23, the hazard function can be derived as:

$$h(t) = \frac{\frac{1}{t\sigma\sqrt{2\pi}} exp\left[-\frac{(log_e at)^2}{2\sigma^2}\right]}{1 - G\left(log_e \frac{at}{\sigma}\right)} \qquad 3.24$$

Watson & Wells (1961) posited that the hazard function of the lognormal distribution increases from 0 initially to a reach maximum and then decreases monotonically, approaching 0 as $t \to \infty$.

Log-logistic Distribution

The log-logistic distribution is an alternative model to the Weibull distribution. It has a fairly flexible functional form. The log-logistic distribution is one of the parametric survival time models whose hazard rate may be decreasing, increasing, as well as hump-shaped, thus it initially increases and then decreases. The survival time T is said to have a log-logistic distribution if InT has a logistic distribution. The probability density function of the log-logistic distribution is given as:

$$f(t) = \frac{\alpha\lambda[\alpha t]^{\lambda-1}}{[1 + [\alpha t]^{\lambda}]^2}, \qquad \alpha > 0, \lambda > 0 \; and \; t > 0 \qquad 3.25$$

The survival function is also given by:

$$S(t) = \frac{1}{1 + [\alpha t]^{\lambda}} \qquad 3.26$$

The hazard function is therefore given as:

$$h(t) = \frac{\alpha\lambda[\alpha t]^{\lambda-1}}{1 + [\alpha t]^{\lambda}} \qquad 3.27$$

When $\lambda > 1$, the hazard function first increases and then reaches maximum at $t = \frac{[\lambda-1]^{\frac{1}{\lambda}}}{\alpha^{\frac{1}{\lambda}}}$ and it decreases. Also when $\lambda = 1$, the hazard begins at $\alpha^{\frac{1}{\lambda}}$ and then declines monotonically and when $\lambda < 1$, the hazard tends to be very large as $t \to 0$ and it then declines towards 0 as $t \to \infty$. The general shape of the log-logistic hazard function is very similar to the lognormal distribution. The log-logistic distribution is therefore useful in describing a hazard that increases initially and then later decreases.

Gamma Distribution

The gamma distribution includes the exponential distribution as a special case. The gamma distribution is mostly limited when applying it in survival analysis because it does not have a closed form expressions for survival and hazard functions. Both include the incomplete gamma integral given as:

$$I_k(x) = \frac{\int_0^x \lambda^{x-1} e^{-x} dx}{\Gamma(k)} \qquad 3.28$$

Also, estimating the parameters using the traditional maximum likelihood is difficult hence requires the calculation of such incomplete gamma integrals. This imposes additional numerical problems in parameter estimation. The gamma distribution with parameters λ and k, denoted $\Gamma(\lambda, k)$, has a probability density function:

$$f(t) = \frac{\lambda^k t^{k-1} e^{-\lambda t}}{\Gamma(k)} \qquad t > 0, \lambda > 0 \; and \; k > 0$$

where k and λ are the shape and scale parameters respectively.

The survival function is also given by:

$$S(t) = 1 - I_k(\lambda t)$$

The hazard function is therefore given as:

$$h(t) = \frac{\lambda^k t^{k-1} e^{-\lambda t}}{\left(1 - I_k(\lambda x)\right)\Gamma(\mathrm{k})} \qquad 3.29$$

When $k = 1$, the gamma distribution reduces to exponential distribution. Also when $\lambda = \frac{1}{2}$ and $k = \frac{1}{2}v$ where v is an integer, the gamma distribution reduces to chi square with v degrees of freedom. With integer k, the gamma distribution is sometimes called a special Erlangian distribution.

In this study, parametric distribution such as the exponential, weibull, Gompertz, lognormal and the log-logistic was considered. These parametric distributions can either be modelled using the Proportional Hazard (PH) model or the Accelerated Failure Time (AFT) model. The PH and the AFT model are discussed in the next sub-section.

3.5.3.1 Proportional Hazard (PH) Model

The proportional hazards model is a regression model with duration as dependent variable. PH model allows for information about known covariates to be included in the model of the survival analysis and is considered as the most used model in the area of survival analysis. Let $h(t, X)$ denote the hazard of an individual at time t with covariate vector $X = (X_1, \cdots, X_k)$. The proportional hazards model is given by:

$$h(t|x) = h_0(t)\exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right) = h_0(t)\exp(\beta' x) \qquad 3.30$$

Where $h_0(t)$ is the baseline hazard function which takes on any form, x is the set of observed values of the explanatory variables $\left(x = (x_1, x_2, \ldots\ x_p)\right)$ and $\beta_1, \beta_2, \ \ldots, \beta_p$ are the unknown parameters of the model. The baseline hazard function is the value of when all of the explanatory variables 0 and the parameters $\beta_1, \beta_2, \ \ldots, \beta_p$ in a proportional hazards model are called the proportional hazards regression coefficients. The proportional hazard models are

built in two main ways that is the baseline hazard function and the effect parameters. The baseline hazard shows how the hazard changes over time at the baseline (baseline is where all the covariates are zero). The effect parameters $\exp(\beta' x)$ however describes how the hazard changes in response to the explanatory covariates. The proportional hazards regression model can be simplified by taking the natural logarithm of the hazard function $h(t|x)$ resulting in linear form:

$$Inh(t|x) = Inh_0(t) + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad 3.31$$

Two main assumptions required in other to make proportional hazards model valid are:

1. Log-linear assumption: The natural logarithm of the hazard function is a linear function of the explanatory variables.
2. Proportionality assumption: The ratio of the hazard function for two observed values at time t depends only on the values of the explanatory variables and not on the time t.

It is important to note that when the all the assumptions are met, the explanatory variables have a multiplicative effect on the hazard function, which is the same for each time t. That is for $x_1 = (x_{11}, x_{21}, \ldots . x_{p1})$ and $x_2 = (x_{21}, x_{22}, \ldots . x_{p2})$, the ratio of the hazard function at x_1 and x_2 at time t under the proportional hazards model is given as:

$$\frac{h(t|x_1)}{h(t|x_2)} = \frac{h_0(t)e^{\beta_1 x_{11} + \beta_2 x_{21} + \cdots + \beta_p x_{p1}}}{h_0(t)e^{\beta_1 x_{12} + \beta_2 x_{22} + \cdots + \beta_p x_{p2}}}$$

$$= \frac{e^{\beta_1 x_{11} + \beta_2 x_{21} + \cdots + \beta_p x_{p1}}}{e^{\beta_1 x_{12} + \beta_2 x_{22} + \cdots + \beta_p x_{p2}}}$$

$$= e^{\beta_1 (x_{11} - x_{12}) + \beta_2 (x_{21} - x_{22}) + \cdots + \beta_p (x_{p1} - x_{p2})}$$

$$= e^{\beta_1 (x_{11} - x_{12})} \times e^{\beta_2 (x_{21} - x_{22})} \times \ldots . \times e^{\beta_2 (x_{21} - x_{22})} \qquad 3.32$$

It can be noticed from equation 3.32 that the ratio of the hazard function for two different sets of explanatory variables does not depend on the time t. This means that the ratio of the hazard function for the two different sets of explanatory variables will be the same for all possible survival times. Also it can be noticed that no assumptions are made about the underlying distribution of the baseline hazard function.

There are several methods of verifying the validity of the assumption of proportional hazards. One way of verifying is by comparing the log-log plots $\left(\text{plots of} - \log[-\log\hat{S}(t)]\right)$ versus the natural logarithm of time for each value of a covariate. The assumption of proportional hazard will be acceptable only if the plots are parallel. Kleinbaum & Klein (2005) noted that a plot of $-\log[-\log\hat{S}(t)]$ against $\ln(t)$ is likely to result in the following:

a. A Parallel straight lines plot indicating that Weibull PH and AFT assumptions are not violated
b. A Parallel straight lines plot with slope of 1 means Exponential PH and AFT assumptions hold
c. A Parallel but not straight line plot means PH model (but not Weibull and not AFT) assumption hold. Thus Cox model can be used.
d. A non-parallel and non-straight line plot suggest PH model is violated but not Weibull.
e. A non-parallel but straight lines plot suggest that Weibull holds, but PH and AFT is violated differently.

Cleves (2008) also described another method for assessing the assumption of proportional hazards graphically that is comparing the plots of survival curves $S(t|x_1)$ versus time that uses the Kaplan-Meier estimator $\hat{S}_{KM}(t|x_i)$ with those from the Cox regression $\hat{S}_0(t)^{\exp(\beta x_i)}$. The assumption of proportional hazards will be satisfied if these plots are very similar.

According to Cleves (2008) a statistical test like the Schoenfeld residual can also be used in checking the proportional hazards assumption. Suppose the PH assumption holds for a particular covariate then the Schoenfeld residual for that covariate will not be related to survival time. This test finds the Pearson product-moment correlation between the scaled Schoenfeld residuals and time. The null hypothesis state that the correlation between the Schoenfeld residuals and the ranked survival time is zero. When the null hypothesis is rejected, it is concluded that the PH assumption is violated. A nonzero slope is an indication of a violation of the proportional hazard assumption.

There are two main ways in dealing with non-proportionality for one or more covariates and they are the stratified Cox model and the Cox regression model with time-dependent variables. Considering a parametric model such as the AFT model is another method of dealing with non-proportionality. The Stratified Cox model stratifies all the predictors that do not satisfy the PH assumption. The data is stratified into subgroups and the model is applied on each stratum. The model is given by

$$h_{ig} = h_{og}(t)\exp(\beta' x_{ig}) \qquad 3.33$$

where g denotes the stratum

The stratified proportional hazards can easily be fitted using major software packages for survival analysis. A major limitation to this method is that the effect of the stratified predictor cannot be identified. The use of this method is therefore useful in situations where the covariate with non-proportionality is categorical and are not of direct interest.

The second approach to consider is the Cox model with time-dependent covariates. Time-dependent covariate are the values of covariates the change over time t. The PH assumption

is said to be violated if the effect of a variables changes with time. To model a time-dependent effect, a time-dependent covariate $X(t)$ must be created, then

$$\beta X(t) = \beta X \times g(t).$$

where $g(t)$ is a function t such as t or $logt$

The Cox model with time-dependent covariates $X_j(t)$ can be expressed as

$$h(t|x(t)) = h_0(t)exp\left[\sum_{i=1}^{p_1} \beta_i x_i + \sum_{i=1}^{p_2} \alpha_j x_j(t)\right]$$

The hazard ratio at time t for the two individuals with different covariates x and x^* is given by

$$\widehat{HR}(t) = exp\left[\sum_{i=1}^{p_1} \hat{\beta}_i(x_i^* - x_i) + \sum_{i=1}^{p_2} \hat{\alpha}_j\left(x_j^*(t) - x_j(t)\right)\right]$$

It can be noticed from the hazard function that, the coefficient $\hat{\alpha}_j$ is not time-dependent.

3.5.3.1.1 Parametric PH model

The parametric proportional hazard model is similar to that of the Cox model. It is considered as the parametric version of the Cox PH model. The main difference between the parametric PH model and Cox PH model is that for the parametric PH model the baseline hazard function assumes a specific distribution when fitted to a data whereas with the Cox model it has no such constraint. Also with the parametric PH model, the parameters are estimated using the maximum likelihood whiles the Cox model uses the partial likelihood in estimating its parameters. It is important to note however that the interpretation of the hazard ratio is the same for the two models and also both assumes the proportionality of hazards. The proportional hazard model is mostly applied to the following parametric distributions: exponential, Weibull and Gompertz distribution.

Weibull PH model

Supposing the survival times are assumed to follow a Weibull distribution with scale parameter λ and shape parameter α and this imposes a specific parametric form on $h_0(t)$. The hazard and survival function of the Weibull distribution are given as

$$h(t) = \lambda\alpha(\lambda t)^{\alpha-1} \ , \qquad S(t) = \exp(\lambda t^{\alpha})$$

Using equation 3.38, the hazard function of a neonate with covariates $(x_1, x_2, \dots, x_p)$ is given by:

$$h(t|x) = \lambda\alpha(t)^{\alpha-1}\exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right) = \lambda\alpha(t)^{\alpha-1}\exp(\beta' x) \quad 3.34$$

From equation 3.41, it can be observed that the survival time of the neonate in the study has a Weibull distribution with scale parameter $\lambda\exp(\beta' x)$ and shape parameter α. The Weibull distribution therefore has the proportional hazards property. This result shows that the effect of the explanatory variables in the model only change the scale parameter of the distribution whereas the shape parameter remains constant. The corresponding survival function is given by:

$$S_i(t) = exp\{-\exp(\beta' x_i)\lambda t^{\alpha}\} \quad 3.35$$

The Weibull PH model is fitted by building a likelihood function of the n observations and maximising this function with respect to the unknown parameters $\beta_1, \beta_2, \dots, \beta_p, \lambda \ and \ \alpha$. The likelihood function for the hazard and survivor function is written as:

$$\prod_{i=1}^{n}\{h_i(t_i)\}^{\delta_i} S_i(t_i) \quad 3.36$$

Taking logarithm of the likelihood function in equation 3.36, we obtain

$$\sum_{i=1}^{n}\{\delta_i log h_i(t_i) + log\, S_i(t_i)\} \quad 3.37$$

Substituting equation 3.36 and 3.37 into equation 3.35, the log-likelihood becomes

$$\sum_{i=1}^{n}[\delta_i\{\beta' x_i + \log(\lambda\alpha) + (\alpha-1)logt_i\} - \lambda\exp(\beta' x_i)t^{\alpha}]$$

which can be written as

$$\sum_{i=1}^{n}[\delta_i\{\beta' x_i + \log(\lambda\alpha) + \alpha logt_i\} - \lambda\exp(\beta' x_i)t^{\alpha}] - \sum_{i=1}^{n}\delta_i logt_i \qquad 3.38$$

Since the last term in equation 3.45 does not have any of the unknown parameters, it is omitted, hence the resulting log-likelihood function is given by

$$\sum_{i=1}^{n}[\delta_i\{\beta' x_i + \log(\lambda\alpha) + \alpha logt_i\} - \lambda\exp(\beta' x_i)t^{\alpha}] \qquad 3.39$$

Exponential PH model

The exponential is a special case of the Weibull model when α=1. The hazard function of the exponential model is assumed to be constant over time. The hazard and survival function are given as:

$$h(t) = \lambda, \quad S(t) = \exp(-\lambda t)$$

The hazard function for the exponential PH model of a particular neonate is given by

$$h(t|x) = \lambda\exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p) = \lambda\exp(\beta' x) \qquad 3.40$$

Gompertz PH model

The hazard and survivor function of the Gompertz distribution are given by

$$h(t) = \alpha e^{\lambda t}, \quad S(t) = exp\left[\frac{\alpha}{\lambda}(1 - e^{\lambda t})\right]$$

where $0 < t < \infty$ and $\alpha > 0$. The parameter λ determines the shape of the hazard function. Similar to the Weibull hazard function, the Gompertz hazard increases or decreases

monotonically. The general Gompertz proportional hazards model for the hazard function of a particular neonate is given by

$$h_i(t) = \exp\left(\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}\right) \alpha e^{\lambda t} = \alpha \exp(\beta' x) \exp(\lambda t) \qquad 3.41$$

where $x_{1i}, x_{2i}, \ldots x_{pi}$ are the values of p explanatory variables and $\beta's, \alpha \; and \; \lambda$ are the unknown parameters. It can easily be seen that the Gompertz distribution has the PH property. The Gompertz model can be fitted by maximizing the likelihood function provided by the expression in equation 3.41.

3.5.3.2 Accelerated Failure Time Model

The accelerated failure time model (AFT) is an alternative to the PH model for the analysis of survival time data. The AFT model covers a wide range of survival time distributions. With this model the explanatory variables has direct effect on the survival time instead of the hazard function as in PH model. This makes the interpretation of the results very easy, because the predictors affect the mean survival time through the regression parameters. Under the accelerated failure time model, the hazard function of the ith individual at time t is given by

$$h_i(t) = e^{-\eta_i} h_0(t|e^{\eta_i}) \qquad 3.42$$

where $\eta_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi}$ is the linear component of the model, x_{ji} is the value of the jth explanatory variable, X_j, $j = 1,2, \ldots, p$ for the ith individual, $i = 1,2, \ldots, n$. $h_0(t)$ is the baseline hazard function (the hazard rate at time t for an individual for whom the values of the p explanatory variables are all equal to zero). The corresponding survivor function for the ith individual is also given by

$$S_i(t) = \{S_0(t|\exp(\eta_i)\} \qquad 3.43$$

where $S_0(t)$ is the baseline survivor function and $\exp(\eta_i)$ is the acceleration factor.

The AFT model expressed in the log-linear form is given by

$$logT_i = \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi} + \sigma\varepsilon_i \qquad 3.44$$

where $\alpha_1, \alpha_2, \ldots, \alpha_p$ are the unknown coefficients of the values of p explanatory variables, $X_1, X_2, \ldots, X_p$ and μ, σ are the intercept and scale parameters respectively. The quantity ε_i is a random variable used to model the deviation of the values of $logT_i$ from the linear part of the model and is assumed to have a particular probability distribution.

The log-linear formation in equation 3.44 can be used to form a general survivor function for the ith individual, given as

$$S_i(t) = P(T_i \geq t) = P(logT_i \geq logt)$$

From equation 3.44,

$$S_i(t) = P\left(\mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi} + \sigma\varepsilon_i \geq logt\right)$$

$$= P\left(\varepsilon_i \geq \frac{logt - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma}\right) \qquad 3.45$$

Now the survival function $S_{\varepsilon_i}(\varepsilon)$ for the random variable ε_i can be expressed from equation 3.45 as

$$S_i(t) = S_{\varepsilon_i}\left(\frac{logt - \mu - \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi}}{\sigma}\right) \qquad 3.46$$

It can be noticed that the survivor function T_i for can be found from the survivor function of the distribution of ε_i. Also the result shows that an accelerated failure time model can be obtained from many probability distributions for ε_i.

The accelerated failure time model can be fitted using the maximum likelihood. The likelihood function can easily be derived from the log-linear representation in equation 3.44

from which the estimates will be obtained using the using iterative methods. The likelihood of the n observed survival times $t_1, t_2, \ldots, t_n$ is given as

$$L(\alpha, \mu, \sigma) = \prod_{i=1}^{n} \{f_i(t_i)\}^{\delta_i} \{S_i(t_i)\}^{1-\delta_i} \quad 3.47$$

Where $f_i(t_i)$ and $S_i(t_i)$ are the density and survival functions of the ith individual at t_i respectively and δ_i is the event indicator for the ith observation. Expressing the likelihood function in terms of the survivor and density function of ε_i yields

$$L(\alpha, \mu, \sigma) = \prod_{i=1}^{n} \{\sigma(t_i)\}^{-\delta_i} \{f_{\varepsilon_i}(z_i)\}^{\delta_i} \{S_{\varepsilon_i}(z_i)\}^{1-\delta_i} \quad 3.48$$

Taking the log of the likelihood function, we have

$$L(\alpha, \mu, \sigma) = \sum_{i=1}^{n} \{-\delta_i \log(\sigma t_i) + \delta_i log f_{\varepsilon_i}(z_i) + (1-\delta_i) log S_{\varepsilon_i}(z_i)\} \quad 3.49$$

Where $z_i = (log t_i - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \ldots - \alpha_p x_{pi})/\sigma$. The estimates of the unknown parameters, $\mu, \sigma, \alpha_1, \alpha_2, \ldots, \alpha_p$ can be obtained by maximizing the log-likelihood function using the Newton-Raphson procedure.

3.5.3.2.1 Parametric AFT model

The accelerated failure time model can be applied distributions such as the exponential, Weibull, log-logistic, log-normal and the gamma distribution. However, parametric accelerated failure time models based on the Weibull, log-logistic and lognormal distributions for the survival times are most commonly used in practice.

Weibull AFT model

If the survival times are assumed to have a Weibull distribution with scale parameter λ and shape parameter γ where the hazard function is

$$h_0(t) = \lambda\gamma t^{\gamma-1}$$

Under the AFT model, the hazard function for the ith individual from equation 3.49 is given as:

$$h_i(t) = e^{-\eta_i}\lambda\gamma(e^{-\eta_i}t)^{\gamma-1} = (e^{-\eta_i})^{\gamma}\lambda\gamma t^{\gamma-1} \qquad 3.50$$

The survival time for the ith individual therefore has a $W(\lambda e^{\gamma\eta_i}, \gamma)$ distribution which also implies that the Weibull distribution possess the accelerated failure time property. If T_i follows a Weibull distribution, then ε_i has an extreme value distribution known as Gumbel distribution and its survivor function is given by

$$S_{\varepsilon_i}(\varepsilon) = \exp(-e^{\varepsilon})$$

for $-\infty < \varepsilon < \infty$. The hazard and cumulative hazard for the Gumbel distribution is also given by $H_{\varepsilon_i}(\varepsilon) = e^{\varepsilon}$ and $h_{\varepsilon_i}(\varepsilon) = e^{\varepsilon}$ respectively.

From equation 3.53, the survival function of T_i under the AFT model is given as

$$S_i(t) = exp\left(-exp\frac{logt - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma}\right) \qquad 3.51$$

Which can also be expressed as

$$S_i(t) = \exp\left(-\lambda t^{1/\sigma}\right)$$

where $\lambda_i = exp\{-(\mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}/\sigma)\}$

The cumulative hazard function is given as

$$H_i(t) = -logS_i(t) = exp\left(-exp\frac{logt - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma}\right)$$

Which can also be written as $\lambda_i t^{1/\sigma}$ and the hazard function is given by

$$h_i(t) = \frac{1}{\sigma t} exp\left(-exp\frac{logt - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma}\right) \quad 3.52$$

$$= \lambda_i \sigma^{-1} t^{\sigma^{-1}-1}$$

From equation 3.42, the survival function of the Weibull model under the PH model is given by

$$S_i(t)\exp\{-exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)\lambda t^{\gamma}\} \quad 3.53$$

Comparing equation 3.51 and 3.52, it can be seen that the parameters λ, γ, β_j in the PH model can be expressed by the parameters μ, σ, α_j in the AFT model

$$\lambda = \exp(-\mu/\sigma), \quad \gamma = 1/\sigma, \qquad \beta_j = -\alpha_j/\sigma$$

Using equation 3.3, the AFT formation of the hazard function of the Weibull model is given by

$$h_i(t) = \frac{1}{\sigma t}\exp\left(\frac{logt - \mu - \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi}}{\sigma}\right) \quad 3.54$$

Assuming that the pth percentile of the survival time distribution for the ith individual is the value $t_i(p)$ such that $S_i\{t_i(p)\} = \frac{100-p}{100}$. From equation 3.56, we can obtain

$$t_i(p) = exp\left[\sigma log\left\{-log\left(\frac{100-p}{100}\right)\right\} + \mu + \alpha' x_i\right]$$

Similarly the pth percentile of the distribution of ε_i, $\varepsilon_i(p)$ is such that

$$exp\{-e^{\varepsilon_i(p)}\} = \frac{100-p}{100}$$

So that

$$\varepsilon_i(p) = log\left\{-log\left(\frac{100-p}{100}\right)\right\}$$

Log-logistic AFT model

Suppose the survival times have a log-logistic distribution with parameters α, λ, then the baseline hazard function is given by

$$h_0(t) = \frac{e^{\alpha}\lambda t^{\lambda-1}}{1 + e^{\alpha}t^{\lambda}}$$

Under the accelerated failure time model, the hazard of death at time t for the i[th] individual is

$$h_i(t) = e^{-\eta_i} h_0(e^{-\eta_i}t)$$

where $\eta_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi}$

consequently,

$$h_i(t) = \frac{e^{\eta_i} e^{\alpha}\lambda(e^{-\eta_i}t)^{-1}}{1 + e^{\alpha-\lambda\eta_i}t^{\lambda}}$$

$$= \frac{e^{\alpha-\lambda\eta_i}\lambda t^{\lambda-1}}{1 + e^{\alpha-\lambda\eta_i}t^{\lambda}} \qquad 3.55$$

It can be seen that the survival time for the ith individual follows a log-logistic distribution with parameters $\alpha - \lambda\eta_i$ and λ. The log-logistic distribution therefore has the accelerated time property. If T_i has a log-logistic distribution, then ε_i has a logistic distribution. The survival function of logistic distribution is given by

$$S_{\varepsilon_i}(\varepsilon) = \frac{1}{1 + \exp(\varepsilon)}$$

Using equation 3.53, the AFT formation of the survival function of the log-logistic model is given by

$$S_i(t) = \left[1 + t^{1/\sigma} exp\left(\frac{-\mu - \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi}}{\sigma}\right)\right]^{-1} \qquad 3.56$$

Using equation 3.32, the survivor function of T_i when T_i has a log-logistic distribution with parameters $\alpha - \lambda\eta_i$, λ where $\eta_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi}$ is given by

$$S_i(t) = \frac{1}{1 + e^{\alpha - \lambda \eta_i} t^{\lambda}} \qquad 3.57$$

Comparing equation 3.55 and 3.56, it can be seen that the parameters α and λ can be expressed in terms of μ and σ. That is

$$\alpha = -\mu/\sigma \text{ and } \lambda = \sigma^{-1}$$

This also indicate that the AFT model with log-logistic survival times can be represented in terms of a log-linear model. The hazard function for the ith individual is given by

$$h_i(t) = \frac{1}{\sigma t}\left\{1 + exp\left[-\frac{logt - \mu - \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi}}{\sigma}\right]\right\}^{-1} \qquad 3.58$$

The pth percentile of the survival distribution for the ith individual is $t_i(p)$ and from equation 3.58 $t_i(p)$ is given by

$$t_i(p) = exp\left[\sigma log\left(\frac{100 - p}{100}\right) + \mu + \alpha' x_i\right]$$

The log-normal AFT model

Assuming the survival times follows a lognormal distribution, then the baseline survivor function is given as

$$S_0(t) = 1 - \Phi\left(\frac{logt - \mu}{\sigma}\right)$$

where μ and σ are the unknown parameters. Under the AFT model, the survivor function for the ith individual is then

$$S_i(t) = S_0(e^{-\eta_i} t)$$

where $\eta_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi}$ is a linear combination of the values of p explanatory variables for the ith individual. Consequently,

$$S_i(t) = 1 - \Phi\left(\frac{logt - \eta_i - \mu}{\sigma}\right) \qquad 3.59$$

which is the survivor function for an individual whose survival times have a lognormal distribution with parameters $\mu + \eta_i$ and σ. The lognormal distribution therefore has the AFT property.

T_i has a lognormal distribution if $logT_i$ is normally distributed. It then follows that ε_i has a standard normal distribution. The survival function of ε_i is given by

$$S_{\varepsilon_i}(\varepsilon) = 1 - \Phi(\varepsilon)$$

The cumulative hazard and the hazard function of ε_i are given by

$H_{\varepsilon_i}(\varepsilon) = -log\{1 - \Phi(\varepsilon)\}$ and $h_{\varepsilon_i}(\varepsilon) = \frac{f_{\varepsilon_i}(\varepsilon)}{S_{\varepsilon_i}(\varepsilon)}$ respectively, where $f_{\varepsilon_i}(\varepsilon)$ is the density function of the standard normal distribution given as

$$f_{\varepsilon_i}(\varepsilon) = \frac{1}{\sqrt{(2\pi)}} exp\{-\varepsilon^2/2\}$$

The random variable T_i in the AFT model is then said to have a lognormal distribution with parameters $\mu + \alpha' x_i$ and σ. The hazard function is given as

$$h_i(t) = \frac{1}{\sigma t} h_{\varepsilon_i}\left(\frac{logt - \mu - \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi}}{\sigma}\right) \qquad 3.60$$

where h_{ε_i} is the hazard function of the distribution of ε_i. The survival function is also given in equation 3.64.

The p^{th} percentile of the distribution of T_i, is given as

$$t_i(p) = exp\{\sigma\Phi^{-1}(p/100) + \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi}\}$$

3.6 Frailty Models

Analysing survival data is mostly premised on the assumption that the study population is homogeneous. That is conditional on the covariates, all neonates have the same risk of experiencing an event (death). However, in practice we may have neonates with different risk

and hazards as well as some form of association between the event times of some subgroups of the population since neonates of these groups share a common characteristics which is unobservable. In the application of Survival analysis, usually only few covariates such sex, age, weight, parity etc. are considered ignoring other factors that can influence the survival of neonates. These factors ignored are mostly unknown hence cannot be included in the analysis.

Vaupel, Manton, & Stallard (1979) therefore introduced the concept of frailty in order to account for unobserved heterogeneity, random effects, and association in univariate survival models. The term frailty suggest that some neonates may be more frail or at risk than the others in the data set though they may appear to be similar considering their observable traits like age, weight, sex etc. Frailty model introduces an additional parameter to the hazard rate in order to account for the random frailties. Frailty model is a hazard model with a multiplicative frailty factor. The main assumption underlying this model is that information about the hidden factors (both internal and external) are included in the shape and structure of the hazard function which is in the form of frailty distribution. These frailties can be specific to individuals (individual frailty) or groups (shared frailty).

3.6.1 Model Development

Considering covariates with a covariate vector X, the Cox-proportional hazard model is given as

$$h(t, X) = h_0(t) \exp(X^T \beta)$$

where $X = [X_1, X_2, \ldots., X_n]$ are the covariate and $\beta = [\beta_1, \beta_2, \ldots., \beta_n]$ are the regression parameter vectors.

Also considering unobserved components denoted by a vector U, the Cox-proportional hazard model is modified as

$$h(t,X,U) = h_0(t)\,exp(X^T\beta + U) = h_0(t)\,exp(U)exp(X^T\beta) \quad 3.61$$

Let $Z = \exp(U)$ then we have:

$$h(t,X,Z) = h_0(t)\,exp(X^T\beta + U) = Zh_0(t)\,exp(X^T\beta) \quad 3.62$$

Equation 3.67 represent the frailty model where Z is a variable representing the frailty term.

3.6.2 Univariate Frailty Models

The univariate frailty model presents the population as a combination where the baseline hazard is common to all neonates but each neonate has their own frailty. Suppose we have a sample of j observations in a study, some of these observations fail earlier than others due to unobserved heterogeneity. Given that an individual neonate $i, i = 1,2,\ldots..n$ has a survival time denoted as t_i the hazard function conditional on both covariates and frailty can be written as

$$h_i(t_i,X_i,Z_i) = Z_i h_0(t_i)\,exp\left(X_i^T\beta\right) \quad 3.63$$

When $Z_i > 1$, it suggests that the ith individual is more frail than an average individual in that given cluster or group. On the contrary if $Z_i < 1$, it suggest that ith individual is less frail and therefore tends to survive longer.

The survival function of ith the individual conditional on frailty is given by

$$S_i(t_i,X|Z_i) = exp\left[-z_i\exp(X_i^T\beta)\int_0^{t_i} h_0(s,X_i|z_i)ds\right] = exp(-z_iH_0(t_i)\exp(X_i^T\beta)) \quad 3.64$$

where $H_0(t_i) = \int_0^{t_i} h_0(s)ds$ is the cumulative baseline hazard function at the individual level.

The unconditional survival function of an individual i at the population level is given as the mean of the survival function conditional on frailty with respect to the frailty distribution:

$$S_i(t_i,X_i) = E[S_i(t_i,X_i|Z_i)] = E\left[exp\left(-z_i exp\left(X_i^T\right)H_0(t_i)\right)\right]$$

3.6.3 Multivariate (Shared) Frailty Models

Multivariate frailty model is an extension of the univariate frailty model which allows individuals in the same cluster or group to share the same frailty value. Shared frailty model assumes that individuals in a subgroup or pair share the same frailty, but frailty from group to group may differ and hence the name shared frailty model. Shared frailty model was introduced by Clayton (1978) and was extensively studied by Hougaard (2012), Therneau, Grambsch & Fleming (1990), Duchateau et al. (2002) and Duchateau & Janssen (2007). Shared frailty model is similar to the individual frailty model except that the frailty is now shared among the n_i observations in the ith group. The hazard function of the jth individual of the ith group is given as:

$$h_{ij}(t) = Z_i h_0(t) \exp\left(X_{ij}^T \beta\right) \qquad 3.65$$

where X_{ij} is a vector of covariates for the individual j in the ith group, Z_i is the unobserved covariates and $h_0(t)$ is the baseline hazard function.

The frailties Z_i are assumed to be distributed to be identically and independently distributed random variables with common density function $f(u, \theta)$, where θ represent the frailty distribution parameter. A semi-parametric shared frailty model is a frailty model with a non-parametric baseline hazard function $h_0(t)$. It important to note that the variability Z_i of determines the degree of heterogeneity among the groups.

3.6.4 The Distributions of Frailty

The gamma distribution is the most widely used frailty distribution due to the fact that it is very tractable. It important to note that frailty can be assumed to follow other distributions such Weibull, lognormal, Inverse Gaussian etc. Hougaard (1986) suggested the gamma and the inverse Gaussian distributions are part of the positive stable family of distributions for the frailty model. Oakes (1989) also suggested the inverse Gaussian and log-normal models for

the distribution of the frailty. These distributions mentioned by the various authors have been discussed in the previous section. In this study the gamma distribution is used as the main frailty distributions. The gamma distribution is one of the most commonly used distributions for frailty. To Hougaard (1995) gamma frailty model is the most commonly used frailty model because of the following reasons:

- The Gamma distribution takes on positive random variables. Since the frailty term is a positive random variable, it makes the Gamma distribution the most suitable choice for the frailty term.
- The Gamma distribution is flexible. Thus the pdf of a two parameter Gamma distribution is given by:

$$f(z,\alpha,\beta) = \frac{\beta^{\alpha} z^{\alpha-1}\exp(-\beta z)}{\Gamma(\alpha)}, \quad \alpha > 0, \quad \beta > 0 \; and \; z > 0$$

The mean and variance are given as $E(Z) = \frac{\alpha}{\beta}$ and $V(Z) = \frac{\alpha}{\beta^2}$ respectively. Where α and β are the shape and scale parameters respectively. When the shape parameter is equal to 1 ($\alpha = 1$), the distribution becomes exponential with parameter β and for large values of α, the distribution assumes a bell shape that is identical to that of the a normal distribution.

- It is analytically tractable and easy to compute because of its simple Laplace transform.

$$L\{f(z)\}(s) = \int_0^{\infty} f(z)\exp(-zs)\,dz$$

$$= \int_0^{\infty} \frac{\beta^{\alpha} z^{\alpha-1} e^{-\alpha z}}{\Gamma(\alpha)} \exp(-zs)\,dz$$

$$= \frac{\beta^{\alpha}}{(s+\beta)}$$

where the parameter s is a complex number:

$$s = a + ib$$

3.6.5 The Shared Gamma frailty model

The gamma frailty model is restricted to one parameter Gamma distribution ($\alpha = \beta$), where the expectation $E(Z) = 1$ and variance $V(Z) = \frac{1}{\alpha}$. Assuming that the frailty term Z is distributed as gamma with $E(Z) = 1$ and $Var(Z) = \theta$, then $\alpha = \beta = \frac{1}{\theta}$. The probability density function of the frailty term Z is given by

$$f(z) = \frac{z^{1/\theta - 1} exp(-z/\theta)}{\Gamma\left(\frac{1}{\theta}\right)\theta^{\frac{1}{\theta}}}, \qquad \theta > 0$$

Let T denote the random variable representing the survival times and Z denote frailty which is distributed as a Gamma. The conditional survival function is given by:

$$S_i(t|z) = exp(-zH_0(t))$$

The unconditional survival function is obtained by integrating out z from the conditional survival function:

$$S_i(t) = E[S(t|z)]$$

$$= \int_0^{\infty} e^{-zH_0 exp(X_i^T \beta)(t)} f(z)dz = L\big(H(t)\big)$$

where L denotes the Laplace transform.

The Laplace transform of a one parameter Gamma distribution is given by:

$$L(S) = [1 + \theta s]^{\frac{-1}{\theta}}$$

The unconditional survival function can therefore be rewritten as:

$$S_i(t) = [1 + \theta H_0(t) exp(X_i^T \beta)]^{\frac{-1}{\theta}}$$

The likelihood function for the individual frailty is given by:

$$L(t, X_i, \beta, \theta) = \prod_{i=0}^{G} \prod_{j=1}^{n_i} z_i^{\delta_i} h_0(t_{ij})^{\delta_i} exp\left(\delta_i X_{ij}^T \beta\right), \left[S(t_{ij})\right]^{1-\delta_i} \quad 3.66$$

where G denotes the total number of clusters in the data set, and n_i is the total number of individuals in cluster i.

The full likelihood function of the shared frailty model is also given as

$$L\left(t_{ij}, X_{ij}, \beta, \theta\right) = \prod_{i=0}^{G} \prod_{j=1}^{n_i} z_i^{\delta_i} h_0(t_{ij})^{\delta_i} exp\left(\delta_i X_{ij}^T \beta\right), \left[S(t_{ij})\right]^{1-\delta_i} \prod_{i=1}^{G} f(z_i) \quad 3.67$$

3.6.6 Estimation of parameters in Shared Gamma frailty model

Given that the ith individual's survival and hazard functions are given as $S_i(t_i)$ and $h_i(t_i)$ respectively and that the probability density function of the frailty term is given by $f(z)$, then the likelihood function for that individual is denoted as $L_i(z_i, t_i, X_i, \theta)$ is given by:

$$L(z_i, t_i, X_i, \theta) = f(z)[S_i(t_i)h_i(t_i)]^{\delta_i}[S_i(t_i)]^{1-\delta_i}$$

Differentiating the likelihood function, we obtain

$$l = (z_i, t_i, X_i, \theta) = In\left(L_i(z_i, t_i, X_i, \theta)\right)$$

$$= In\left(f(z)[S_i(t_i)h_i(t_i)]^{\delta_i}[S_i(t_i)]^{1-\delta_i}\right)$$

Now we find the observed likelihood, since the frailty term is unobserved. The observed likelihood is found by integrating the frailty terms out with respect to its distribution

$$l_{obs}(z_i, t_i, X_i, \theta) = \int_0^{\infty} f(z_i) l_i(z_i, t_i, X_i, \theta) dz_i \quad 3.68$$

We now derive the parameter estimates by differentiating equation 3.69 with respect to all the parameters in the model. The resulting equations are then solved simultaneously. With frailty models it is very difficult to solve the equations simultaneously especially when latent

variables are present. The presence of this latent variables in addition to the unknown parameters requires that we use a more advanced method like the expectation-maximisation algorithm (EM-Algorithm), the Markov Chain Monte Carlo (MCMC), the Monte Carlo EM (MCEM) or the penalised partial likelihood (PPL). The Expectation-maximisation algorithm and penalised partial likelihood are discussed below.

3.6.6.1 The Expectation-Maximization (EM) Algorithm

Nguti (2003) posited that since the baseline hazard of the semi parametric model is unspecified and the frailties are unobserved, it becomes very difficult to maximize the likelihood to estimate the parameters The Expectation-Maximization (EM) algorithm which is typically used in the presence of unobserved (latent) information is one solution to this kind of problem**.** The EM algorithm iterates between the expectation and maximization step.

Expectation step

With the expectation step, the expected values of the unobserved frailties conditional on the observed information and the current parameter estimates are obtained.

Maximization step

With the maximization step, the expected values found in the E-step are taken to be the true information and new estimates of the parameters of interest are obtained by maximizing the likelihood, given the expected values.

EM algorithm application is based on two main conditions. That is the expected values for the unobserved information must be easy to find and secondly the maximizing the likelihood conditional on the expected values of the unobserved information must be straightforward. It is important to state that carrying out EM algorithm on the computer is intensive and slow.

3.6.6.2 The Penalized Partial Likelihood (PPL)

The Penalized Partial Likelihood (PPL) is an alternative estimation method proposed by Therneau et al. (1990). Under the PPL, the random effect is treated as a penalty term. The PPL is very fast in its application as compared to the EM algorithm and is mostly implemented in standard software. The PPL approach use the random effects u_i instead of the frailty term $z_i = \exp(u_i)$.

Assuming a univariate frailty model, the hazard function and the survival function of the individual is given by:

$$h_i(t_i) = h_0(t_i)exp\left(X_i^T\beta + u_i\right)$$

$$S_i(t_i) = exp\left(-H_0(t_i)\, expX_i^T\beta + u_i\right)$$

The full likelihood is given as:

$$L_{ifull}(u_i, \theta, \beta) = f_i(u_i)[h_i(t_i)S_i(t_i)]^{\delta_i}[S_i(t_i)]^{1-\delta}$$

where θ is the variance of the $u_i s$

The log of the full likelihood is given as:

$$L_{ifull}(u_i, \theta, \beta) = Inf_i(u_i) + h_0(t_i) + \delta_i X_i^T\beta + \delta_i u_i - H_0(t_i)expX_i^T\beta + u_i \qquad 3.69$$

According to Duchateau and Janssen (2007), equation 3.69 can be written in two parts, thus the first consisting of the conditional likelihood of the data given the frailties and the second part corresponding to the distribution of the frailties.

$$L_{ifull}(u_i, \theta, \beta) = l_{part1_i}(\beta, u_i) + l_{part2_i}(\theta, u_i) \qquad 3.70$$

From equation 3.70, $l_{part2_i}(\theta, u_i)$ is known as the penalty term, hence from the full likelihood in equation 3.69,

$$l_{part1_i}(\beta, u_i) = \delta_i Inh_0(t_i) + X_i^T\beta + u_i - H_0(t_i)exp(X_i^T\beta + u_i)$$

$$l_{part2_i}(\theta, u_i) = Inf(u_i)$$

When the actual value of the random effect is far away from its mean, the logarithm of the absolute value of the probability density function at that value of the random effect ($Inf|u_i|$) takes on a large value which this denotes that the penalty term has a large negative contribution to the likelihood. The penalised partial likelihood is given as:

$$l_{ppl}(u_i, \theta, \beta) = l_{part}(\beta, u_i) - l_{pen}(\theta, u_i)$$

$$l_{peni}(\theta) = -Inf(u_i)$$

$$l_{pen}(\theta) = \sum_{i=1}^{n} -In\, f(u_i)$$

If $\eta_i = X_i^T + u_i$

$$l_{part}(\beta, u_i) = \sum_{i}^{n} \delta_i \left[\eta_i - In\left(\sum_{qw \epsilon R(t_i)} \exp(\eta_i qw)\right)\right]$$

where $R(t_i)$ represents the risk set at time t_i. That is with all the contributing terms defined, l_{ppl} gives the penalised partial likelihood used to draw inference on the parameters θ and β.

3.7 Model Checking

Graphical methods can be used to check if the parametric model fits the observed data. For example, if the survival time follows an exponential distribution, then we can plot $log[-logS(t)]$ versus $log\ t$ and this should yield a straight line with slope of 1. If the plots are parallel but not straight, it suggests that the PH assumption holds but not the Weibull. The Weibull assumption is said to be valid if the lines for two groups are straight but not parallel. The log-logistic assumption can also be evaluated graphically by plotting the $log[(1 -$

$S(t)/S(t))]$ versus $\log t$. The log-logistic assumption is said to be valid if the resulting plot is a straight line. For the log-normal distribution, a plot of $\Phi^{-1}[1 - S(t)]$ versus $\log t$ should be linear. It must be noted however that graphical methods are not very reliable in practice because the conclusion is based on the researcher's observation.

Residual plots can also be used to check the goodness of fit of the model. The Cox-Snell residuals is one of the most useful plots for comparing the distributions. The Cox-Snell residual for the ith individual with observed time t_i is defined as

$$r_{c_i} = \hat{H}(t_i|x_i) = -log\left[\hat{S}(t_i|x_i)\right]$$

where t_i is the observed survival time for individual i, x_i is the vector of covariate values for individual i, and $\hat{S}(t_i)$ is the estimated survival function on the fitted model. The estimated survival function of the ith individual is given by

$$\hat{S}_i(t) = S_{\varepsilon_i}\left(\frac{\log t - \hat{\mu} - \hat{\alpha}x_i}{\hat{\sigma}}\right)$$

where μ, $\hat{\alpha}$ and $\hat{\sigma}$ are the maximum likelihood estimator of μ, α and σ respectively, $S_{\varepsilon_i}(\varepsilon)$ is the survival function of ε_i in the AFT model, and $r_{s_i} = \frac{\log t - \hat{\mu} - \hat{\alpha}x_i}{\hat{\sigma}}$ is known as the standard residual. The Cox-Snell residual can be applied to any parametric model. Under the Weibull AFT model since $S_{\varepsilon_i}(\varepsilon) = \exp(-e^{\varepsilon})$, the Cox-Snell residual for Weibull is then given as

$$r_{c_i} = -log\{\hat{S}(t_i)\} = -logS_{\varepsilon_i}(r_{s_i}) = \exp(r_{s_i})$$

Similarly, with the log-logistic AFT model, since $S_{\varepsilon_i}(\varepsilon) = (1 + e^{\varepsilon})^{-1}$, the Cox-Snell residual for the log-logistic is then given as

$$r_{c_i} = log\left[1 + exp(r_{s_i})\right]$$

Also under the lognormal AFT model, $S_{\varepsilon_i}(\varepsilon) = 1 - \Phi(\varepsilon)$, hence the Cox-Snell residual for the lognormal becomes

$$r_{c_i} = log[1 - \Phi(r_{s_i})]$$

If the fitted model is appropriate, the plot of $\log\left(-logS(r_{c_i})\right)$ versus $\log r_{c_i}$ is a straight line with the unit slope through the origin.

The quantile-quantile plot can be used to assess the potential of an accelerated failure time model. for any value of p in the interval (0,100), the pth percentile is

$$t(p) = S^{-1}\left(\frac{100 - p}{100}\right)$$

Let $t_0(p)$ and $t_1(p)$ be the pth percentile estimated from the survival functions of the two groups of survival data. The percentiles for the two groups can be formulated as

$$t_0(p) = S_0^{-1}\left(\frac{100 - p}{100}\right), \quad t_1(p) = S_1^{-1}\left(\frac{100 - p}{100}\right)$$

where $S_0(t)$ and $S_1(t)$ are the survival functions for the two groups. We therefore obtain

$$S_1[t_1(p)] = S_0[t_0(p)]$$

From the AFT model, $S_1(t) = S_0(t|\eta)$, hence

$$S_1[t_1(p)] = S_0[t_1(p)/\eta]$$

We therefore obtain

$$t_0(p) = \eta^{-1} t_1(p)$$

The percentiles of the survival distributions for two groups can be estimated by the K-M estimates of the individual survival functions. We then plot the percentiles of the K-M estimated survival function from one group against the other. The AFT model is said to be appropriate if the plot gives an appropriate straight line through the origin. The slope of the line will be an estimate of the acceleration factor η^{-1}.

The chi-square test or large sample Z tests can be used to test proportional hazard models. This test is used to check if the hazard ratio of two observations is constant. Thus each of the covariate is tested individually for a p-value say greater than 0.10. The PH assumption is said to be valid if the p-values are greater than 0.10. On the other hand if the p-value obtained is less than 0.05, it suggests that the covariate being test does not satisfy PH assumption.

3.8 Model Selection

For the purposes of comparing the parametric and semi-parametric models, the study adopted the Akaike Information Criterion (AIC) and it is given by

$$AIC = -2 * \log(likelihood) + 2(p + k)$$

Where p is the number of parameter, $k = 1$ for the exponential model and $k = 2$ for the Weibull, logistic and lognormal models. A smaller AIC values indicate the best likelihood or fitting model. The AIC values of the parametric models for the PH model are compared separately and then compared to the semi-parametric model. The AFT models are also compared and the best model is selected based on the AIC values.

3.9 Data Analysis

The study made use of statistical software's such as the STATA and R package. The data obtained was first entered and collated in the Excel spreadsheet. The statistical software's were then used for the data analysis. The R package was used to perform the Kaplan Meier and log-rank test whiles the STATA package was used for the PH, AFT and frailty modelling.

3.10 Ethical Consideration

This study had approval from the University of Ghana. Also, consent was sought from the management of St. Jude Hospital before data was obtained for the study. The researcher assured the management of the hospital of respondent's privacy, confidentiality and anonymity in the study. Thus any variable that would be connected to the identity of the neonates and their mothers were deleted from the database that was available for this analysis.

CHAPTER FOUR

RESULT AND DISCUSSION

4.1 Introduction

This chapter presents the analysis and the discussion of the results obtained from the study. The chapter is organised into two main sections excluding the introductory section. The first section presents the description of the variables and provide a general description of the data using basic statistics and Kaplan-Meier approach. The second section present the models used in the study along with their evaluation and diagnostics.

4.2 Description of variables

The response variable of interest for this study was the time until death of a neonate in days and the covariates were birth weight of child, Apgar score 1 and 5, gestational age, sex of child, maternal age, parity, mode of delivery, complications during delivery and place of residence. For easy exploration of the data some of the continuous variables such as birth weight, Apgar score, gestational age and parity were categorized into groups, however these continuous variables was used in fitting the models. The covariates with their description and categorization are presented in Table 4.1.

Table 4.1: Description and categorization of variables

Variables	Description	Categories
Birth weight	The birth weight of neonate at birth	Very Low (<1.5 kg); Low (1.5kg – 2.499kg); Normal (> 2.5kg)
Apgar score 1 and 5	Measurement of the physical condition of the neonate in the 1st and 5th minutes after birth	High risk (≤ 6) Adequate (≥ 7)
Gestational age	Duration of pregnancy in weeks	Preterm (< 37 weeks) Full term (≥ 37 or < 42)
Sex of child	Sex of the child	Male =1 and Female = 2
Maternal age	Age of the mother at time of birth	High risk (< 20 years or >34 years); Adequate (20-24 years)
Parity	Number of live offspring	High risk (none or more than Two); Adequate (one or two)
Mode of delivery	Delivery either by Caesarean section or SVD	SVD=1; CS=2
Complications	Complications of delivery	Yes = 1; No = 0
Place of residence	Place of residence during birth	1=Urban; Rural=2

The survival status of neonates was coded as

$$Status = \begin{cases} 1 & Death\ of\ neonate \\ 0 & Neonate\ is\ censored \end{cases}$$

4.3 Preliminary Analysis

The data used was a secondary data with 632 neonates considered for the study. The data spans a period ranging from 2012 to 2015. Table 4.2 presents the descriptive statistics of the variables of interest. Table 4.2 revealed that out of the 632 neonates sampled for the study, 320 of them were males and 312 were females. It was noticed that more of the females experienced the event than the males (Male dead =33, Female =40) from the sample. The Pearson chi-square obtained for sex of child indicates that there are no significant difference between the survival status of male and female ($\chi^2 = 0.974, p - value = 0.324$). Considering the mode of delivery, 484 of the mothers went through the spontaneous vaginal delivery (SVD) that is giving birth without medical intervention while 148 of the mothers undergo caesarean section delivery (CS). Out of the 484 mothers who gave birth through SVD, 64 of their neonates experienced the event and also out the 148 mothers who underwent CS 11 of their neonates died. This seems to suggest that neonates born through CS are more likely to survive than those born through SVD. The Pearson Chi-square however revealed that there are no significant differences between the survival status of neonates born through SVD or CS ($\chi^2 = 3.208, p - value = 0.073$).

The average weight of the neonates recorded was 3.15 with the minimum and maximum been 0.8 and 5.3 respectively. In all 3 of the neonates weighed below 1.5Kg signifying a very low weight, 36 weighed between 1.5-2.5Kg indicating a low weight and 590 of them weighed above 2.5Kg signifying a normal weight. One of the neonates who experienced the event had a weight less than 1.5Kg, 11 of the neonates who experienced the event had a weight between 1.5-2.5Kg and 60 of the neonates of died had a weight above 2.5Kg. The Pearson chi-square revealed that there are significant differences between the weight groups ($\chi^2 = 15.336, p - value = 0.000$).

The data showed that 571 of the mothers did not experience complications during delivery while 61 of them had complications during delivery. Again, it was noticed that with 8 of the neonates who experienced the event, their mothers had some form of complications during delivery while with 65 of the neonates who died, their mothers did not experience any complications. The Pearson chi-square revealed that there are no significant difference between neonates whose mothers had complications during delivery and those who did not have complications ($\chi^2 = 0.162, p - value = 0.688$). Table 4.2 also revealed that 208 of the mothers were from the urban centres of Obuasi while 571 of them were from the rural parts of Obuasi. Out of the 208 mothers from the urban centres, 13 of their neonates experienced the event and out of the 571 mothers from the rural parts, 60 of their neonates experienced the event. The Pearson chi-square established a significant difference among place of residence group ($\chi^2 = 8.526, p - value = 0.004$).

Considering maternal age, the data indicated that 124 of the mothers were below the ages of 20 years or above the ages of 34 years (indicating a high risk of giving birth) while 72 of the mothers were between the ages of 20-24 (signifying an adequate age of giving birth). It can nbe seen that with 15 of the neonates who experienced the event, their mothers ages were less than 20 years or above 34 years while with 57 of the neonates who died, their mothers ages were between 20-24 years. The Pearson chi-square suggested that there is no significant difference among the maternal age group ($\chi^2 = 0.072, p - value = 0.789$). The gestational age of the mothers were also considered it was revealed that the gestational age of 89 mothers were below 37 weeks (pre-term) while the gestational age of 543 mothers were either below 42 weeks or above 37 weeks (full-term).

The average gestational age among the mothers was 38 weeks with the minimum and the maximum been 20 and 47 weeks respectively. It was observed that 8 of the neonates who experienced the event were born before the 37 weeks of pregnancy was completed (pre-term) while 65 of the neonates who died were born after the pregnancy lasts for the normal length of time (full term). The Pearson chi-square did not show any significant difference among the gestational age group ($\chi^2 = 0.665, p - value = 0.415$). From the variable Apgar score (1st and 5th minute), 6 of the neonate who experienced the event had an Apgar score less than 6 meaning their physical condition was in a high risk while 54 of them who died had an Apgar score greater than 7 meaning their physical condition was adequate. The Pearson chi-square showed a significant difference among the Apgar score group ($\chi^2 = 37.72, p - value = 0.000$). Finally the average live offspring a neonate mother had was 1 with the minimum and maximum been 0 and 8 respectively.

Table 4.2: Descriptive Statistics

Variable	Dead	Censored	Min	Max	Mean	Std. Deviation	χ^2	P-value
Sex of child			1	2	1.4937	0.5003	0.973	0.324
Male	33	287						
Female	40	272						
Mode of delivery			1	2	1.234	0.4238	3.208	0.073
SVD	62	422						
CS	11	137						
Weight			0.8	5.3	3.1542	0.5259	15.336	0.000
<1.5Kg	1	2						
1.5-2.5 Kg	11	25						
>2.5Kg	60	530						
Complications			0	1	0.9652	0.2955	0.162	0.688
Yes	8	53						
No	65	506						
Place of residence			1	2	1.6709	0.4702	8.526	0.004
Urban	13	195						
Rural	60	364						
Maternal age			13	45	27.39	5.9496	0.072	0.789
< 20 years or > 34 years	15	109						
Between 20-24 years	57	15						
Gestational age			20	47	38.137	1.6516	0.665	0.415
<37 weeks	8	81						
≥37 weeks or < 42 weeks	65	478						
Apgar 1 and 5			1	2	1.8035	0.3977	37.717	0.000
≤ 6	6	28						
≥ 7	54	531						
Parity			0	8	1.4388	0.4238		

The Kaplan-Meier approach was also used to provide a summary of the distribution of the censored variables. The mean and median of overall survival time were 25.84 and 28.0 days respectively with a 95 percent confidence limit of $[25.3276, 26.3591]$ days. Also the minimum and maximum survival times of the neonate were 0 and 29 days. The result suggest that on the average neonates experience the event (death) after the 26^{th} day. Also neonates experience the event as early as day zero meaning still birth while some also experience the event as late as 29^{th} day. Figure 4.1 depict the Kaplan-Meier probability of the survival of the neonates with a 95 percent confidence band. The Kaplan-Meier uses the actual survival times and the probability of survival over time T. The K-M plot revealed that about 2 percent of the neonates experienced the event within day zero. It was also observed that more of the neonates were censored throughout the time period of study.

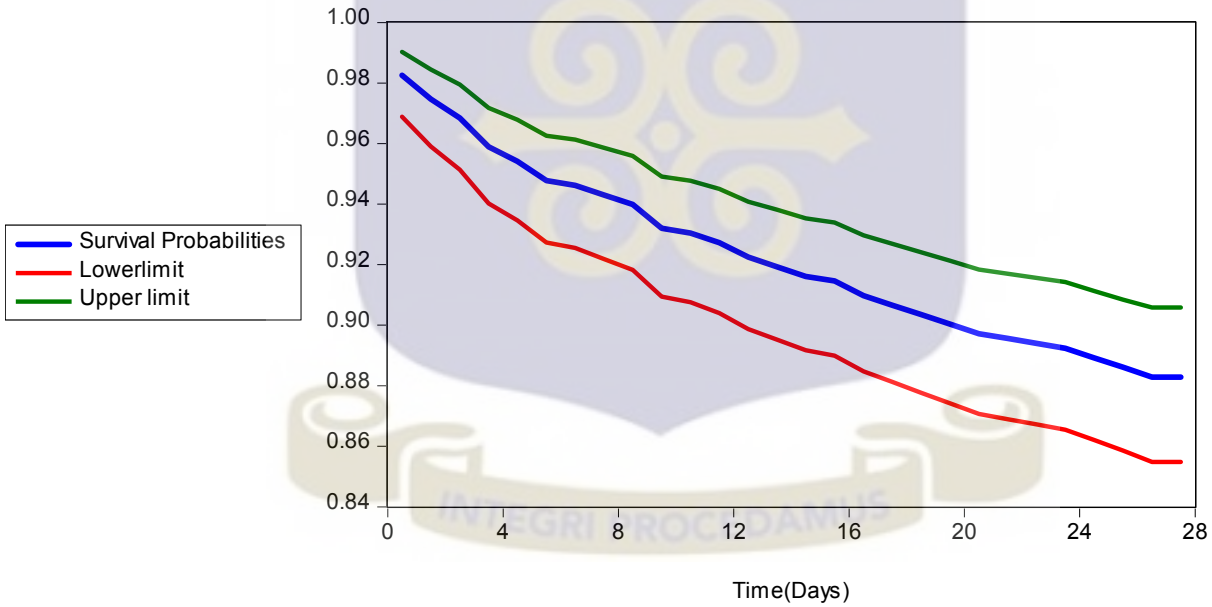

Figure 4.1: Kaplan Meier Survival Plot of time to neonate death in days

Plots of the Kaplan Meier curves for each category of the covariates sex of child, mode of delivery, birth weight of infants, complications, place of residence, complications of delivery and Apgar score are shown in Figure 4.2 to 4.7 respectively (See Appendix A). It can be noticed that the Kaplan Meier curve for females neonates tend to have shorter survival than

males. A careful look at the graph suggest that the difference between the two curves is very small. Also it can be noticed from Figure 4.3 that neonates born by SVD have lower chance of surviving as compared to those born by CS. This result is contrary to literature; however in recent times caesarean section has proven to be safe due to the advancement in technology. Neonates from the rural areas have shorter survival time than those from the urban areas. Considering the birth weight, the graph in Figure 4.4 shows that neonates with very low birth weight have lower chances of survival compared to those with low and normal weight. Furthermore neonates with adequate Apgar score (greater than or equal to 7) have longer survival compared to neonates with high risk Apgar score (less than 6). Finally, neonates born with complications have longer survival than those born without complications during delivery. However the curve crossed each other, suggesting no survival difference among the categories. Generally, it could be observed that graphically most of the KM curves indicate a difference between the covariate categories except the graph of complications.

A formal test was carried using the Log rank and Wilcoxon test to compare difference between each categorical variable. The general hypothesis states that there is no difference against there are differences among the groups. Thus we wish to test that:

H_0 : The survival times of neonates among the groups are not different

H_1: The survival times of neonates among the groups are different

From Table 4.3, the Log rank and Wilcoxon test conducted for mode of delivery showed no significant difference between the various survival experiences among the categories since we failed to reject the null hypothesis. Similarly, the Log rank and Wilcoxon test performed for sex of child, complications of delivery, maternal age, gestational age and Apgar score revealed a significant difference between the various survival experience among the categories. However the Log rank and Wilcoxon test carried out on place of residence, Apgar

score and birth weight indicated that there are significant difference between the various survival experience among the categories since we rejected the null hypothesis stating that the survival times of neonates are different among groups. Though the KM curve portrayed differences in the groups of covariates such as mode of delivery, sex of child complications of delivery, maternal age and gestational age, the statistical test performed using the Log rank and Wilcoxon test indicated that the differences are not significant.

Table 4.3: Log rank and Wilcoxon Test (Comparing categorical variables of neonates)

Variables	Log Rank		Wilcoxon		
	χ^2	P-value	χ^2	P-value	Decision
Mode of Delivery	2.30	0.1297	2.48	0.1153	Fail to reject H_0
Sex of child	2.09	0.1483	2.20	0.1378	Fail to reject H_0
Complications of delivery	0.22	0.6361	0.23	0.6280	Fail to reject H_0
Place of residence	7.19	0.0073	7.04	0.0008	Reject H_0
Maternal age	0.07	0.7885	0.04	0.8408	Fail to reject H_0
Gestational Age	0.15	0.7027	0.18	0.6678	Fail to reject H_0
Apgar score	13.08	0.0003	13.06	0.0003	Reject H_0
Birth Weight	15.04	0.0005	15.28	0.0005	Reject H_0

4.4 Further Analysis

This section presents the comparison of the PH models in the parametric form to the semi-parametric model, the cox regression model. Also comparison of the AFT model using distributions such as the exponential, Weibull, log-logistic and the log-normal was also presented. Comparisons of the best models from the PH model and AFT model was also presented. The Akaike information criterion (AIC) was used for comparing the efficiency of parametric and semi-parametric models. Finally, the semi-parametric frailty model was presented to check if there are unobserved heterogeneity among the neonates. The preliminary analysis revealed that variables such weight of child, place of residence, Apgar score 1 and 5 are likely to be the risk factors associated to neonatal death. It is important to mention however that all the variables explored were used in the modelling since literature consider them to be significant variables.

4.4.1 Checking the PH Assumptions

Before comparing the PH models in the parametric form to the semi-parametric model, we need to ensure that the PH assumptions are not violated. The proportional hazards models assume that the hazard ratio is proportional over time. The PH assumption was conducted for each of the covariates. The PH assumption test based on the Schoenfeld residual was performed to evaluate the PH assumption. This test checks the assumption on each of the covariates and also give a global test. From Table 4.4, it can be observed that all the p-values for the covariates and the global test are all greater than 0.05. This means that we have enough evidence to conclude that proportionality assumption has not been violated. Additionally, a plot of the Kaplan- Meier observed survival curves against the Cox predicted curves for the same variable was used to assess the PH assumption. When the predicted and observed curves are close together, it means that the PH assumption has not been violated.

Figure 4.8 to 4.15 presents the plots of Kaplan-Meier observed survival curve against the Cox predicted curves (see Appendix A). A careful look at the figures showed that observed Kaplan-Meir survival curve were very close to the cox predicted curves for place of residence, sex of child and Apgar score 1 and 5. Other covariates like the birth weight, maternal age, complications of delivery, gestational age and mode of delivery also saw their observed curves and predicted to be fairly close together. This result suggests that the PH assumption is less likely to be violated.

Table 4.4: Proportional hazard assumption test for the covariates

	Rho	Chi-square	Difference	P-value
Maternal age	-0.04449	0.13	1	0.7213
Parity	-0.03108	0.07	1	0.7910
Mode of delivery	0.33775	6.24	1	0.1025
Apgar score 1	-0.03850	0.08	1	0.7811
Apgar score 5	0.00627	0.00	1	0.9640
Gestational age	-0.10854	0.64	1	0.4222
Place of residence	0.09235	0.50	1	0.4784
Sex of child	-0.14390	1.16	1	0.2810
Complications of delivery	-0.07660	0.38	1	0.5378
Birth weight	0.02733	0.05	1	0.8189
Global test		10.96	10	0.3608

4.4.2 Comparing the PH models and the AFT models

In this section we first compared the parametric PH models to the cox regression model and selected the best fit model. We also compared the AFT models and selected the best fit

model. The best fit models for the PH and AFT models was then compared to ascertain the overall best fit model for the dataset.

Table 4.5 present the hazard ratios for the cox regression and PH parametric models (Exponential, Weibull and Gompertz). Results of the three PH models are presented in Appendix B. The hazard ratios for the PH models were compared for consistency. Generally, it can be observed that the values of the hazard ratio for the cox model and parametric PH models looks very similar. Due to the similarities in the hazard ratios, the log-likelihood values and AIC method was used to adjudge the best-fitting model.

Table 4.5: Comparison of Hazard Ratios for the PH models

Covariates	Cox		Exponential		Weibull		Gompertz	
	HR	Std. Err	HR	Std. Err	HR	Std. Err	HR	Std. Err
Maternal Age	0.9595	0.0313	0.9594	0.3260	0.9594	0.3810	0.9595	0.0312
Parity	1.3010	0.1533	1.3037	0.1176	1.3028	0.1409	1.3016	0.1531
Apgar score 1	0.6643	0.1193	0.6575	0.1804	0.6602	0.2161	0.6638	0.1201
Apgar score 5	1.2043	0.2091	1.2061	0.1730	1.2061	0.2019	1.2026	0.2089
Gestational age	1.0394	0.0536	1.0401	0.5143	1.0396	0.5975	1.0392	0.0537
Birth weight	0.4722	0.1099	0.4592	0.2373	0.4649	0.2897	0.4672	0.1102
delivery (SVD)	0.6930	0.5137	0.6807	0.3519	0.6856	0.4089	0.6898	0.5103
residence (Urban)	2.9260	0.1187	3.0183	0.3547	2.9855	0.4350	2.9542	0.1198
Deliveries (Yes)	1.1315	0.3629	1.1432	0.4127	1.1372	0.4765	1.1294	0.3655
Sex (Male)	1.2506	0.2117	1.2523	0.2650	1.2517	0.3075	1.2470	0.2126

From Table 4.6, it can be seen the Gompertz model has the smallest AIC value of 508.2979 compared to the other PH models. The log- likelihood value obtained also confirms that

Gompertz model is the best since the value of the log-likelihood is smaller than the other PH models. This means that Gompertz model is the best fitting model for the data under the PH modelling.

Table 4.6: Comparison of the PH models using Log-likelihood and AIC

Models	No. of Parameters	Difference	Log-likelihood	AIC
Cox	None	10	-354.5778	729.1556
Exponential	1	11	-244.1363	510.2720
Weibull	2	12	-243.454	510.9084
Gompertz	2	12	-242.1489	508.2979

We now use the AFT model with the exponential, weibull, log-logistic and Gompertz distributions to model the data. Table 4.7 present the comparison of standard errors for the AFT models. Results of the four AFT models are presented in Appendix B. It can be observed that overall the standard errors for all the models are quite similar.

Table 4.7: Comparison of Standard Errors for AFT models

Covariates	Exponential		Weibull		Log-logistics		Log-normal	
	Coef.	Std. Err	Coef.	Std. Err	Coef.	Std. Err	Coef.	Std. Err
Maternal Age	0.0414	0.0326	0.0479	0.0381	0.0481	0.3827	0.0533	0.0399
Parity	-0.2650	0.1176	-0.0305	0.1409	-0.3197	0.1448	-0.0367	0.1480
Apgar score 1	0.4192	0.1804	0.4792	0.2160	0.4576	0.2302	0.4920	0.2669
Apgar score 5	-0.0188	0.1730	-0.2163	0.2019	-0.1883	0.2219	-0.2142	0.2640
Gestational age	-0.0393	0.0514	-0.0449	0.5975	-0.0449	0.0626	-0.5912	0.7021
Birth weight	0.7783	0.2373	0.8839	0.2898	0.8990	0.2936	0.9264	0.3167
delivery (SVD)	-0.3846	0.3519	-0.4356	0.4089	-0.4495	0.4095	-0.4759	0.4208
residence (Urban)	1.1045	0.3547	1.2624	0.4349	1.2059	0.4285	1.1852	0.4202

Comp. (Yes)	0.1339	0.4127	0.1484	0.4765	0.1365	0.4925	0.2797	0.5170
Sex (Male)	0.2250	0.2649	0.2591	0.3075	0.2695	0.3144	0.2950	0.3317

The exponential AFT model, weibull AFT model, log-logistic AFT model and log-normal AFT model are compared using statistical criteria AIC and log likelihood ratio (the smaller AIC is the better). From Table 4.8, the Lognormal AFT model seems to be a suitable AFT model fitting the data according to AIC values compared with other AFT models. Thus the Lognormal AFT model is better than exponential, weibull and log-logistic under the AFT modelling.

Table 4.8: Comparison of the AFT models using Log-likelihood and AIC

Models	No. of Parameters	Difference	Log-likelihood	AIC
Exponential	1	11	-244.364	510.273
Weibull	2	12	-243.454	510.908
Log-logistics	2	12	-243.355	510.709
Lognormal	2	12	-242.683	509.365

We further used the log-likelihood ratios and AIC to select the overall best model for the data. Table 4.9 shows log-likelihood ratios and Akaike's Information Criterion (AIC) for the best PH and AFT models. According to this criterion, among the desired models, a model that has the lowest AIC is the best and the most efficient. Comparing the AIC values it can be observed that Gompertz model is the overall best fitted model to the data. Thus the data is best fitted using the Gompertz PH model.

Table 4.9: Comparison of the PH and AFT models using Log-likelihood and AIC

Models	No. of Parameters	Difference	Log-likelihood	AIC
Gompertz	2	12	-242.149	508.298
Lognormal	2	12	-242.683	509.365

The final model using the Gompertz PH is therefore presented in Table 4.10. The table reports on the hazard ratios, the standard errors, p-values and confidence intervals for the Gompertz PH model. It could be observed that parity, Apgar score 1, birth weight and place of residence are all significant at 5 percent significance. Other variables such as maternal age, Apgar 5, gestational age mode of delivery, complications of deliveries and sex of child were found not to be statistically significant. For the covariate parity, a unit increase in the number of live offspring by the neonate's mother leads to an increase in the hazard risk by 30.16% since the hazard ratio is greater than 1 ($HR = 1.3016,\ p-value = 0.025$). This means that a unit increase in the number of children a neonate mother has, decreases the probability of death of the neonate by 30 percent. This finding is consistent with the findings of Niragire et. al. (2011) who found that mothers who had more than one child expose their children to a high risk of experiencing the event (death) before reaching five years. Also the chance of a neonate dying before the 28^{th} day decreases by 33.62% for a unit increase in Apgar score 1 ($HR = 0.6638,\ p-value = 0.024$). Thus a unit increase in the Apgar score 1 increases the probability of death of the neonate by 34 percent. The hazard ratio for the birth weight was less than 1 ($HR = 0.4672,\ p-value = 0.001$) implying that a unit increase in the birth weight of a neonate decreases the hazard rate by 53%. In other words a unit increase in the birth weight of a neonate increases the probability of death of the neonate by 53%. The result is in congruence with the findings of Lanfranchi et al. (2011) whose finding showed that birth weight is significantly associated to neonatal death. Considering place of residence covariate, it could be observed that at 5 percent significance level the hazard ratio for urban is

0.3385. The hazard ratio means that the risk of death for neonates whose mothers are from the urban centres is 66% less than those from the rural areas. Thus neonates from urban centres have better chance of survival and their survival times are higher than neonates from rural areas. The result confirms the findings of Ezeh et al. (2014) whose finding revealed that neonates born to mothers residing in rural areas are at higher risk as compared to urban residents.

Table 4.10: The Full Model of the Gompertz PH

Covariates	Haz. Ratio	Std. Err.	Z	P>z	[95% Conf.	Interval]
Maternal age	0.95956	0.0312	-1.27	0.205	.9002827	1.022801
Parity	1.3016	0.1531	2.24	0.025	1.033608	1.639149
apgarscore1	0.6638	0.1201	-2.26	0.024	.4655193	.9465105
apgar5	1.2027	0.2089	1.06	0.288	.8556535	1.690407
Gestational age	1.0393	0.0537	0.75	0.455	.9392578	1.149972
Birth weight	0.4672	0.1102	-3.23	0.001	.2943084	.741647
Delivery (SVD)	1.4497	0.5103	1.05	0.291	.7271532	2.890092
Residence (Urban)	0.3385	0.1198	-3.06	0.002	.1690697	.6777315
Complications (Yes)	0.8854	.3654615	-0.29	0.768	.3942989	1.98833
Sex of child (Male)	0.8019	.2125761	-0.83	0.405	.4769632	1.348245
Intercept	0.1283	.2932737	-0.90	0.369	.0014557	11.31331
Gamma	-0.0326	.0165914	-1.97	0.049	-.0651427	-.0001056

4.5 Shared Frailty Model

The following analysis is based on the existence of unobserved heterogeneity among neonates from actual localities. It is important to mention that the localities (rural and urban) is

considered to be a random effect rather than a fixed effect. This is because neonates from their individual localities are not of interest by itself, rather the interest is in the heterogeneity between the localities. The application of the gamma frailty distribution to the survival distributions will be able to ascertain the existence or absence of the unobserved characteristics. The analysis made use of the Akaike information criterion to select the best frailty model for this analysis and subsequent analysis will be based on the best selected model.

Table 4.11: Log Likelihood, AIC, BIC for the gamma frailty of the survival distributions

Survival Distribution	Log Likelihood	AIC	BIC
Gompertz	-247.49162	518.9832	572.0036
Exponential	-247.94702	517.894	566.4961
Weibull	-247.89871	519.7974	572.8178
Lognormal	-247.45682	518.9136	571.934
Log logistic	-247.89435	519.7887	572.8091
Cox frailty	-360.13602	738.272	778.0373

Table 4.11 presents the AIC values for the shared gamma frailty of the survival models of interest. The model with the lowest value of AIC is considered the best model to fit the survival data when considering frailty. From Table 4.11, the AIC for the Exponential was the lowest among the five models. Hence the exponential with shared gamma frailty model will fit the data well when frailty or unobserved heterogeneity is of interest when considering the localities of the neonates. Therefore further analysis based on frailty will be centred on the application of the exponential with shared gamma frailty model. Results of the Cox with shared gamma frailty, Weibull, Lognormal, Log-logistic and Gompertz are present in Appendix B.

Table 4.12 presents the full model of the shared gamma frailty model with exponential as the baseline distribution. The model fitted was the exponential baseline hazard model with place of residence (Urban or Rural) effect as the frailty term. Using the likelihood ratio test with a null hypothesis that the variance of the frailty term is zero ($\theta = 3.6388$), the chi-square test statistic $\chi^2 = 4.35$ with a $p - value\ of\ 0.019$. The p-value less than the level of significance (0.05) hints the existence of unobserved heterogeneity between the localities. Thus in each category of neonates by locality has different values of random effects and there is heterogeneity of risks between locality categories (rural and urban). This means that neonates from the urban and rural residence both have different risk and hazards.

Table 4.12: Shared Gamma frailty model for the Exponential Baseline Hazard

Covariates	Haz. Ratio	Std. Err.	Z	P>z	[95% Conf.	Interval]
Maternal age	.9577005	.0374153	-1.11	0.269	.8871051	1.033914
Parity	1.375102	.2090662	2.10	0.036	1.020754	1.85246
Apgar score1	.6386289	.1678128	-1.71	0.088	.3815733	1.068856
Apgar score 5	1.181118	.3113012	0.63	0.528	.7046065	1.979886
Birth weight	.4014994	.1221756	-3.00	0.003	.2211392	.7289605
Mode of delivery	.6150875	.259509	-1.15	0.249	.2690358	1.406254
Complications	1.206944	.6548393	0.35	0.729	.4167342	3.495545
Sex of child (male)	1.313656	.4345925	0.82	0.410	.6868811	2.51236
Constant	.1342997	.3821733	-0.71	0.480	.000508	35.50738
Ln of theta	1.291654	.6337162	2.04	0.042	.0495933	2.533715
Theta	3.638801	2.305967		1.050844	12.60023	

Likelihood-ratio test of theta=0: chi-square (01) = 4.35 p-value = 0.019

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATION

5.1 Introduction

This chapter presents a summary of the findings from the study as well as the conclusions, recommendations and the areas for future research. Thus the main results and findings on the performance of the survival models are presented.

5.2 Summary

The main aim of this study was to model neonatal mortality using survival analysis. The preliminary results revealed that out of the 632 neonates studied, 73 of them experienced the event whiles 559 of them were censored. Thus the mortality rate was 11.6 percent. It was observed that the survival times of neonates based on covariates such as sex of child, mode of delivery, complication during delivery, maternal age and gestational age were found not be statistically different. The survival time for Apgar score, birth weight and place of residence was however found to be significant. The mean and median of overall survival time was found to be 25.84 and 28.0 days respectively. The Kaplan Meier plot indicated that about 2 percent of the neonates experienced the event within day zero with majority of the neonates experiencing censoring throughout the time period of study. The K-M plot for each categorical covariate revealed that female neonates have shorter survival as compared to male. Also neonates born by SVD have lower chance of surviving as compared to those born by CS.

Again from the KM plot, neonates from the rural areas tend to have shorter survival time than those from the urban areas and neonates with very low birth weight were seen to have lower chances of survival compared to those with low and normal weight. Furthermore neonates

with adequate Apgar score (greater than or equal to 7) had longer survival compared to neonates with high risk Apgar score (less than 6) and finally neonates born with complications had longer survival than those born without complications during delivery. The Log-rank test carried out however suggested that the differences between covariates such as mode of delivery, sex of child complications of delivery, maternal age and gestational age were not statistically significant.

In comparing the PH models with the Cox regression model using the log-likelihood ration and AIC, it was revealed that the Gompertz PH model was the best fitted model for the data under the PH model since its AIC values were smaller compared to the exponential, weibull and cox model. A comparison of the AFT model with distributions such the exponential, weibull, lognormal and log-logistics was also carried out and was revealed that the lognormal AFT model best fit the data under the AFT model. The Gompertz PH model was also compared to the Lognormal AFT model using their AIC values obtained. This was done to select in other to select the overall best model for the dataset. It was found that the Gompertz PH model was the best model for data since its AIC value was smaller than the Lognormal AFT model.

Modelling with the Gompertz PH model, it was revealed that parity, Apgar score 1, birth weight and place of residence were all significant at 5 percent significance level. Other variables such as maternal age, Apgar 5, gestational age mode of delivery, complications of deliveries and sex of child were found not to be statistically significant. Thus changes in the covariates correspond to changes in the survival time of the neonates. For instance the results showed that a unit increase in the number of children a neonate mother decreases the survival time of the neonate by 30 percent. Also a unit increase in the Apgar score 1 increases the

survival time of a neonate by 34 percent and unit increase in the birth weight of a neonate increases the survival time of the neonate by 53 percent. Finally, the risk of failure of neonates from urban centres was less than neonates from the rural areas by 66 percent.

Finally, the cox proportional hazard model with neonate's locality effect as the frailty term revealed that there are observed heterogeneity among neonates from the urban and rural centres.

5.3 Conclusion

Based on the results obtained, the study concludes that the average survival time for neonates considered was 25.84 days. Thus on the average neonates are expected to survive up to the 25^{th} day. The study also concludes that the Gompertz PH regression model best fit the data, hence can provide an accurate results than the Cox model and AFT models such as the exponential, weibull, lognormal and log-logistic. The study provides some evidence for association between some selected explanatory variables and the survival time of neonates. Neonate whose mother has more live offspring is less likely to survive as compared to neonate whose mother has less live offspring. Similarly, an increase in the weight of a neonate corresponds to an increase in the survival time of the neonate. Also neonates from rural areas are less likely to survive as compared to neonates from the rural areas. Finally, the study concludes that there were unobserved heterogeneity or frailty effect in categories of neonates based on place of residence. Hence neonates in the same urban or rural areas share the same unobserved frailty. The study finally concludes that there are unobserved characteristics among neonates from the urban and rural centres.

5.4 Recommendations

Based on the results from the study the following recommendations were made:

- Though most researchers apply the Cox model for the survival analysis, parametric model has however proven to have the ability to present better results than Cox model. The study therefore recommends that parametric models should be considered when fitting a survival model to a data.
- Since the findings from the study suggest that there are some unobserved characteristics were not accounted for, the study recommends that further studies must be carried out to explore those characteristics and their impact on neonatal mortality.
- The study further recommends that mothers should be educated on family planning methods since it has been shown that mothers with more children expose their neonates to a higher risk of death before reaching the 28^{th} day.
- Finally the study recommends a follow up study which is a good alternative to secondary data. Thus a cohort of mothers may be followed up for 28^{th} day and the death and birth histories of their neonates recorded.

5.5 Strength and Limitation of study

The main strength of this work is its large sample size and this helped in obtaining a more reliable and accurate estimates as well as estimation of standard errors. However, there were some limitations imposed during the conduct of the study. The data used for the study had lots of neonates censored. Nardi and Schemper (2003) indicated that to obtain an appropriate fit for parametric models, the percentage of censoring should not exceed 50 percent. The data used for the however had about 88 percent censoring. Also relevant information about neonate mothers such as educational level, visit to antenatal care and socioeconomic status were not available.

REFERENCES

Aalen, O., Borgan, O., & Gjessing, H. (2008). *Survival and event history analysis: a process point of view*: Springer Science & Business Media.

AFZAL, A. R., & ALAM, S. Analysis and Comparison of Under Five Child Mortality between Rural and Urban Area in Bangladesh. *Journal Of Applied Quantitative Methods*, 1.

Araújo, B. F. d., Tanaka, A. C. d. A., Madi, J. M., & Zatti, H. (2005). Newborn mortality study in the neonatal intensive care unit of Caxias do Sul General Hospital, Rio Grande do Sul. *Revista Brasileira de Saúde Materno Infantil, 5*(4), 463-469.

Black, R. E., Morris, S. S., & Bryce, J. (2003). Where and why are 10 million children dying every year? *The lancet, 361*(9376), 2226-2234.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, 89-99.

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.

Cleves, M. (2008). *An introduction to survival analysis using Stata*: Stata Press.

Collett, D. (2015). *Modelling survival data in medical research*: CRC press.

Cox, D. R. (1972). lRegression Models and Life Tables, mJournal of the Royal Statistical Society. *Series B, 34*(2), 187&220.

Dahiru, T. (2015). Surviving the First Day in Nigeria: Risk Factors and Protectors. *American Journal of Public Health, 3*(4A), 19-26.

Dos Santos, S., & Henry, S. (2008). Rainfall variation as a factor in child survival in rural Burkina Faso: the benefit of an event-history analysis. *Population, Space and Place, 14*(1), 1-20.

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

Duchateau, L., Janssen, P., Lindsey, P., Legrand, C., Nguti, R., & Sylvester, R. (2002). The shared frailty model and the power for heterogeneity tests in multicenter trials. *Computational Statistics & Data Analysis, 40*(3), 603-620.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American statistical association, 72*(359), 557-565.

Engmann, C., Walega, P., Aborigo, R. A., Adongo, P., Moyer, C. A., Lavasani, L., Hodgson, A. (2012). Stillbirths and early neonatal mortality in rural Northern Ghana. *Tropical Medicine & International Health, 17*(3), 272-282.

Ezeh, O. K., Agho, K. E., Dibley, M. J., Hall, J., & Page, A. N. (2014). Determinants of neonatal mortality in Nigeria: evidence from the 2008 demographic and health survey. *BMC public health, 14*(1), 1.

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

Folasade, I. B. (2000). Environmental factors, situation of women and child mortality in southwestern Nigeria. *Social Science & Medicine, 51*(10), 1473-1489.

Gandotra, M., & Das, N. (1990). Contraceptive choice shift and use continuation: a prospective study in Gujarat. *Journal of Family Welfare, 36*(3), 54-69.

Ghana Statistical Service (2015). *Ghana demographic and health survey, 2015*.

Goldenberg, R. L., Culhane, J. F., Iams, J. D., & Romero, R. (2008). Epidemiology and causes of preterm birth. *The lancet, 371*(9606), 75-84.

Haghighi, L., Nojomi, M., Mohabbatian, B., & Najmi, Z. (2013). Survival predictors of preterm neonates: Hospital based study in Iran (2010-2011). *Iranian journal of reproductive medicine, 11*(12), 957.

Hanagal, D. D. (2011). *Modeling survival data using frailty models*: CRC Press.

Health, W. H. O. R. (2003). *Pregnancy, childbirth, postpartum, and newborn care: a guide for essential practice*: World Health Organization.

Hosmer, D. W., Lemeshow, S., & May, S. (2008). Model development. *Applied Survival Analysis: Regression Modeling of Time-to-Event Data, Second Edition*, 132-168.

Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika, 73*(2), 387-396.

Hougaard, P. (1995). Frailty models for survival data. *Lifetime data analysis, 1*(3), 255-273.

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

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association, 53*(282), 457-481.

Kassar, S. B., Melo, A. M., Coutinho, S. B., Lima, M. C., & Lira, P. I. (2013). Determinants of neonatal death with emphasis on health care during pregnancy, childbirth and reproductive history. *Jornal de pediatria, 89*(3), 269-277.

Kayode, G. A., Ansah, E., Agyepong, I. A., Amoakoh-Coleman, M., Grobbee, D. E., & Klipstein-Grobusch, K. (2014). Individual and community determinants of neonatal mortality in Ghana: a multilevel analysis. *BMC pregnancy and childbirth, 14*(1), 1.

Kleinbaum, D., & Klein, M. (2005). Survival analysis: A self-learning approach. *Springer, New York, USA*.

Kojo, K. (2012). *Modelling the risk factors of neonatal mortality in ghana using logistic regression.* department of Mathematics, Kwame Nkrumah University of Science and Technology.

Lanfranchi, L. M. M., Viola, G. R., & Nascimento, L. F. C. (2011). The use of Cox regression to estimate the risk factors of neonatal death in a private NICU. *Revista Paulista de Pediatria, 29*(2), 224-230.

Lawn, J., McCarthy, B. J., & Ross, S. R. (2013). The Healthy Newborn. *A Reference Manual for Program Managers The WHO Collaborating Center in Reproductive Health CDC Care-CDC Health Initiative The Health Unit Care*.

Lawn, J. E., Cousens, S., Zupan, J., & Team, L. N. S. S. (2005). 4 million neonatal deaths: when? Where? Why? *The lancet, 365*(9462), 891-900.

Li, C., Yan, H., Zeng, L., Dibley, M. J., & Wang, D. (2015). Predictors for neonatal death in the rural areas of Shaanxi Province of Northwestern China: a cross-sectional study. *BMC public health, 15*(1), 1.

Liu, L., Johnson, H. L., Cousens, S., Perin, J., Scott, S., Lawn, J. E., & Li, M. (2012). Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. *The lancet, 379*(9832), 2151-2161.

Manda, S. O. (1999). Birth intervals, breastfeeding and determinants of childhood mortality in Malawi. *Social Science & Medicine, 48*(3), 301-312.

Matthews, Z., Channon, A., Neal, S., Osrin, D., Madise, N., & Stones, W. (2010). Examining the "urban advantage" in maternal health care in developing countries. *PLoS Med, 7*(9), e1000327.

Mekonnen, Y., Tensou, B., Telake, D. S., Degefie, T., & Bekele, A. (2013). Neonatal mortality in Ethiopia: trends and determinants. *BMC public health, 13*(1), 1.

Mercer, A., Haseen, F., Huq, N. L., Uddin, N., Khan, M. H., & Larson, C. P. (2006). Risk factors for neonatal mortality in rural areas of Bangladesh served by a large NGO programme. *Health Policy and Planning, 21*(6), 432-443.

Mesike, C. G., & Mojekwu, J. N. (2012). Environmental determinants of child mortality in Nigeria. *Journal of Sustainable Development, 5*(1), 65.

Mosley, W. H., & Chen, L. C. (1984). An analytical framework for the study of child survival in developing countries. *Population and development review, 10*, 25-45.

Nardi, A., & Schemper, M. (2003). Comparing Cox and parametric models in clinical studies. *Statistics in Medicine, 22*(23), 3597-3610.

Nasejje, J. B., Mwambi, H. G., & Achia, T. N. (2015). Understanding the determinants of under-five child mortality in Uganda including the estimation of unobserved household and community effects using both frequentist and Bayesian survival analysis approaches. *BMC public health, 15*(1), 1.

Nguti, R. W. (2003). *Random effects survival models applied to animal breeding data*: LUC.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American statistical association, 84*(406), 487-493.

Organization, W. H. (2012). Born too soon: the global action report on preterm birth.

Osei-Kwakye, K., Otupiri, E., Dabo, E. O., Browne, E., & Adjuik, M. (2010). Determinants of Under-Five Mortality in Builsa District, Upper East Region, Ghana. *Journal of Science and Technology (Ghana), 30*(1).

Parlato, R., Darmstadt, G. L., Tinker, A. G., & Lives, S. N. (2004). *Saving Newborn Lives: Tools for Newborn Health: Qualitative Research to Improve Newborn Care Practices*: Saving Newborn Lives Initiative.

Peristat, I. (2008). EURO-PERISTAT project in collaboration with SCPE, EUROCAT and EURONEONET. European perinatal health report. Better statistics for better health for pregnant women and their babies in 2004. 2008.

Pickett, G., & Hanlon, J. (1990). Philosophy and purpose of public health. *Public Health Administration and Practice. 9th Edition. St. Louis: Times Mirror/Mosby College Publishing*, 3-20.

Rajaratnam, J. K., Marcus, J. R., Levin-Rector, A., Chalupka, A. N., Wang, H., Dwyer, L., & Murray, C. J. (2010). Worldwide mortality in men and women aged 15–59 years from 1970 to 2010: a systematic analysis. *The lancet, 375*(9727), 1704-1720.

Rinne, H. (2008). *The Weibull distribution: a handbook*: CRC Press.

Safer, M. P. (2007). Neonatal and perinatal mortality.

Siakwa, M., Kpikpitse, D., Laryea, T., Ankobil, A., Dare, S., & Ebu, N. A five-year neonatal mortality trend in a Ghanaian Teaching Hospital after the implementation of strategies to achieve the Millenium Development Goal (MDG) 4.

Stevenson, M., & EpiCentre, I. (2009). An introduction to survival analysis. *Palmerston North, NZ: EpiCentre, Massey University*.

Therneau, T. M., Grambsch, P. M., & Fleming, T. R. (1990). Martingale-based residuals for survival models. *Biometrika, 77*(1), 147-160.

UNICEF. (2008). *The state of the world's children 2009: maternal and newborn health* (Vol. 9): Unicef.

Van den Broek, N., & Graham, W. (2009). Quality of care for maternal and newborn health: the neglected agenda. *BJOG: An International Journal of Obstetrics & Gynaecology, 116*(s1), 18-21.

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

Victora, C. G., Black, R. E., & Bryce, J. (2007). Learning from new initiatives in maternal and child health. *The lancet, 370*(9593), 1113-1114.

Watson, G., & Wells, W. (1961). On the possibility of improving the mean useful life of items by eliminating those with short lives. *Technometrics, 3*(2), 281-298.

Wintrebert, C., Putter, H., Zwinderman, A. H., & Van Houwelingen, J. (2004). Centre-effect on Survival after Bone Marrow Transplantation: Application of Time-dependent Frailty Models. *Biometrical journal, 46*(5), 512-525.

World Health Organization (2003). *Pregnancy, childbirth, postpartum, and newborn care: a guide for essential practice*.

World Health Organization (2012). *Born too soon: the global action report on preterm birth*.

Worku, Z. (2009). *Factors that affect under-five mortality among South African children: analysis of the South African Demographic and Health Survey Data Set.* Paper presented at the Proceedings of the World Congress on Engineering and Computer Science.

Yehuala, S., Ayalew, S., & Teka, Z. (2015). Survival Analysis of Premature Infants Admitted to Neonatal Int ensive Care Unit (NICU) in Northwest Ethiopia using Semi-Parametric Fr ailty Model. *Journal of Biometrics & Biostatistics, 2015*.

Yinger, N. V., & Ransom, E. I. (2003). Why invest in newborn health. *Population Reference Bureau: Washington, DC*.

Appendix A

Graphs for Chapter Four

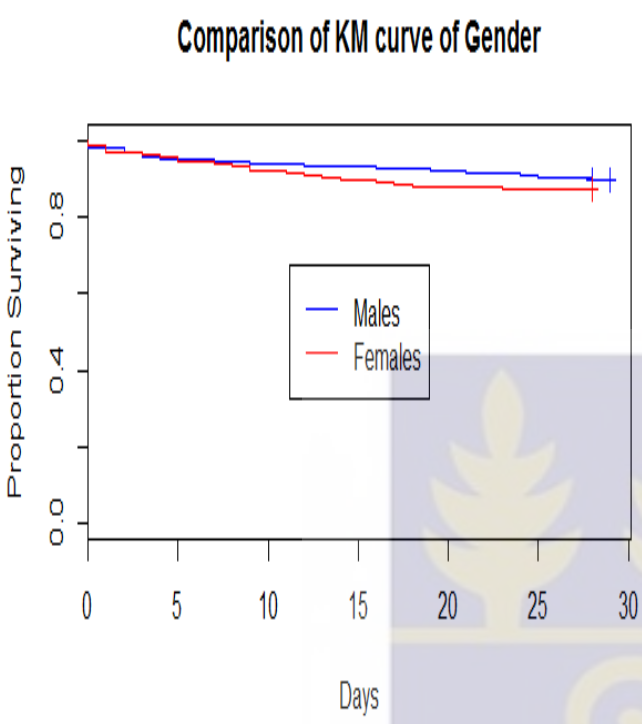

Figure 4.2: KM curve of Sex

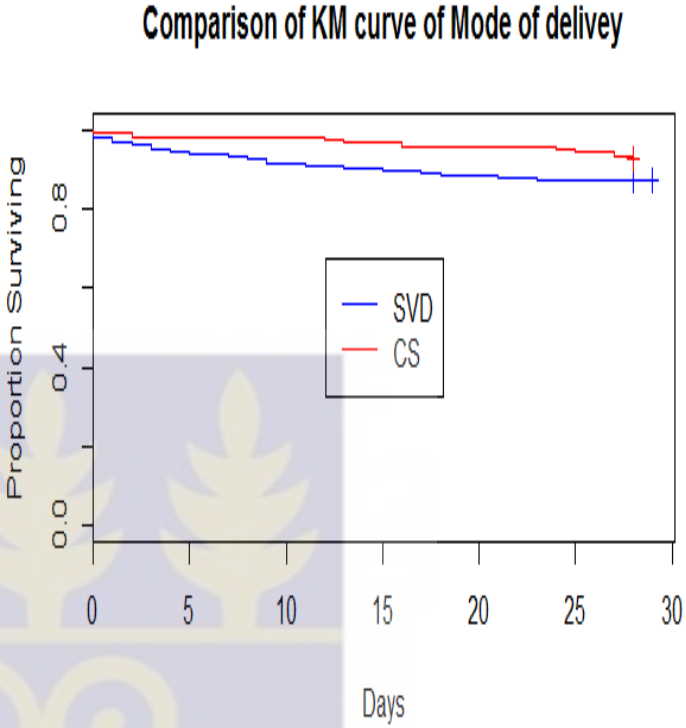

Figure 4.3: KM curve of Mode of Delivery

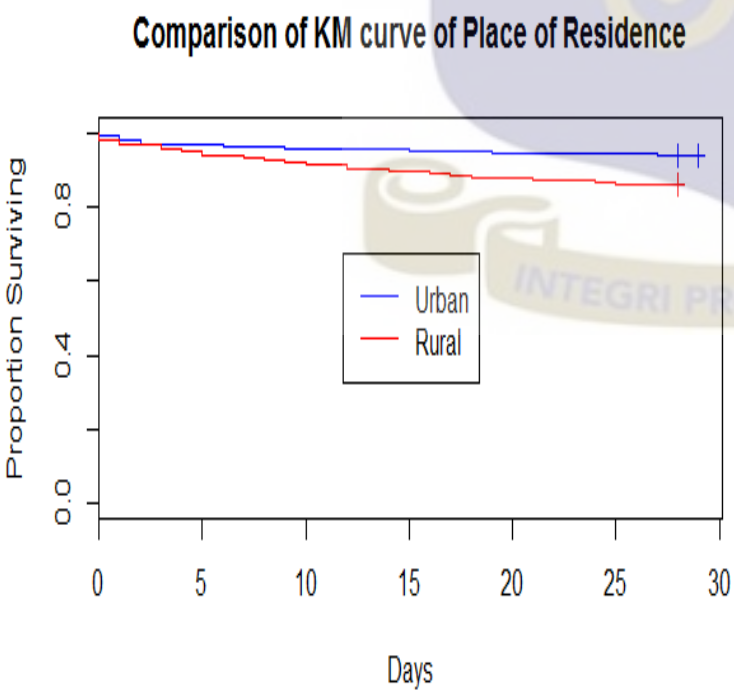

Figure 4.4: KM curve of Place of residence

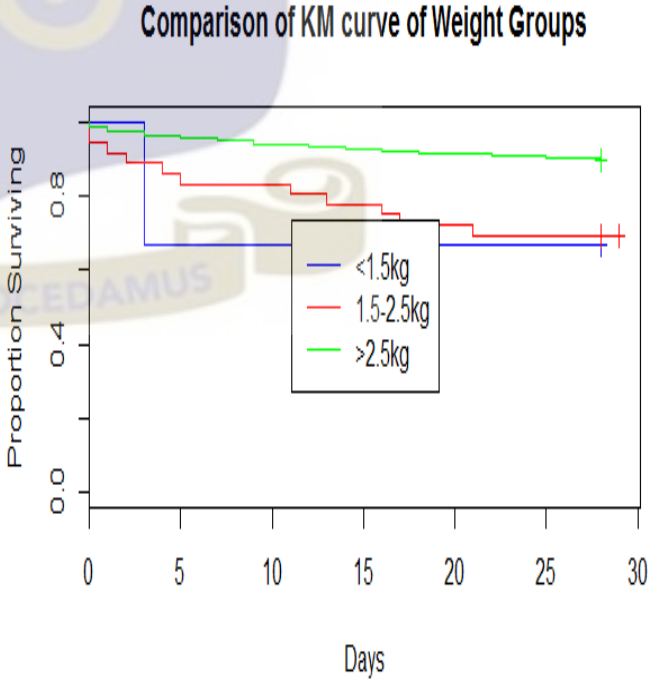

Figure 4.5: KM curve of Birth Weight

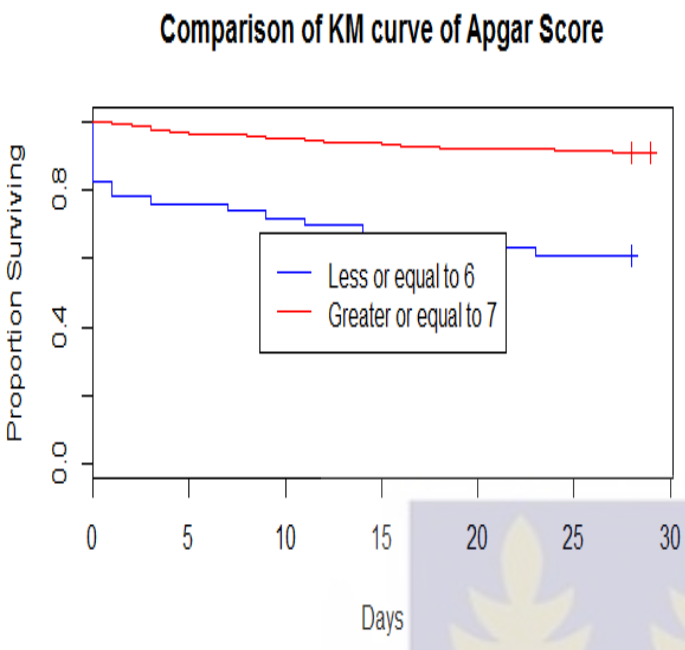

Figure 4.6: KM curve of Apgar score

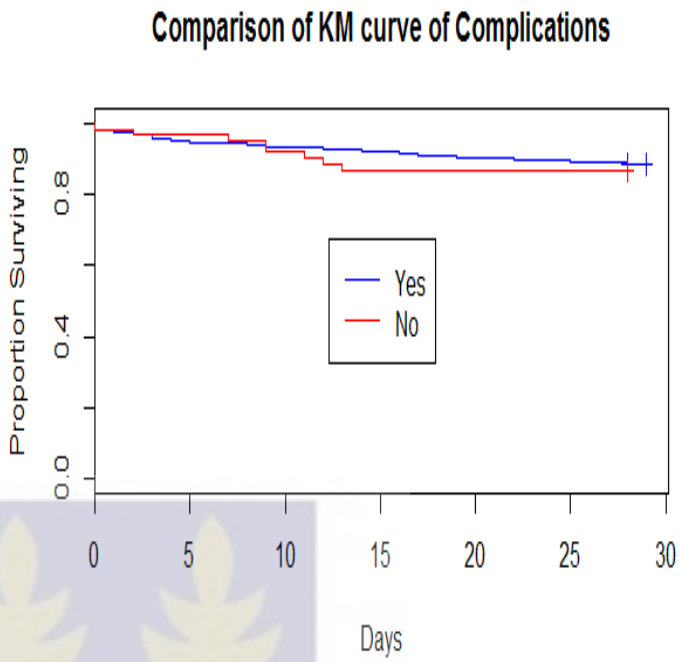

Figure 4.7: KM curve of complications

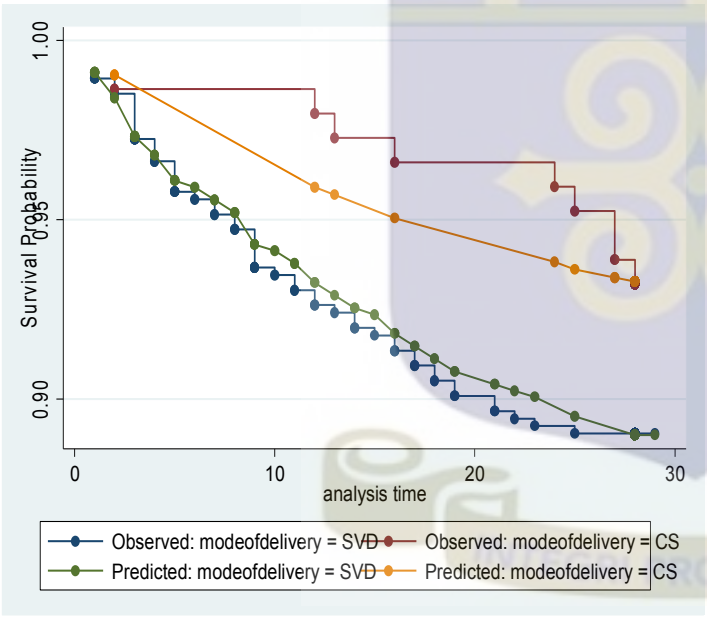

Figure 4.8: KM curve for Mode of delivery

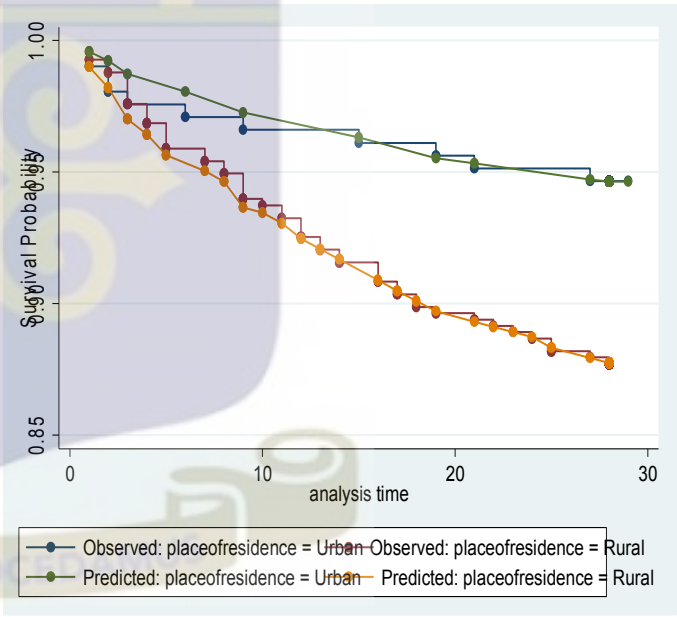

Figure 4.9: KM curve for Place of residence

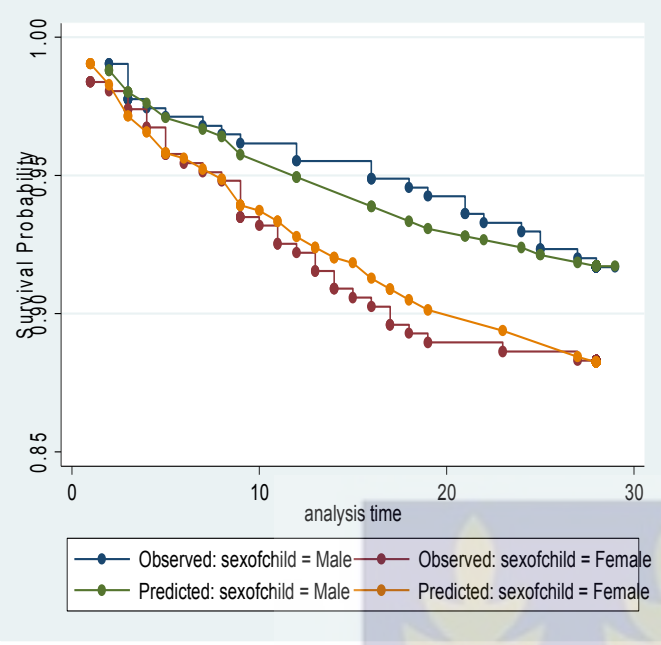

Table 4.10: KM curve for Sex of neonate

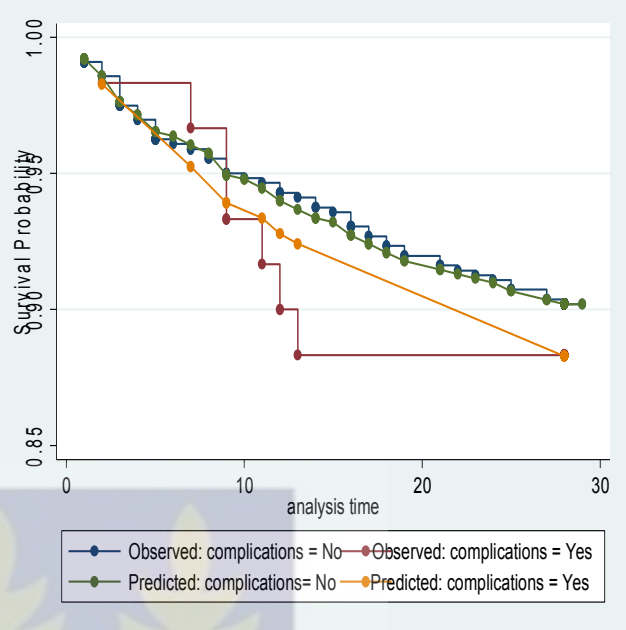

Figure 4.11: KM curve for complications

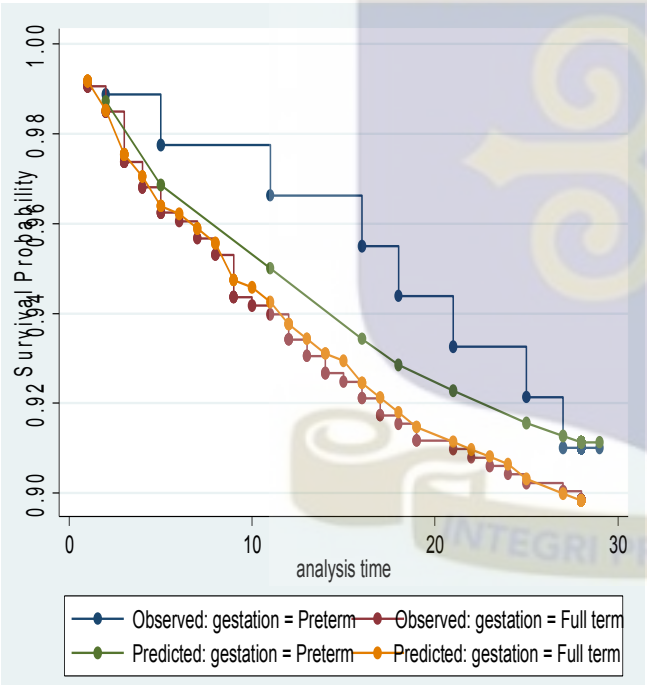

Figure 4.12:KM curve for gestational Age

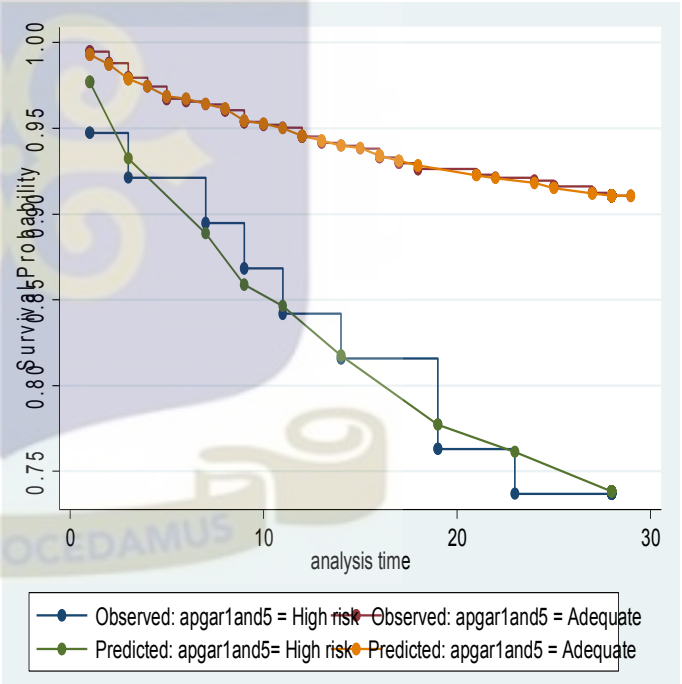

Figure 4.13: KM curve for Apgar 1 and 5

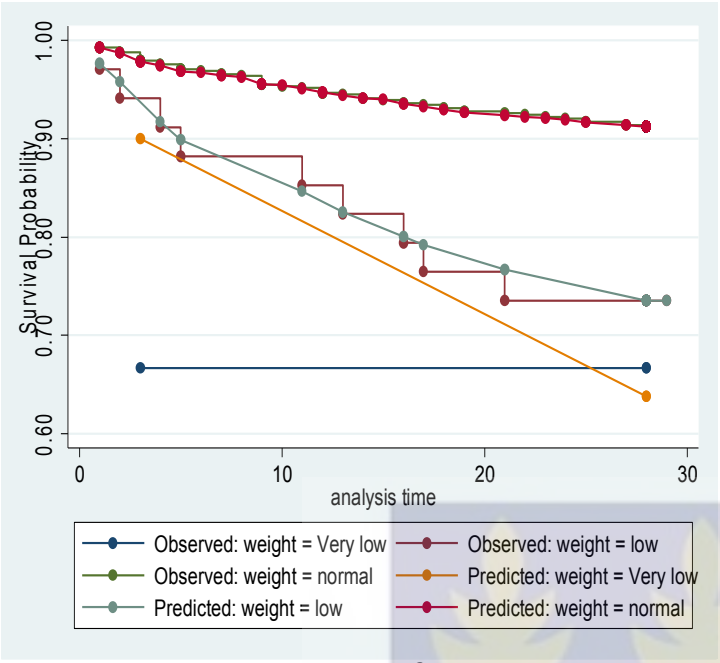

Figure 4.14: KM curve for Birth weight

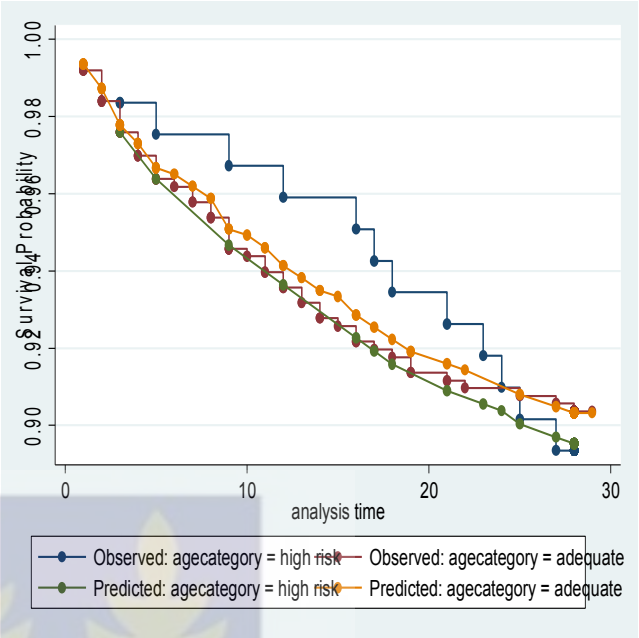

Figure 4.15: KM curve for Maternal Age

APPENDIX B

Tables for Chapter Four

Table 4.13: The Cox PH Full Model

Covariate	Haz. Ratio	Std. Err.	Z	P>z	[95% Conf.	Interval]
Maternal Age	.9594739	.0312304	-1.27	0.204	.9001751	1.022679
Parity	1.300974	.1530728	2.24	0.025	1.033038	1.638405
Apgar score 1	.6643372	.1203846	-2.26	0.024	.4657387	.9476214
Apgar score 5	1.204269	.2094558	1.07	0.285	.856399	1.693445
Gestational age	1.03937	.0538076	0.75	0.456	.9390825	1.150367
Birth weight	.4721516	.1109354	-3.19	0.001	.2979104	.7483028
Mode of delivery (SVD)	.693045	.2439821	-1.04	0.298	.3476172	1.381725
Place of residence (Urban)	2.926044	1.035345	3.03	0.002	1.462492	5.85421
Complications (Yes)	1.131509	.4670165	0.30	0.765	.5038892	2.540862
Sex of child (Male)	1.250687	.3315366	0.84	0.399	.7438909	2.102753

Table 4.14: The Exponential PH Full Model

Covariates	Haz. Ratio	Std. Err.	Z	P>z	[95% Conf.	Interval]
Maternal Age	.9594471	.0312788	-1.27	0.204	.9000594	1.022753
Parity	1.303737	.1533622	2.25	0.024	1.035287	1.641795
Apgar score 1	.6575741	.118617	-2.32	0.020	.4617427	.9364602
Apgar score 5	1.206773	.208782	1.09	0.277	.859727	1.693912
Gestational age	1.040133	.0534982	0.77	0.444	.9403902	1.150455
Birth weight	.4591984	.1089774	-3.28	0.001	.2883981	.7311531
Mode of delivery (SVD)	1.468977	.5169513	1.09	0.274	.7369987	2.927949
Place of residence (Urban)	.3313147	.1175311	-3.11	0.002	.1653043	.6640447
Complications (Yes)	.8747046	.3610018	-0.32	0.746	.3895469	1.964098
Sex of child (Male)	.7985119	.2115948	-0.85	0.396	.4750338	1.342265
_cons	.0911155	.2074322	-1.05	0.293	.0010514	7.896336

Table 4.15: The Weibull PH Full Model

Covariates	Haz. Ratio	Std. Err.	Z	P>z	[95% Conf.	Interval]
Maternal Age	.959379	.0312623	-1.27	0.203	.9000217	1.022651
Parity	1.30283	.1532644	2.25	0.025	1.034553	1.640676
Apgar score 1	.6602256	.1193246	-2.30	0.022	.4632892	.9408766
Apgar score 5	1.206145	.2091456	1.08	0.280	.8586205	1.69433
Gestational age	1.039634	.0536232	0.75	0.451	.939672	1.150229
Birth weight	.4649223	.1099046	-3.24	0.001	.2925242	.7389225
Mode of delivery (SVD)	1.458515	.5133719	1.07	0.284	.7316493	2.907496
Place of residence (Urban)	.3349545	.1187285	-3.09	0.002	.167212	.6709713
Complications (Yes)	.8793714	.3629064	-0.31	0.755	.3916438	1.974483
Sex of child (Male)	.7989343	.2117084	-0.85	0.397	.4752832	1.342981
_cons	.1355409	.3119954	-0.87	0.385	.0014884	12.34274
/ln_p	-.1433699	.1258857	-1.14	0.255	-.3901014	.1033616
P	.8664335	.1090716		0.6769	1.108892	
1/p	1.154157	.1452918		0.9018	1.477131	

Table 4.16: The Exponential AFT Full Model

Covariates	Coef.	Std. Err.	Z	P>z	[95% Conf.	Interval]
Maternal Age	.0413981	.0326008	1.27	0.204	-.0224983	.1052946
Parity	-.2652344	.1176328	-2.25	0.024	-.4957904	-.0346784
Apgar score 1	.4191979	.1803858	2.32	0.020	.0656482	.7727475
Apgar score 5	-.1879501	.1730085	-1.09	0.277	-.5270405	.1511403
Gestational age	-.0393485	.051434	-0.77	0.444	-.1401572	.0614603
Birth weight	.7782729	.2373209	3.28	0.001	.3131324	1.243413
Mode of delivery (SVD)	-.3845665	.3519124	-1.09	0.274	-1.074302	.3051691
Place of residence (Urban)	1.104686	.3547415	3.11	0.002	.4094059	1.799967
Complications (Yes)	.133869	.4127127	0.32	0.746	-.6750331	.9427711
Sex of child (Male)	.2250054	.2649864	0.85	0.396	-.2943585	.7443693
_cons	2.395627	2.276586	1.05	0.293	-2.066399	6.857654

Table 4.17: The Weibull AFT Full Model

Covariates	Coef.	Std. Err.	Z	P>z	[95% Conf.	Interval]
Maternal Age	.0478618	.0381025	1.26	0.209	-.0268178	.1225414
Parity	-.3053196	.1409163	-2.17	0.030	-.5815104	-.0291287
Apgar score 1	.4791754	.2160552	2.22	0.027	.0557149	.9026358
Apgar score 5	-.216323	.2019287	-1.07	0.284	-.6120959	.1794499
Gestational age	-.0448602	.0597513	-0.75	0.453	-.1619706	.0722501
Birth weight	.8839513	.2897499	3.05	0.002	.3160519	1.451851
Mode of delivery (SVD)	-.4356007	.4089619	-1.07	0.287	-1.237151	.3659499
Place of residence (Urban)	1.262371	.43498	2.90	0.004	.4098261	2.114916
Complications (Yes)	.1483645	.4764789	0.31	0.756	-.7855171	1.082246
Sex of child (Male)	.2590811	.3075344	0.84	0.400	-.3436753	.8618375
_cons	2.306561	2.631943	0.88	0.381	-2.851953	7.465074
/ln_p	-.1433699	.1258857	-1.14	0.255	-.3901014	.1033616
P	.8664335	.1090716		.6769	1.108892	
1/p	1.154157	.1452918		.9018	1.477131	

Table 4.18: The Lognormal AFT Full Model

Covariates	Coef.	Std. Err.	Z	P>z	[95% Conf.	Interval]
Maternal Age	.0533049	.0399084	1.34	0.182	-.0249142	.131524
Parity	-.3671012	.1479785	-2.48	0.013	-.6571337	-.0770687
Apgar score 1	.4919935	.2668607	1.84	0.065	-.0310439	1.015031
Apgar score 5	-.2142249	.2640447	-0.81	0.417	-.7317431	.3032932
Gestational age	-.0591229	.0702126	-0.84	0.400	-.196737	.0784913
Birth weight	.9264385	.3167293	2.93	0.003	.3056605	1.547216
Mode of delivery (SVD)	-.4758637	.4207663	-1.13	0.258	-1.30055	.3488231
Place of residence (Urban)	1.185203	.420146	2.82	0.005	.3617321	2.008674
Complications (Yes)	.2797424	.5169707	0.54	0.588	-.7335016	1.292986
Sex of child (Male)	.2950255	.331715	0.89	0.374	-.3551241	.945175
_cons	2.724229	2.985173	0.91	0.361	-3.126603	8.575061
/ln_sig	.8272501	.1124078	7.36	0.000	.6069348	1.047565
Sigma	2.287021	.257079			1.8347	2.850702

Table 4.19: The Log-Logistics AFT Full Model

Covariates	Coef.	Std. Err.	Z	P>z	[95% Conf.	Interval]
Maternal Age	.0481086	.0382734	1.26	0.209	-.0269059	.1231231
Parity	-.3197273	.1448188	-2.21	0.027	-.603567	-.0358877
Apgar score 1	.457617	.2301865	1.99	0.047	.0064597	.9087742
Apgar score 5	-.1882795	.22192	-0.85	0.396	-.6232347	.2466758
Gestational age	-.0449438	.0626862	-0.72	0.473	-.1678065	.0779189
Birth weight	.8990328	.2935816	3.06	0.002	.3236235	1.474442
Mode of delivery (SVD)	-.4495148	.4094669	-1.10	0.272	-1.252055	.3530255
Place of residence (Urban)	1.205904	.4284885	2.81	0.005	.3660818	2.045726
Complications (Yes)	.1364848	.4925404	0.28	0.782	-.8288766	1.101846
Sex of child (Male)	.2695916	.3143682	0.86	0.391	-.3465587	.885742
_cons	1.997416	2.707233	0.74	0.461	-3.308663	7.303496
/ln_gam	.084913	.1236659	0.69	0.492	-.1574677	.3272936
Gamma	1.088622	.1346254		.8543	1.387209	

Table 4.20: Univariate Cox shared gamma frailty model

Covariates	Hazard Ratio	Std. Err	P-value	[95% Conf.	Interval]	chibar2 (01)	Prob>= chibar2
Maternal Age	1.008659	.02151	0.686	.9673	1.0517	2.72	0.050
Theta	.1196684	.16503					
Parity	1.16405	.08601	0.040	1.0071	1.3454	3.55	0.030
Theta	.141509	.18513					
Apgar score 1	.775297	0.0382	0.000	0.7039	0.8539	3.02	0.041
Theta	.1273888	0.1720					
Apgar score 5	.8044742	0.0376	0.000	0.7339	0.8818	3.69	0.027
Theta	.1542073	0.2005					
Gestation	1.027446	0.0578	0.630	0.9202	1.1471	3.07	0.040
Theta	.1286427	. 1732					
Birth weight	.4634895	0.1072	0.001	0.2945	0.7294	3.08	0.040
Theta	.130064	0.1748					
Sex of child (Male)	.6907069	0.1777	0.151	0.4170	1.1438	3.03	0.041
Theta	.1277023	0.1723					
Complications	.8222175	0.3299	0.626	0.3744	1.8054	3.04	0.041
Theta	.1278718	0.1724					

Table 4.21: Multivariable Cox Shared Gamma Frailty model with Place of residence effect

Covariates	Haz. Ratio	Std. Err.	z	P>z	[95% Conf.	Interval]
Maternal age	.9719105	.0299099	-0.93	0.355	.9150211	1.032337
parity	1.232765	.1376381	1.87	0.061	.9904759	1.534324
Mode of delivery	.6667374	.2340556	-1.15	0.248	.3350764	1.326679
apgarscore1	.6830554	.1260053	-2.07	0.039	.4758085	.9805724
apgar5	1.179539	.2136776	0.91	0.362	.8270164	1.682327
Gestational age (weeks)	1.037843	.05715	0.67	0.500	.9316635	1.156123
Sex of child	1.306497	.3466444	1.01	0.314	.7767206	2.197618
Complications of deliveries	1.139827	.4719772	0.32	0.752	.5062622	2.566272
Weight (KG)	.5100957	.1142881	-3.00	0.003	.3288047	.7913441

Likelihood-ratio test of theta=0: chibar2 (01) =0.19 Prob>=chibar2 = 0.329

Table 4.22: Shared Gamma frailty model for the Weibull Baseline Hazard

_t	Haz. Ratio	Std. Err.	z	P>z	[95% Conf.	Interval]
Maternal age	.9551704	.039996	-1.10	0.273	.8799103	1.036868
Parity	1.409093	.2560381	1.89	0.059	.9868995	2.011901
apgarscore1	.6285143	.1805423	-1.62	0.106	.357936	1.103634
apgar5	1.189789	.3388862	0.61	0.542	.6808055	2.079298
Birth weight	.4038306	.1324639	-2.76	0.006	.2123193	.7680845
Mode of delivery	.5994196	.2702798	-1.14	0.256	.2476978	1.450574
Complications	1.258193	.7618748	0.38	0.704	.3839885	4.122647
Gestation	1.063223	.0842756	0.77	0.439	.9102372	1.241922
Sex of child (male)	1.328702	.4648115	0.81	0.417	.6693556	2.637534
_cons	.1162164	.3528843	-0.71	0.478	.0003024	44.65722
/ln_p	.0617252	.2046368	0.30	0.763	-.3393555	.462806
/ln_the	1.532961	.9735343	1.57	0.115	-.3751316	3.441053
p	1.06367	.217666		.7122292	1.588525	
1/p	.9401412	.1923875		.6295147	1.404042	
theta	4.631869	4.509283		.6871988	31.2198	

Likelihood-ratio test of theta=0: chibar2(01) = 2.84 Prob>=chibar2 = 0.046

Table 4.23: Shared Gamma frailty model for the Lognormal Baseline Hazard

_t	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
Maternal age	.0447345	.039744	1.13	0.260	-.0331624	.1226314
Parity	-.3093212	.1491184	-2.07	0.038	-.6015878	-.0170545
apgarscore1	.4716848	.2751102	1.71	0.086	-.0675213	1.010891
apgar5	-.1856267	.2740625	-0.68	0.498	-.7227792	.3515259
Birth weight	.8950168	.316707	2.83	0.005	.2742824	1.515751
Mode of delivery	.4999359	.4219219	1.18	0.236	-.3270158	1.326888
Gestation	-.0677819	.0726152	-0.93	0.351	-.210105	.0745413
Complications	-.2387538	.5277583	-0.45	0.651	-1.273141	.7956334
Sex of child (male)	-.3054765	.3341984	-0.91	0.361	-.9604933	.3495403
_cons	3.381775	2.982476	1.13	0.257	-2.46377	9.22732
/ln_sig	.8255905	.1486136	5.56	0.000	.5343131	1.116868
/ln_the	-1.151108	5.958211	-0.19	0.847	-12.82899	10.52677
Sigma	2.283229	.3393189			1.706276	3.05527
Theta	.3162863	1.884501			2.68e-06	37300.88

Likelihood-ratio test of theta=0: chibar2(01) = 0.03 Prob>=chibar2 = 0.430

Table 4.24: Shared Gamma frailty model for the Log-logistic Baseline Hazard

_t	Coef.	Std. Err.	z	P>z	[95% Conf.	Interval]
Maternal age	.0437778	.0383003	1.14	0.253	-.0312894	.1188451
Parity	-.3210094	.1498458	-2.14	0.032	-.6147018	-.0273169
apgarscore1	.4325057	.2616212	1.65	0.098	-.0802625	.9452739
apgar5	-.1592359	.2589511	-0.61	0.539	-.6667707	.3482989
Birth weight	.9107612	.2978443	3.06	0.002	.3269971	1.494525
Mode of delivery	.4638822	.412418	1.12	0.261	-.3444422	1.272207
Gestation	-.0524693	.0692095	-0.76	0.448	-.1881175	.0831789
Complications	.1526024	.5261919	0.29	0.772	-.8787147	1.18392
Sex of child (male)	-.254111	.3241563	-0.78	0.433	-.8894457	.3812236
_cons	1.893426	2.772308	0.68	0.495	-3.540198	7.327051
/ln_gam	-.0332261	.1753413	-0.19	0.850	-.3768888	.3104365
/ln_the	.9084783	1.16633	0.78	0.436	-1.377487	3.194443
gamma	.9673198	.1696111		.6859923	1.36402	
theta	2.480545	2.893134		.2522116	24.39659	

Likelihood-ratio test of theta=0: chibar2 (01) = 1.41 Prob>=chibar2 = 0.118

Table 4.25: Shared Gamma frailty model for the Gompertz Baseline Hazard

_t	Haz. Ratio	Std. Err.	z	P>z	[95% Conf.	Interval]
Maternal age	.9637319	.034988	-1.02	0.309	.8975396	1.034806
Parity	1.308672	.1912939	1.84	0.066	.9826688	1.742827
apgarscore1	.6601706	.1504411	-1.82	0.068	.4223597	1.031882
apgar5	1.173065	.2631251	0.71	0.477	.7557757	1.820755
Birth weight	.4441491	.1298505	-2.78	0.006	.2504225	.7877427
Mode of delivery	.6408759	.2488615	-1.15	0.252	.2993918	1.371854
Gestation	1.045579	.0657991	0.71	0.479	.9242517	1.182834
Complications	1.139032	.5404626	0.27	0.784	.4494142	2.886855
Sex of child (male)	1.291013	.38633	0.85	0.393	.7181454	2.32086
_cons	.1265971	.3316938	-0.79	0.430	.0007451	21.50969
/gamma	-.0221849	.0222252	-1.00	0.318	-.0657455	.0213756
/ln_the	.568989	1.287382	0.44	0.659	-1.954234	3.092212
theta	1.76648	2.274135		.141673	22.02574	

Likelihood-ratio test of theta=0: chibar2(01) = 0.82 Prob>=chibar2 = 0.182

APPENDIX C

Data Analysis codes

R codes

```
neonatal5<- read.csv("C:/Users/Dan/Desktop/neodata5.csv")
head(neonatal5)
attach(neonatal5)
library(splines)
library(MASS)
library(survival)

neoKaplan5<-survfit(Surv( Survival.Time..Days.,Survival.status)~1,data=neonatal5)
summary(neoKaplan5)
##Kaplan Meier Plots##
plot(neoKaplan5,main="Kaplan Meier Estimated Survival Function Plot",ylab="Survival
Probability",xlab="Time(Days)")
par(mfrow=c(1,1))
##KM GENDER##
KMneogender<-
survfit(Surv(Survival.Time..Days.,Survival.status)~Sex.of.child,data=neonatal5)
KMneogender
summary(KMneogender)
plot(KMneogender,conf.int=FALSE,col=c('blue','red'),xlab="Days",ylab="Proportion
Surviving")
legend('center',c("Males","Females"),col=c('blue','red'),lty=1)
title("Comparison of KM curve of Gender")
##KM Mode of Delivery##
KMneomoded<-
survfit(Surv(Survival.Time..Days.,Survival.status)~mode.of.delivery,data=neonatal5)
KMneomoded
summary(KMneomoded)
plot(KMneomoded,conf.int=FALSE,col=c('blue','red'),xlab="Days",ylab="Proportion
Surviving")
```

```
legend('center',c("SVD","CS"),col=c('blue','red'),lty=1)
title("Comparison of KM curve of Mode of delivey")
##KM Place of Residence##
KMneoplaceR<-
survfit(Surv(Survival.Time..Days.,Survival.status)~Place.of.residence,data=neonatal5)
KMneoplaceR
summary(KMneoplaceR)
plot(KMneoplaceR,conf.int=FALSE,col=c('blue','red'),xlab="Days",ylab="Proportion
Surviving")
legend('center',c("Urban","Rural"),col=c('blue','red'),lty=1)
title("Comparison of KM curve of Place of Residence")
##KM Weight of Neonates##
KMneoweightR<-
survfit(Surv(Survival.Time..Days.,Survival.status)~Birth.weight..Category.,data=neonatal5)
KMneoweightR
summary(KMneoweightR)
plot(KMneoweightR,conf.int=FALSE,col=c('blue','red','green'),xlab="Days",ylab="Proportio
n Surviving")
legend('center',c("<1.5kg","1.5-2.5kg",">2.5kg"),col=c('blue','red','green'),lty=1)
title("Comparison of KM curve of Weight Groups")
##KM apgar score##
KMneoapgarR<-
survfit(Surv(Survival.Time..Days.,Survival.status)~Apgar.1.and.5..Category.,data=neonatal5)
KMneoapgarR
summary(KMneoapgarR)
plot(KMneoapgarR,conf.int=FALSE,col=c('blue','red'),xlab="Days",ylab="Proportion
Surviving")
legend('center',c("Less or equal to 6","Greater or equal to 7"),col=c('blue','red'),lty=1)
title("Comparison of KM curve of Apgar Score")
##KM Complications##
KMneocompR<-survfit(Surv(Survival.Time..Days.,Survival.status)~
Complications.of.deliveries,data=neonatal5)
KMneocompR
summary(KMneocompR)
```

```
plot(KMneocompR,conf.int=FALSE,col=c('blue','red'),xlab="Days",ylab="Proportion Surviving")
legend('center',c("Yes","No"),col=c('blue','red'),lty=1)
title("Comparison of KM curve of Complications")
##Log Rank for Categorical variables sex of child##
H<-with(neonatal5,Surv(Survival.Time..Days.,Survival.status==1))
logRanksex<-survdiff(H~factor(Sex.of.child),data=neonatal5,rho=0)
logRanksex
##Log Rank for Categorical variables Mode of delivery##
logRankmod<-survdiff(H~factor(mode.of.delivery),data=neonatal5,rho=0)
logRankmod
##Log Rank for Categorical variables place of residence##
logRankplaceR<-survdiff(H~factor(Place.of.residence),data=neonatal5,rho=0)
logRankplaceR
##Log Rank for Categorical variables Weight GR.##
logRankweightR<-survdiff(H~factor(Birth.weight..Category.),data=neonatal5,rho=0)
logRankweightR
##Log Rank for Categorical variables Apgar Score.##
logRankapgarR<-survdiff(H~factor(Apgar.1.and.5..Category.),data=neonatal5,rho=0)
logRankapgarR
```

Stata codes

PH modelling in Stata

```
# declare data as survival data.
stset time,(failure==1)scale(1)
#Models
```

1. Cox PH

```
stcox maternalage parity modeofdelivery apgarscore1 apgar5 gestationalageweeks placeofresidence sexofchild complicationsofdeliveries weightofbabykg
-AIC value
estat ic
#Testing for Proportionality Assumption
estat phtest, detail
```

2. Exponential PH

streg maternalage parity modeofdelivery apgarscore1 apgar5 gestationalageweeks placeofresidence sexofchild complicationsofdeliveries weightofbabykg, dist(exponential)

-AIC value

estat ic

3. Weibull PH

streg maternalage parity modeofdelivery apgarscore1 apgar5 gestationalageweeks placeofresidence sexofchild complicationsofdeliveries weightofbabykg, dist(weibull)

-AIC value

estat ic

4. Gompertz PH

streg maternalage parity modeofdelivery apgarscore1 apgar5 gestationalageweeks placeofresidence sexofchild complicationsofdeliveries weightofbabykg, dist(gompertz)

-AIC model

estat ic

5. Exponential AFT

streg maternalage parity modeofdelivery apgarscore1 apgar5 gestationalageweeks placeofresidence sexofchild complicationsofdeliveries weightofbabykg, dist(exponential) time

-AIC value

estat ic

6. Weibull AFT

streg maternalage parity modeofdelivery apgarscore1 apgar5 gestationalageweeks placeofresidence sexofchild complicationsofdeliveries weightofbabykg, dist(weibull) time

-AIC value

estat ic

7. Log-normal AFT

streg maternalage parity modeofdelivery apgarscore1 apgar5 gestationalageweeks placeofresidence sexofchild complicationsofdeliveries weightofbabykg, dist(lognormal)

-AIC value

estat ic

8. Log-logistic AFT

streg maternalage parity modeofdelivery apgarscore1 apgar5 gestationalageweeks placeofresidence sexofchild complicationsofdeliveries weightofbabykg, dist(loglogistic)

-AIC value

estat ic

9. Exponential frailty

streg maternalage parity modeofdelivery apgarscore1 apgar5 gestationalageweeks sexofchild complicationsofdeliveries weightofbabykg, dist(exponential) frailty(gamma)

-AIC value

estat ic

10. Weibull frailty

streg maternalage parity modeofdelivery apgarscore1 apgar5 gestationalageweeks sexofchild complicationsofdeliveries weightofbabykg, dist(weibull) frailty(gamma)

-AIC value

estat ic

11. Gompertz frailty

streg maternalage parity modeofdelivery apgarscore1 apgar5 gestationalageweeks sexofchild complicationsofdeliveries weightofbabykg, dist(gompertz) frailty(gamma)

-AIC value

estat ic

12. Lognormal frailty

streg maternalage parity modeofdelivery apgarscore1 apgar5 gestationalageweeks sexofchild complicationsofdeliveries weightofbabykg, dist(lognormal) frailty(gamma)

-AIC value

estat ic

13. Log-logistic frailty

streg maternalage parity modeofdelivery apgarscore1 apgar5 gestationalageweeks sexofchild complicationsofdeliveries weightofbabykg, dist(loglogistic) frailty(gamma)

-AIC value

estat ic

14. Cox frailty

stcox maternalage parity modeofdelivery apgarscore1 apgar5 gestationalageweeks sexofchild complicationsofdeliveries weightofbabykg, shared(localities)

-AIC value

estat ic