ADDIS ABABA UNIVERSITY

GRADUATE STUDIES PROGRAMME

DEPARTMENT OF STATISTICS

Survival analysis of infant mortality in Ethiopia

Abebu Abebaw

A Thesis Submitted To

The Department Of Statistics

Presented In Partial Fulfillment of the Requirements

For the Degree of Masters of Science in Statistics

Addis Ababa University

Addis Ababa, Ethiopia

June, 2013

Addis Ababa University

School of Graduate Studies

This is to certify that the thesis prepared by Abebu Abebaw, entitled: **Survival analysis of infant mortality in Ethiopia** and submitted in partial fulfillment of the requirements for the Degree of Master of Science in Statistics complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

Examiner Eshetu W. Signature _____ Date 28/06/2013

Examiner Mekonnen Tadesse Signature _____ Date 28/06/2013

Advisor M. K. Sharma Signature _____ Date _____

_____ Eshetu W.

Chair of Department or Graduate Program Coordinator

ABSTRACT

Survival analysis of infant mortality in Ethiopia

Abebu Abebaw

Addis Ababa University, 2013

Mortality is one of the components of population change. Infant and child mortality are among the best indicators of health development and socioeconomic status. Because a society's life expectancy at birth is determined by the survival chances of infants and children. That is why reduction of infant and child mortality is a worldwide target and one of the most important key indicators of the Millennium Development Goals (MDGs). Hence its indication is very important for evaluation and public health strategy.

The objective of this study was to examine the impact of maternal, socio economic and sanitation variables on infant and child mortality in Ethiopia and identify which of these factors had a pronounced impact for the reduction of infant and child mortality. The data in this study were obtained from the 2011 Ethiopia Demographic and Health survey (EDHS, 2011) conducted by the Central Statistical Agency (CSA). To analyze the data descriptive statistics, univariate and multivariable analyses were used. The descriptive analysis indicates that a death proportion is lower for females (20.6%) than for males (24.85%). Kaplan-Meier survival curves and log-Rank test were used to compare the survival experience of different groups. Cox's regression model was employed to identify the covariates that had a statistical significant effect on the survival time of infants. The estimation of the model parameters was done by partial maximum likelihood procedures. Mothers' educational level, birth order, sex and types of birth were identified as the risk factors for the death of infants. Furthermore it was found that the survival probabilities of infants with multiple birth, first birth order, non educated mothers and male children were low.

Acknowledgement

I would like to express my sincere appreciation to my advisor Prof.M.K Sharma for his valuable suggestions, comments and guidance in the completion of this work.

I would like to express my great debt which I owed to my family for providing financial, technical and moral support.

Furthermore, I wish to thank to Ato Workneh, Ato Tenaw Endalamaw, Ato Temesgen, Ato Bogale, Kalkidan, and their friends, who have provided me help and moral support in one way or another to accomplish this study.

Finally, I extend my gratitude to all staff of the Department of Statistics in Addis Ababa University for their assistance in various ways. I would like to thank the Ethiopian Central Statistics Agency for providing me with all the relevant secondary data used in this study and my sincere thanks also goes to all friends and colleagues who helped in my way to complete this work.

List of Acronym

CSA	Central Statistical Agency
EDHS	Ethiopian Demographic and Health Survey
HIV/AIDS	Human Immunodeficiency Virus/ Acquired Immunodeficiency Syndrome
MDGs	Millennium Development Goals
MOFED	Ministry of Finance and Economic Development
UNICEF	United Nations International Children Emergency Fund
USAID	United States Agency for International Development
WHO	World Health Organization

Table contents

List of Figures

List of Tables

List of Appendixes

CHAPTER ONE

1. INTRODUCTION

1.1. Background of the study

For a long time demographers have been interested in the study of mortality, which is one of the components of population change. Infant and child mortality are among the best indicators of health development and socioeconomic status. Because a society's life expectancy at birth is determined by the survival chances of infants and children. That is why reduction of infant and child mortality is a worldwide target and one of the most important key indicators of the Millennium Development Goals (MDGs). Hence its indication is very important for evaluation and public health strategy.

One of the most important targets of Millennium Development Goals (MDGs) that had been adopted in 2000 at the United Nations Millennium Summit was reducing infant and under-five child mortality rates by two-thirds from the 1990 levels by 2015. In 2000 the Ethiopian government announced the intention by signed the millennium declaration committing to achieve the Millennium Development Goals (MDGs) by 2015, many of which overlap with the 2015 national policy goals, which introduced by the federal government in 1991 and a policy action has been continued. For instance, in 2004 the Ethiopian government prepared child survival strategy and implementation plan to reduce child mortality of 140/1,000 live births to 67/1,000 live births by 2015, (FMOH, 2005).

Africa is the only continent that has seen rising numbers of deaths among infant and child since the 1970s [UNICEF, 2008]. Mortality rates among infant and children remain strikingly high throughout the majority of sub Saharan Africa (UNICEF, 2010). Sub-

Saharan Africa has achieved only around a 30 percent reduction in child mortality, less than half that is required to reach Millennium Development Goal Number 4 (MDG 4). Of the thirty countries with the world's highest child mortality rates, twenty-seven are in sub-Saharan Africa (UNICEF, 1999).The region's under-five mortality in 1998 was 173 per 1,000 live births (UNICEF, 2,000) compared to the minimum goal of 70/1,000 internationally adopted in the 1990 World Summit for Children. According to the UNICEF report (2010), in sub-Saharan Africa 1 in 8 children dies before his/her fifth birthday nearly 20 times the average for developed regions (1 in 167). Thus, infant and child mortality remains a big issue for these developing countries, especially as researchers attempt to distinguish what factors contribute to the high levels.

Ethiopia is the second largest country in Africa and the least developing country with high fertility and rapid population growth rates (MoFED, 2008). About 472,000 Ethiopian children die each year before their fifth birthdays (National Strategy for Child Survival in Ethiopia, 2005). This tragic fact places Ethiopia sixth among the countries of the world in terms of the absolute number of child deaths. Of every 100 children in Ethiopia, 14 will not live to celebrate their fifth birthday (Child Health in Ethiopia, 2004).

Infant and child mortality in Ethiopia had shown a continuous decline since 1960 with a more pronounced reduction in the recent decades. The trend of infant mortality rates has been about 200 per 1000 live births in 1960, 153 per 1000 live births in 1970, 110 per 1000 live births in 1984 (CSA, 1991, 1993), 97 per 1000 live birth in 2000 , 77 per 1000 live births in 2005 and 59 per 1000 live births in 2011 , this means that infant mortality declined by 20.6 percent between 2000 and 2005 (DHS, 2005) and declined by 39 percent between 2000 and 2011 (EDHS,2011) and under-five mortality rate is more than

200 per 1000 live births in 1960 and continued the reduction to 166 per 1000 live births in 2000, 123 per 1000 live births in 2005 and 88 per 1000 live births in 2011,this means that under five mortality declined by 47 percent between 2000 and 2011(EDHS,2011). In a great deal of this gains was achieved through the progressive implementation of Millennium Development Goals (MDGs) and Ethiopian health sector development program, has made a great strides to improve infant and child mortality. However, yet over all infant and under-five mortality rates remain very high; one in every 17 Ethiopian children did not survive to celebrate its first birth day and one in every 11 children died before its fifth birth day (EDHE, 2011). It was noticed that the decline of infant and child mortality had been achieved through the intervention of disease oriented programs. In recent decades the awareness of maternal, environmental, behavioral, sanitation and socioeconomic factors increased and recognized as additional important factors of child mortality. Understanding the current determinants of child mortality is essential to inform policies and strategies to accelerate the reduction of child mortality. Child mortality is often associated with poverty, maternal education, maternal fertility characteristics, maternal under-nutrition, intervals between births, access to adequate safe water and basic curative health services [MoFED, 2010].This study will consider most of the variables corresponding to the categories of these determinants and will attempt to identify the major factors that contribute to the death of infant and children in Ethiopia.

1.2. Statement of the problem

Infant and child mortality rate is a basic development indicator of health and socioeconomic status of a country, and also indicates life quality of a given population, as measured by life expectancy. The current levels of infant and child mortality in Ethiopia

are 59 per 1,000 live births and 31 per 1,000 live births respectively, which is higher than the target of the minimum Millennium Development Goals (MDGs). It is not well understandable why infant and child mortality rates remain high and far from desired in Ethiopia, despite the intervention made. In recent years, it has been established that, HIV/AIDS epidemic, poverty, economic crises, political unrest and civil war have a strong impact on the reduction of infant and child mortality. Several researchers investigated the regional variation of infant and child mortality. It was noticed that a decline of infant and child mortality was achieved through the intervention of disease oriented programs. In recent decades the awareness of maternal, environmental, behavioral and socioeconomic factors increased and these are recognized as additional important factors of infant and child mortality [MoFED, 2008].

However, infant and child mortality remains a major problem in Ethiopia still now. So that studying the determinants of infant and child mortality is crucial. Despite numerous interventions and action plans, very little evidence exists about why infant and child mortality rates in Ethiopia have not declined as desired. If Ethiopia is committed to achieving the MDG on child mortality, it is necessary to understand clearly the factors that are contributing to the high levels of mortality. This study therefore is an attempt to finds out the major contributing factors for infant and child mortality in Ethiopia.

1.3. Objective of the study

1.3.1. General objective

The general objectives of this study is to examine the impact of maternal, socio economic and sanitation variables on infant in Ethiopia and identify which of these factors have more pronounced impact for the reduction of infant and child mortality.

1.3.2. Specific objectives

- ✓ To identify determinants of infant mortality by using Cox proportional hazards regression.
- ✓ To estimate the survival time of infants.

1.4. Significance of the study.

This study has the purpose of identifying the major contributing factors for infant and child mortality in Ethiopia. So the result of this study

- ➢ Might provide information to government and other concerned bodies in setting policies, strategies and further investigation for reduction of infant mortality.
- ➢ Could provide base-line data for detail and further studies in the future.

1.5. Limitation of the study

The study has different limitations the major limitation of the study goes with the problems related to the use of secondary data. The study is conducted based on secondary data which might have incomplete and biased information.

Generating accurate estimates of infant mortality poses a considerable challenge because of the limited availability of high-quality data. Vital registration systems are the preferred source of data on infant mortality because they collect information as events occur (minimize recall errors from women) and cover the entire population. However, there are lack of vital registration systems that accurately record all births and deaths.

CHAPTER TWO

2. LITERATURE REVIEW

The literature selected and discussed in this section are those that are more related and relevant to this study.

Kombo and Ginneken (2009) using the result of 2005-06 Zimbabwean DHS investigate the maternal, socioeconomic and sanitation factors on infant and child mortality by using Cox regression model. They found an evidence of birth order (6+) with short preceding interval significantly associated with high risk of infant and child mortality. Multiple births tend to increase infant and child mortality. On the other hand socioeconomic determinants are rather small and insignificant effect on infant and child mortality. They suggest that the influence of birth order, preceding birth intervals, maternal age, type of birth and sanitation factors are more pronounced on infant mortality while weak effect on child mortality. The association of maternal, socioeconomic and sanitation factors with infant and child mortality weak as compared to 1994 and 1999 DHS surveys, show that those determinants are highly correlated and significant impact on infant and child mortality.

Wang (2003), using the results from the 2000 Ethiopia DHS examines the environmental determinants of child mortality. She used three hazard models, the Weibull, the Piece-wise Weibull and the Cox model to examine three age-specific mortality rates: neonatal (under one month), infant (under one year), and under-five mortality by location (urban/rural), female education attainment, religion affiliation, income quintile, and access to basic environmental services (water, sanitation and electricity). The study

showed that children born in rural areas face much higher mortality risk compared with those born in urban areas.

Mahfouz et al (2009) used primary data to estimate the levels of infant and under-five mortality and to determine the socioeconomic, demographic and environmental factors contributing to infants and child mortality in Malakal town, southern Sudan. It was found that child interval, child immunization, family size, family income, and mother's education, have significant influence on infant and under-five mortality

Klaauw and Wang (2003) developed a flexible parametric framework for analyzing infant and child mortality. This framework is based on widely used hazard rate models, which extend with two features. First, the model allows individual characteristics and household's socio-economic and environmental characteristics to have different impacts on infant and child mortality at different ages. Second, they allow for frailty at multiple levels, which can be correlated with each other. The first feature seems to be particularly relevant in describing infant and child mortality. Child specific and household's socio-economic and environmental characteristics have significantly different impacts on mortality rates at different ages of the child. They also used the estimated model to perform a number of policy experiments. The policy experiments show that, infant and child mortality rates can be reduced substantially by improving the household's socio-economic and environmental characteristics. Their model predicts that a significant number of under 5 years deaths can be averted by providing access to electricity, improving the education of women, providing sanitation facilities and reducing indoor air pollution. In particular, reducing indoor air pollution and increasing the educational level of women have substantial impacts on child mortality.

Balk et al (2003) used data from DHS for 12 countries in West Africa. It was shown that the determinants that influence child survival are mother's age at birth and the birth order of the child. These maternal factors have differential impacts on infants and children: infant deaths among mothers under age 20 typically occur in early infancy; young motherhood had less impact for children age 1–4 years. Birth order is closely related to mother's age at birth. First births are less likely to survive infancy than higher order births. The impact of birth order on survival is greatly reduced for children age 1–4 years. Multiple births are associated with much higher risk of death, especially during infancy. Maternal education has been observed to have a strong impact on child survival. Infants and children of mothers with no education both have only an 89 percent chance of survival at 12 months and at 59 months. Infants and children born to mothers with secondary or higher education have greatly improved chances of surviving, 95 percent and 97 percent, respectively. Infants and children residing in urban areas have, on average, better survival chances than those in residing in rural areas.

Desta (2011) using data from 2000 and 2005 EDHS employed logistic regression analysis to examine the socioeconomic, demographic and biological factors of infant and child mortality in Ethiopia. The study showed that marital status, birth order, type of births and preceding birth intervals are a significant proximate determinants of infant and child mortality. Breast feeding had an important significant effect on infant mortality but not on child mortality. Children born to women not currently married, first born children, multiple birth, children born within 18 months of the previous birth and children who were breastfed less than 6 months were exposed to the high risk of infant and child mortality.

Uddin (2009) used data from Bangladesh DHS. Cross-tabulation and multiple logistic regression techniques were used to estimate the predictors of child mortality. The cross-tabulation analysis shows that parent's education is the vital factor associated with child mortality risk but in logistic regression analysis only the father's education was found significant to reducing child mortality. The occupation of father was found a significant characteristic in both analyses. Breast feeding status and birth order have substantial impact on child mortality.

According to Caldwell (1979) infant and child mortality are highly associated with mother's education that increases the awareness of how to care her children before birth and after birth and enables her to change feeding and child care practices by shaping and modifying the traditional familial relationships. Education plays an important role to improve knowledge of medical and health care, particularly mother's education enhance to improve more effective preventative and health care practice.

Aguirre (1995) identified that the mothers 'education is the most important factor that directly affects child mortality. He used full and partial hazard in the Cox proportional model specification and found that there is a strong association between the instantaneous risk of dying and education in the face of other controls.

Using data from the first round of Demographic and Health Surveys for 22 developing countries, Desai and Alva (1998) examined the effect of maternal education on three markers of child health: infant mortality, children height for- age, and immunization status and found that there was a consistent negative relationship between maternal education and the probability of infant death. Children of mothers who attended primary

school are less likely to die than are children of mothers with no education. Children of mothers with a secondary-school education are the least likely to experience infant deaths. Among the 22 countries, this effect is statistically significant in 11 countries for primary education and in 10 countries for secondary education. The education variables are jointly significant in 14 countries.

Ezra and Gurum (2002) employed a logistic regression model to investigate the impact of birth interval on infant and child mortality in the context of communities characterized by high reproduction, prolonged breast feeding practice and poor living conditions in Ethiopia. They found that a short birth interval (<18 months) is significantly associated with infant and child mortality as compared to longer birth intervals (>24 months), implying the influence of short birth interval are more pronounced on infant mortality but weaker impact on child mortality. They observed that mother's in the age groups 15-19 and 35-49 have a significant effect on infant and child mortality as compared to with children born to mothers in the age category 25-34. Education is also a significant determinant of infant and child mortality.

Kumar and Gemechis (2010) uses data from Ethiopia DHS survey (2005) and employed cross tabulation technique to examine the selected socioeconomic, bio-demographic and maternal health care factors that determine child mortality in Ethiopia. The result showed that among socioeconomic variables birth interval with preceding birth and mother's education had significant impact to lowering the risk of child mortality. The result conformed that the child mortality risk associated with children of less than 2 years of birth interval with previous child was highest (15 percent) and lowest (4. 2 percent) for the children whose birth interval was 4+ years. On the other hand, they also reported that

mother's educational levels are significantly correlated to the low risk of child mortality relative to children born from illiterate mothers and fathers with primary educational level. Birth order and place of residence also an important determinates of child mortality in Ethiopia.

Manda (1999) used data from the 1992 DHS in Malawi to study the relationship between infant and child mortality and birth interval, maternal age at birth and, birth order, with and without controlling for other relevant explanatory variables. He also investigated the direct and indirect (through its relationship with birth intervals) effects of breastfeeding on childhood mortality. The study employed proportional hazards models. The study found that birth interval and maternal age effects are limited to the period of infancy.

Baker (1999) applying the Brass indirect estimation of the level of child mortality by using the data that was gathered by the Malawi Diffusion and Ideological change project (1998) from the three administrative region of Malawi: the north, canter and south to examine the pattern of regional variation of child mortality and selected maternal, socioeconomic and environmental factors. He found that the significant variation of child mortality between north and canter, between north and south but not between south and canter. Educational variations between those regions contributed for this regional variation of infant and child mortality. However, education associated with high child mortality variation if health service not readily available. On the other hand from the analysis sanitation and wealth index unexpectedly not contribute for the regional variation of child mortality in Malawi. However, the later result indicated that source of drink water and sanitation facilities highly correlated with the reduction of infant deaths.

Mturi and Curtis (1995) used data from 1991/92 DHS in Tanzania to study the determinants of infant and child mortality by using hazard model found that short birth interval, adolescent pregnancy and previous child mortality associated with increased risk of infant and child mortality while no significant effect of socioeconomic status (i.e. maternal education, partner's education, urban/rural residence and presence of radio in the household) of the population on infant and child mortality. They conclude that demographic and biological factors such as short birth interval (less than 2 years), teenage pregnancies (<20 years) and previous child death were all have an impact on infant and child mortality and socioeconomic mortality differential are not significant.

Similarly in Kenya, Mustafa and Odimegwu (2008) using 2003 DHS data set for children by using logistic regression models. They examined socioeconomic determinants of infant mortality rate both urban and rural setting. They found similar result like in the case of Tanzania that the regional variation exist in infant and child mortality between the differences provinces of Kenya. Most of the socioeconomic factors are not associated with the risk of infant and child mortality while children born in the richest household has lower probability of infant mortality relative to children born in the poorest households. However ethnicity and breast feeding in both urban and rural areas have a significant influence on infant mortality and sex of the child in urban areas and birth order and birth interval in rural areas are important determinants for the risk of infant mortality. Although they found that the incidences of HIV/AIDS in both urban rural areas increase the risk of dying at infancy period.

On the other hand it was observed that in Zimbabwe infant and child mortality differentials exist between urban-rural residence due to regional differences in health

infrastructure, and communication and disease prevalence conditions and also sanitation problem and low pipe water (poor of safe drink water) also highly affect infant and child mortality in Zimbabwe (Zimbabwe Central Statistical Office/ Macro International Inc, 2007).

CHAPTER THREE

3. MATERIALS AND METHODS

3.1. Sources of data

The data in this study were taken from the 2011 Ethiopia Demographic and Health survey (EDHS, 2011) conducted in Ethiopia as part of the worldwide demographic and health survey project. The 2011 Ethiopia Demographic and Health Survey were conducted by the Central Statistical Agency (CSA) with the support of the Ministry of Health. This is the third Demographic and Health Survey (DHS) conducted in Ethiopia, under the worldwide MEASURE DHS project, a USAID-funded project providing support and technical assistance in the implementation of population and health surveys in countries worldwide.

In this survey approximately 18,500 households were selected, 16,515 women aged 15-49 and 14,110 men aged 15-59 were interviewed. The primary objectives of the 2011 EDHS are to provide up-to-date information for planning, policy formulation, monitoring, and evaluation of population and health programs in the country. The survey was intentionally planned to be responded at the beginning of the last term of the MDG reporting period to provide data for the assessment of the Millennium Development Goals (MDGs). The data relating to family planning, fertility levels and determinants, fertility preferences, infant, child, adult and maternal mortality, maternal and child health, nutrition, women's empowerment, and knowledge of HIV/AID were collected for the nine regional states and two city administrations.

Information on infant mortality is found from the birth history of women who were included in the survey. The focus of this study was about infant (from birth to the age of 12 months).

3.2. Variables in the Study

3.2.1. The Response Variable

The response or outcome variable in this study is duration of survival (survival time of infants), measured in months.

Since there are censored observations the coding of the status variable is 1 for uncensored (event) observation and 0 for censored observations (for those who died after celebrating their first birthday or those who are alive at the time of the survey).

3.2.2. Predictor Variables

The predictor variables in survival data analysis are called covariates. These covariates (explanatory variables) in this study are classified into three groups: maternal, socioeconomic and sanitation variables.

Maternal (and related) factors:

- Infants birth order
- Maternal age
- infants sex
- Type of birth

Socioeconomic variables:

- Maternal education
- Paternal education

- Wealth index
- Region
- Religion
- Area of residence

Sanitation variables:

- Source of drinking water
- Types of toilet facility

Table 3.1: Detailed description of socioeconomic, maternal (bio-demographic) and sanitation Variables are presented as follows

NO.	Description and Name	Categories
1	Place of residence(RESIDENC)	(0). Urban (1). Rural
2	Region/Administrative city(REGION)	(1). Tigray (2). Affar (3). Amhara (4). Oromiya (5). Somali (6). Ben-Gumuz (7). SNNP (8). Gambela (9). Harari (10). Dire-dawa (11). Addis Ababa
3	Mothers education level (EDUCMOTH)	(0). no education (1). primary (2). secondary and higher
4	Fathers education level	(0). no education (1). primary (2). secondary and higher
5	Source of drinking water (WATER)	(0).piped water (1). otherwise
6	Wealth index (Economic status)	(0). Poor (1). Medium (2). Rich

7	Religion (RELIGION)	(1). Orthodox (2). catholic (3). Protestant (4).Muslim (5).others
8	Availability of toilet facility (TOILET)	(1). Flush toilet (2). otherwise
9	Infants birth order	(1). First births (2). 2-4 (3). 5+
10	Maternal age	(0). <20 years (1). 20-29 years (2).>= 30 years
11	Child's sex	(1). Male (2). female
12	Type of birth	(0). single (1). multiple

3.3. Methodology

3.3.1. Survival Analysis

Survival analysis is a collection of statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs. By time, we mean years, months, weeks, or days from the beginning of follow-up of an individual until an event occurs; alternatively, time can refer to the age of an individual when an event occurs. By event, we mean death, disease incidence, relapse from remission, recovery (e.g., return to work) or any designated experience of interest that may happen to an individual.

The term survival analysis applies to techniques in which the data being analyzed represent the time it takes for a certain event to occur. Survival analysis is the most important method when there is no time-to-event record.

In reality such situation can occur due to the following reasons:

• When an individual survive beyond the study period or the individual does not experience the event.

• Lost to follow-up, that is, an individual may drop out, transfer to other place, etc.

• Deaths due to other causes different from that/those specified in the study.

Therefore, survival data are almost always incomplete. The statistical terminology for such data is censoring. Censoring is common in survival analysis and it is considered as an important feature of survival data. Survival analysis is well suited to for such data which are very common in medical research since studies in medical areas have a special feature that follow-up studies could start at a certain observation time and could end before all experimental units had experienced an event. The most common encountered form of a censored observation is one in which observation begins at the defined time, say t=0, and terminates before the outcome of interest is observed. Since the incomplete nature of the observation occurs in the right tail of the time axis, such observations are said to be right censoring. The other mechanism that can lead to incomplete observation of time is truncation. A truncated observation is one which is incomplete due to a selection process inherent in the study design.

Several methods have been developed for the analysis of survival data. Some of these are:

• Descriptive statistics which include life tables, survival distribution, and Kaplan-Meier survival estimation which is used for the estimation of the distribution of survival time from a sample.

• Nonparametric tests are available for comparing the survival experience between two or more groups. The most common and widely used of these tests are the log-rank test, Generalized Wilcoxon test and Peto-Prentice test.

• The multivariable Method uses Cox-proportional hazards model. It is considered as the most interesting survival modeling in the interest of examining the relationship between survival and one or more predictors.

3.3.2. Descriptive Methods for Survival data

This method is especially important if individuals are homogeneous at least within groups. In such situation it is appropriate to use the Kaplan-Meier survival estimator. An initial step in the analysis of a set of survival data is to present numerical or graphical summaries of the survival times in a particular group. Usual applications of standard measures of central tendency and variability will not yield estimates of the desired parameters when the data include censored observations. In summarizing survival data, the two common functions of applied are the survivor function and the hazard function (Hosmer and Lemeshow, 1999).

3.3.2.1. Survivor function S (t)

The survivor function is defined to be the probability that the survival time of a randomly selected subject is greater than or equal to some specified time. Thus, it gives the probability that an individual surviving beyond a specified time. Moreover, the distribution of survival time is characterized by three functions: the survivorship function, the probability density function, and the hazard function.

Let T be a random variable associated with the survival times, t be the specified value of the random variable T and $f(t)$ be the underlying probability density function of the survival time T.

The cumulative distribution function $F(t)$, which represents the probability that a subject selected at random will have a survival time less than some stated value t, is given

$$F(t) = P(T<t) = \int_{-\infty}^{t} f(u)du, \; t \geq 0$$

By using this equation the survivor function, $S(t)$, can be given as

$$S(t) = P(T \geq t) = 1 - F(t), \; t \geq 0$$

From the above two equations the relationship between $f(t)$ and $S(t)$ can be derived as

$$f(t) = \frac{d}{dt}F(t) = \frac{d}{dt}\big(1 - S(t)\big) = \frac{-d}{dt}S(t), t \geq 0$$

Theoretically, as t ranges from 0 to infinity, the survivor function can be graphed as a smooth curve.

Survivor functions have the characteristics that:

- they are non-increasing
- at time $t = 0$, $S(t) = S(0) = 1$; that is, at the start of the study, since no one has experienced the event yet, the probability of surviving past time 0 is one and
- at time $t \to \infty$, $S(t) = S(\infty) \to 0$; that is, theoretically, if the study period increased without limit, eventually nobody would survive, so the survivor curve must eventually converge to zero.

3.3.2.2. Hazard function h (t)

The hazard function, denoted by $h(t)$, gives the instantaneous potential per unit time for the event to occur given that the individual has survived up to time t.

In contrast to the survivor function, which focuses on not failing, the hazard function focuses on failing; in other words, the higher the average hazard, the worse the impact on survival. The hazard is a rate, rather than a probability. Thus, the values of the hazard functions range between zero and infinity.

It is also known as the conditional failure rate in reliability, the force of mortality in demography, the intensity function in stochastic process, the age specific failure rate in epidemiology, the inverse of the Mill's ratio in economics or simply the hazard rate. Thus, in some sense, the hazard function can be considered as giving the opposite side of the information given by the survivor function.

The hazard function $h(t) \geq 0$, is given as:

$$h(t) = \lim_{\Delta t} \rightarrow 0 \left\{ \frac{p(t \leq T \leq \Delta t / T > t)}{\Delta t} \right\}$$

From this definition the relationship between the survivor and hazard function, can be expressed as:

$$h(t) = f(t) / S(t) = \frac{-d\{\log S(t)\}}{dt}$$

Where, f(t) is the probability density function of T.

3.3.2.3. Estimation of the survivor function

The survival and hazard functions are estimated using the Kaplan-Meier method as a preliminary analysis. This method is non-parametric or distribution-free, since it does not require specific assumption to be made about the underlying distribution of the survival times.

To apply the Kaplan-Meier method supposes that there are n independent individuals in a random sample with observed survival time's t_1, t_2..., t_n. The distinct ordered failure times observed among the n individuals are $t_{(1)}$, $t_{(2)}$, ..., $t_{(r)}$, r<n as there are more than one individual with the same observed survival time and some of the observations may be right-censored, i.e., the survival status of the individual might not be known at the time of the analysis. The probability of survival at time $t_{(j)}$, P ($t_{(j)}$) is then estimated by

$$P(t_{(j)}) = (n_j - d_j)/n_j$$

Where n_j is the number of individuals who are alive just before time $t_{(j)}$ and d_j is the number who die at this time.

Consequently the estimated probability of surviving beyond $t_{(j)}$, $S(t) = P(T \geq t)$ is defined as:

$$\hat{s}(t) = \prod_{j/t(j)\leq t} \frac{n_j - d_j}{n_j}$$

with the convention that $\hat{s}(t) = 1$ for $t < t(1)$.

The variance of the KM survival estimator which is also known as the Greenwood's formula is

$$\text{var}(\hat{s}(t)) = (\hat{s}(t))^2 \sum \frac{d_i}{n_i(n_i - d_i)}$$

With the approximated standard error given by:

$$\text{s.e}\{\widehat{s(t)}\} = \hat{\text{s}}(\text{t}) \left\{ \sum_{\text{j}=1}^{\text{k}} \frac{\text{d}_\text{i}}{\text{n}_\text{j}(\text{n}_\text{j} - \text{d}_\text{j})} \right\}^{1/2}$$

3.3.2.4. Comparison of Survivorship Functions

When comparing groups of subjects, it is always a good idea to begin with a graphical display of the data in each group. The figure in general shows if the pattern of one survivorship function lying above another which means the group defined by the upper curve lived longer, or had a more favorable survival experience, than the group defined by the lower curve. Now the statistical question is whether the observed difference seen in the figure is significant. The general form of this test statistic is given by

$$Q = \frac{[\sum_{i=1}^{m} w_i(d_{1i} - \hat{e}_{1i})]^2}{\sum_{i=1}^{m} w_i{}^2 \hat{v}_{1i}} \qquad \hat{e}_{1i} = \frac{n_{1i}d_i}{n_i} \qquad \text{And} \quad = \hat{v}_{1i} = \frac{n_{oi}n_{1i}(n_i - d_i)}{n_i{}^2(n_i - 1)}$$

Where

m is the number of rank-ordered failure (death) times

n_{oi} is the number at risk at obeserved survival time $t_{(i)}$ in group 0

n_{1i} is the number at risk at obeserved survival time $t_{(i)}$ in group 1

d_{0i} is the number of obeserved deaths in group 0

d_{1i} is the number of obeserved deaths in group 1

n_i is the total number of indivituals or risk before time $t_{(i)}$

d_i is the total number of deaths at $t_{(i)}$

w_i is the weight for censor adjustment at failure time t(i) .

The contribution to the test statistic depends on which of the various tests is used, but each may be expressed in the form of a ratio of weighted sums over the observed survival times.

Under the null hypothesis that the two survivorship functions are the same, and assuming that the censoring experience is independent of group, and that the total number of observed events and the sum of the expected number of events is large, Q follows a chi-square distribution with one degree of freedom. We can also use the above test to compare k groups.

In this study we use the log rank test which is special cases of Q.

The Log rank test

The log rank test, sometimes called the Cox-Mantel test, is the most well known and widely used test statistic. This test is based on weights equal to one, i.e. $=w_i = 1$. Therefore, the log rank test statistic becomes

$$Q_{LR}= \frac{[\sum_{i=1}^{m}(d_{1i}-\hat{e}_{1i})]^2}{\sum_{i=1}^{m}\hat{v}_{1i}}$$

3.3.3. Regression Models for Survival Data

In most medical studies which give rise to survival data, the relationship between survival experience of individuals and various explanatory variables have to be investigated.

In the analysis of survival data, interest centers on the risk of hazard of failure at any time after the time origin of the study. As a consequence, the hazard function is modeled directly in survival analysis. There are two broad reasons to model survival data. One objective of the modeling process is to determine which combinations of potential explanatory variables affect the form of the hazard function. Another reason for modeling the hazard function is to obtain an estimate of the hazard function itself for an individual from a set of explanatory variables

A variety of models and methods have been developed for doing this sort of survival analysis using either parametric or semi-parametric approaches. Cox proportional hazard model (Cox, 1972) is one of the most popular types of regression models used in survival analysis.

Cox regression model is some what different in form from linear models encountered in regression analysis and in the analysis of data from designed experiments, where the dependence of the mean response, or some function of it, on certain explanatory variables is modeled. However, many of the principles and procedures used in linear modeling carry over to the modeling of survival data. The Cox regression model can be used for data that contain censored observations. The model also takes into account the fact that the probability of experiencing an event differs with duration of exposure to risk.

3.3.3.1. Cox-proportional hazard model

The Kaplan-Meier and log-rank methods described are useful in the analysis of a single sample of survival data, or in the comparison of two or more groups of survival times. However, the relationship between the outcome variable and the explanatory variables is

identified by fitting a regression model. The basic model to be considered here is the proportional hazards model.

The assumption of proportional hazards is that the hazard of death at any given time for an individual in one group is proportional to the hazard at that time for an individual in the other group. When there are covariates in the analysis which are time dependent, this assumption may not hold.

The Cox Proportional Hazard Model is a multiple regression method used to evaluate the effect of multiple covariates on the survival. Cox (1972) proposed a semi-parametric model for the hazard function that allows the addition of covariates, while keeping the baseline hazards unspecified and can take only positive values. With this parameterization the Cox hazard function is

$$\lambda(t,\mathbf{x},\boldsymbol{\beta}) = \lambda_0(t)e^{\boldsymbol{\beta}'\mathbf{x}}$$

Where;

$\lambda_o(t)$ is the baseline hazard function that characterizes how the hazard function changes as a function of survival time,

$\lambda(t,\mathbf{x},\boldsymbol{\beta})$ represents the hazard function at time t with covariates $\boldsymbol{X} = (x_1, \ldots, x_p)'$.

$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is column vector of p regression parameters,

$e^{\beta'X}$ Characterizes how the hazard function changes as function of subject covariates,

t is the failure time.

The survival time of each member of the sample is assumed to follow its own hazard function. In such a case, the above model can equivalently be written as:

$$\lambda i(t,\mathbf{X_i},\boldsymbol{\beta}) = \lambda_o(t)\exp(\beta_1 X_{i1}+\ldots+\beta_p X_{ip})$$

i = 1,…,n where n is total number of observation in the study.

$\mathbf{X_i} = (x_{i1},\ldots,x_{ip})$ is a column vector of measured covariates for the i^{th} individual which are expected to affect the survival probability.

The proportional hazards estimation method computes a coefficient for each predictor variable that indicates the direction and degree of flexing that the predictor has on survival.

The proportional hazard model is the most popular regression method for analysis of censored survival data. The popularity is because:

- ✓ It allows flexible choice of covariates (we can accommodate time varying, time independent, continuous and discrete covariates).
- ✓ It is fairly easy to fit.
- ✓ Standard software packages are programmed to handle proportional hazards model such as SPSS, SAS, STATA, etc.
- ✓ Does not make any assumption about the underlying survival distribution (does not require the knowledge of the shape of the survival distribution).
- ✓ Does not require estimation of the baseline hazards rate, $\lambda_0(t)$, to estimate the regression parameters.

The Cox proportional hazard model is formulated as the hazard function which measures the risk to death or rate of failure at time t.

3.3.3.2. Assumption of Cox proportional hazard model

(1) The baseline hazard, $\lambda_o(t)$ depends on t, but not on covariates $x_1,\ldots,x_P$.

(2) The hazard ratio, i.e.,$e^{\beta' X}$, depends on the covariates $X = (X_1,\ldots,X_P)$,but not on time t.

(3) The covariate x_i does not depend on time t.

Assumption (2) is what led us to call this a proportional hazards model. To express this mathematically, consider two distinct values of the covariates X, say, x_1 and x_2.

$$\lambda(t,X,\beta) = \lambda_o(t)e^{\beta' X}$$

Then, the hazard ratio becomes:

$$\frac{\lambda(t,x_1,\beta)}{\lambda(t,x_2,\beta)} = \frac{\lambda o(t)e^{\beta X_1}}{\lambda o(t)e^{\beta X_2}}$$

$$= \frac{e^{\beta x_1}}{e^{\beta x_2}}$$

$$= e^{\beta'(x_1 - x_2)}$$

is independent of time t .

This shows that the ratio of the hazard functions for two individuals with different covariate values does not vary with time.

3.3.3.3. Estimation of Parameters in proportional hazard model

In Cox proportional hazards model we can estimate the vector of parameters **β** without having any assumptions about the baseline hazard $\lambda_o(t)$. As a consequence, this model is more flexible and an estimate of the parameters can be obtained easily.

Consider n independent individuals, the data that we need for the Cox proportional hazard model is represented by (t_i,δ_i,X_i) i = 1, 2,…, n,

Where t_i = the survival time for the i^{th} individual

δ_i = an indicator of censoring for the i^{th} individual given by 0 for censored and 1 for event/death.

$\mathbf{X_i}$ = a vector of covariates for individual i $(x_{i1}, x_{i2,\ldots,} x_{ip})$.

The full likelihood for right censored data can be constructed as

$$L(\beta) = \prod_{i=1}^{n} \lambda(t_i, x_i, \beta)^{\delta i} S(t_i, x_i, \beta)$$

Where $\lambda(t_i, x_i, \boldsymbol{\beta}) = \lambda o(t_i) e^{\boldsymbol{\beta}' \mathbf{X}}$ is the hazard function for individual i.

$S(t_i, x_i, \beta) = (So(ti))^{\exp(\boldsymbol{\beta}' \mathbf{x_i})}$ is the survival function for individual i.

It follows that $L(\beta) = \prod_{i=1}^{n} (\lambda o(t_i) \exp(\beta' X_i))^{\delta i} (So(t_i))^{\exp(\beta' X_i)}$

The full maximum likelihood estimator of β can be obtained by differentiating the right hand side of equation L (β) with respect to the components of β and the base line hazard $\lambda_o(t)$. This implies that unless we explicitly specify the base line hazard, λo(t), we cannot obtain the maximum likelihood estimators for the full likelihood.

To avoid the specification of the base line hazard, Cox (1972) proposed a partial likelihood approach that treats the baseline hazard as a nuisance parameter and removes it from the estimating equation

Partial likelihood

Instead of constructing a full likelihood, we consider the probability that an individual experiences an event at time t_i given that an event occurred at that time.

Let R_i denote the set of individuals at risk at time just prior to $t_{(i)}$. Assume that for the present case there is only one failure at time t_i, i.e., no ties. Then the probability that individual *i* with covariates X_i the one who experience the event at time $t_{(i)}$.

=P (individual i has experiences an event at time $t_{(i)}$ | one event at time $t_{(i)}$)

$$= \frac{\lambda(t,x_i)}{\sum_{j\varepsilon Rt(i)} \lambda(t,x_j)}$$

And under the proportional hazards assumption on using equation the ratio

$$\frac{\lambda 0(t)exp(\beta' x_i)}{\sum_{j\varepsilon Rt(i)} \lambda 0(t,x_j)}$$

Shows the contribution to the partial likelihood at each death time t (i) by the individuals with covariate X_i in the risk set $R_{t(i)}$. Where $R_{t(i)}$ is the overall subjects in the risk set at time t(i) .

By eliminating the base line hazards function, in the numerator and denominator, this equation becomes

$$\frac{exp(\beta' x_i)}{\sum_{j\varepsilon Rt(i)} exp(\beta' x_j)}$$

Thus the partial likelihood is the product over all failure time t(i) for $i = 1,2,...,m$ of the conditional probability of this equation to give the partial likelihood

$$\text{Lp } (\beta) = \prod_{i=1}^{m} \frac{\exp (\beta' x_i)}{\sum_{j\varepsilon Ri(i)} \exp (\beta \cdot x_j)}$$

We obtain the maximum partial likelihood estimator by differentiating the right hand side of this equation with respect to the component of β, setting the derivative equal to zero and solving for the unknown parameters.

The partial likelihood derived above is valid when there are no ties in the data set.

3.3.3.4. Interpretation of the coefficients of the Cox-regression model

The estimated coefficients for the predictor variables represent the slope or rate of change of a function of the outcome variable per unit of change in the predictor variable by keeping the remaining predictor variables fixed. Thus interpretation involves two issues, determining the functional relationship between the outcome variable and the covariate and appropriately defining the unit of change for the predictor variable (Hosmer-Lemeshow, 1999).

For example, for a dichotomous covariate with value 1 and 0, the hazard ratios of being in the Category of interest for the j^{th} subject, becomes $\frac{\lambda o(t)\exp(\hat{\beta}_i * 1)}{\lambda o(t)\exp(\hat{\beta}_{i*0})} = \exp(\hat{\beta}_i)$ fixing the other covariates constant. It is interpreted as the hazard rate, or rate of death in our case, among subjects with i^{th} covariate value equals 1 is $\exp(\hat{\beta} i)$ time higher than subjects with i^{th} covariate value equals zero, i = 1,2,..., p and j = 1,2,..., n . For covariates having L levels (L>2), similarly interpretations can made by taking one of the L-levels as a reference category.

3.3.4. Model development

In any applied setting, performing a proportional hazard regression analysis of survival data requires a number of critical decisions. It is likely that we will have data on more covariates than we can reasonably expect to include in the model, so we must decide on a method to select a subset of the total number of covariates. When selecting a subset of the covariates, we must consider such issues as clinical importance and statistical significance Hosmer and Lemeshow, 1999).

3.3.4.1. Selection of covariates

The methods available to select a subset of covariates to include in a proportional hazards regression model are essentially the same as those used in any other regression model. There are three methods of selection of influential covariates. These are purposeful selection, stepwise selection (forward selection and backward elimination) and best subset selection. Survival analysis using Cox regression method begins with a thorough univariable analysis of the association between survival time and all important covariates (Hosmer and Lemeshow, 1999).

Recommendable procedure in selecting variables in the study

Hosmer and Lemeshow (1999) and Collett (2003) recommended the following procedure in variable selection.

1. The first step is to fit models that contain each of the variables one at a time. The values of $-2\log\hat{l}$ for these models are then compared with that for the null model. The null model is a model to determine which variables on their own significantly reduce the value of this statistic.

   Include all variables that are significant in the univariable analysis at the 25 percent level and also any other variables which are presumed to be clinically important to fit the initial multivariable model.

2. The variables that appear to be important from step 1 are then fitted together in a multivariable model. In the presence of certain variables others may cease to be important. Consequently, backward elimination is used to omit non-significant

variables from the model. Once a variable has been dropped, the effect of omitting each of the remaining variables in turn should be examined.

3. Variables, that were not important on their own, and so were not under consideration in step 2, may become important in the presence of others. These variables are therefore added to the model from step 2, with forward selection method. This process may result in terms in the model determined at step 2 ceasing to be significant.
4. A final check is made to ensure that neither significant variable is eliminated from the model nor non-significant variable is included in the model. At this any of the main effects currently in the model can be considered for inclusion if the inclusion significantly modifies the model.

3.3.4.2. Testing for the form of linearity of covariates

The assumption of linearity can be checked by using the plot of martingale residuals. The plot of martingale residuals obtained from fitting the model, excluding the covariate whose functional form needs to be determined, against the excluded covariate display the functional form required for the covariate. If the resulting plot is random showing no systematic pattern this indicates that the covariate is linear in the model.

3.3.5. Assessment of Model Adequacy

The methods for assessment of a fitted proportional hazards model are essentially the same as for other regression models. In general requirements for model assessment are

1. Methods for testing the assumption of proportional hazards

2. subject-specific diagnostic statistics that extend the notations of leverage and influence to the proportional hazards model, and

3. Overall summary measures of goodness of fit.

3.3.5.1. Checking for proportionality assumption

In order to use the Cox model, it has to be checked that the assumption of whether the effects of covariates on hazard ratio remain constant over time. This is a vital assumption of proportional hazards model and must be assessed for each covariate. Several procedures of graphical techniques and tests are proposed to investigate the proportionality assumptions in fitting the Cox model (Cox, 1972). The Schoenfeld residuals are employed to assess this assumption.

The Schoenfeld residuals graphical technique can be used to assess Cox model proportionality assumption. The technique is based on individual contributions to the log partial likelihood and measures the difference between the covariate for the i^{th} individual and a weighted average of the covariate over the risk set at each event. To check the proportionality assumption for each covariate, we plot the scaled Schoenfeld residuals against log of survival time. If the proportional hazards assumption is satisfied, the distribution of residuals over time is random, that is, does not show a particular trend, and the smoothed plot called Locally Weighted polynomial regression (Lowess) line summarizing the residuals should be a straight line and close to the horizontal reference line. Otherwise, a plot of scaled Schoenfeld residuals for a given covariate may reveal a violation of the proportional hazards assumption (Schoenfeld, 1982).

Formal tests need to detect any time dependency in particular covariates, after allowing for the effects of explanatory variables that are known. Testing the dependency of covariates on time is equivalent to testing for a non-zero slope in a generalized linear regression of the scaled Schoenfeld residuals on functions of time. A non-zero slope is an indication of a violation of the proportional hazard assumption. The Grambsch-Therneau test of non-proportionality uses partial residuals for the test of proportional hazards assumption. In order to use this test for the i^{th} covariate, Grambsch and Therneau (1994) propose a time-varying coefficient as:

$$\beta_i(t) = \beta_i + \gamma_i g_i(t)$$

where $\beta_i(t)$ is time varying coefficient, β_i is constant, and $g_i(t)$ is some specified function of time, usually $g_i(t) = \ln(t)$;

Then, the Cox proportional hazard model for time varying coefficient with $g_i(t) = \ln(t)$ is defined as:

$h(t, x_i, \beta_i(t)) = h_0(t)\exp(\beta_i(t)x)$, by substituting for $\beta_i(t)$ and $g_i(t)$becomes.

$$= h_0(t)\exp(\beta_i + \gamma_i(lnt)x)$$

$$= h_0(t)\exp(\beta_i x + \gamma_i(lnt)X)$$

This equation is the proportional hazards model with the interaction term, $x \ln(t)$ and main effect x_i . To test the significance of the interaction term $x\ln(t)$, we perform the test:

$H_O: \gamma = 0$ versus $H_1: \gamma \neq 0$ and we use likelihood-based tests like Wald test. If $\gamma = 0$ is not rejected, β_i's are not time varying coefficients and hence the proportional hazards assumption is satisfied. If $\gamma = 0$ is rejected then the proportional hazards assumption is not satisfied, that leads to the need of other methods that cope with time-dependency (Schoenfeld, 1982).

3.3.5.2. Identification of influential and poorly fit subjects

Another important aspect of model evaluation is through diagnostic statistics in order to identify which subjects have an unusual configuration of covariates or observations that have influence on the estimates of the parameters or on the fit of the model. In other words a fitted model is particularly sensitive to one or more observations in the data set.

Such observations can be termed as influential observations. Conclusions from survival analyses are often framed in terms of estimates of the relative hazard, which depends on the estimated values of the coefficients in the Cox regression model. Thus, it is desirable to examine the influence of each observation on these estimates. The interest is about observations that influence estimate of hazard functions and the complete estimate of the model and identifications of these observations. This could be done by fitting the model to all n observations in the data set, and then fitting the same model to the sets of n -1 observations obtained by omitting each of the n observations in turn. The interest is to determine if the result would change when a particular observation is removed from the analysis (Collett, 2003)

Suppose that l_P (β) is log partial likelihood and $\hat{\beta}_j$ is the corresponding j^{th} parameter estimate of the model containing all the n observations and l_P (β_{-i}) be the log partial likelihood and $\hat{\beta}_{j(i)}$ is the j^{th} parameter estimate of the model containing only the $n-1$ observations after deleting the i^{th} observation, respectively. Then, the statistic;

$\Delta_i \hat{\beta}_j = \hat{\beta}_j - \hat{\beta}_j(-i)$, which is known as DFBETA, can be used as a measure of how the j^{th} parameter estimate would change if the i^{th} observation was deleted from the data set. On the other hand, the statistic, $LD_i = 2l_p(\beta) - 2l_p(\beta_{-i})$,which is called

the likelihood displacement statistic, can be used as a measure of how the maximized partial log likelihood changes if the i^{th} observation was deleted from the data set. Observations that influence a particular parameter estimate have a large absolute value of DFBETA than other observations in the data set. Observations that do influence the overall fit of the model are those which have large values of likelihood displacement statistics than the other observations in the data set (Collett, 2003).

3.3.5.3. Overall goodness of fit

If the fitted model is satisfactory (appropriate), the Cox-Snell residuals will behave as n observations from a unit exponential distribution. Thus, the plot of the estimated hazard rate of the Cox- Snell residuals ($\widehat{H}_r(r_i)$), versus r_i will give a straight line through the origin with slope unity if the fitted model is satisfactory. However, the drawback is that they do not indicate the particular departure from the model fitted, if there is any.

3.3.5.4. Residual analysis

Under the proportional hazards model, residuals play a central role in evaluating the model assessment and adequacy. Many model checking procedures are based on quantities known as residuals. Residuals are values that can be calculated for each observation and have the feature that their behavior is known, at least approximately, when the fitted model is satisfactory. The following residuals have been proposed for use.

Cox-Snell residuals: The Cox-Snell residual for the i^{th} individual is given by

$$r_{Ci} = \exp(\hat{\beta}' x_i)\widehat{H}_0(t_i) = \widehat{H}_0(t_i) = -\log\hat{S}_{i(t_i)}$$

Where $\widehat{H}_0$ (t_i) is an estimate of the baseline cumulative hazard function at time t_i, the observed survival time of that individual, $\widehat{H}_i$ (t_i) and $\hat{S}_{i(t_i)}$ are the estimated values of the cumulative hazard and survivor functions of the i^{th} individual at t_i.

Martingale residuals are modified Cox-Snell residuals and, defined as

$$r_{Mi} = C_i - r_i$$

Where, C_i is censoring indicator and r_i is the Cox-Snell residual.

It can be shown that these residuals sum to zero and, in large sample, the martingale residual are uncorrelated with one another and have an expected value of zero. In this respect, they have properties similar to those possessed by residuals encountered in linear regression analysis.

Schoenfeld residuals: Schoenfeld residuals are useful to check the proportionality of the covariates over time, that is, to check the validity of the proportional hazards assumption. If the model fits well then the residuals are randomly distributed without any systematic pattern around the zero line, the reference line.

The i^{th} schoenfeld residual for x_j, the j^{th} explanatory variable in the model, is given by

$$r_{pji} = c_i\{x_{ji} - \hat{a}_{ji}\}$$

Where x_{ji} is the value of j^{th} explanatory variable, j = 1, 2 ...p, for the i^{th} individual

$\hat{a}_{ji} = \frac{\sum_{L\varepsilon R(t_i)} x_{ji}\exp(\hat{\beta}' x_i)}{\sum_{L\varepsilon R(t_i)} \exp(\hat{\beta}' x_i)}$ and $R(t_i)$ is the set of all individuals at risk at time t_i.

CHAPTER FOUR

4. STATISTICAL DATA ANALYSIS AND DISCUSSION

4.1. Introduction

In this chapter we present results of data analysis, discussion and interpretation. The first part presents summary statistics of factors considered in this study. The second part presents descriptive survival analysis and compares the survival time in different groups. The third part is about fitting the model. Then, the adequacy of the model is investigated. Finally, the results are discussed and interpreted. The statistical packages STATA and SAS are employed to analyze the data.

4.2. Summary Statistics

The total number of live births in this study was 2878. The total number of deaths was 656(22.79%). A death proportion seems lower for females (20.6%) than for males (24.85%). The distribution of some explanatory variables over the total sample at risk is presented in table 4.1 below.

Table 4.1: distribution of maternal (demographic), socioeconomic and sanitation factors of infant mortality.

Summary of the number of event and censored values					
Maternal (and related) factors	Category	Total	Event/death	censored	Percent censored
Maternal age	(0). <20 years	1780	433	1347	75.67
	(1). 20-29 years	1054	213	841	79.79
	(2).>= 30 years	44	10	34	77.27
infants sex	(1). Male	1485	369	1116	75.15
	(2). Female	1393	287	1106	79.40

infants birth order	(1). First births (2). 2-4 (3). 5+	545 1591 742	140 344 172	405 1247 570	74.31 78.38 76.82
type of birth	(0). single (1). multiple	2747 131	566 90	2181 41	79.40 31.30

Socioeconomic Variables	Category	total	Event/death	censored	Percent censored
Maternal education	(0).no education (1). Primary (2).secondary and higher	1941 785 152	469 169 18	1472 616 134	75.84 78.47 88.16
paternal education	(0). no education (1). Primary (2).secondary and higher	1432 1095 351	354 239 63	1078 856 288	75.28 78.17 82.05
wealth index	(0). Poor (1). Medium (2). Rich	1461 486 931	348 108 200	1113 378 731	76.18 77.78 78.52
religion	(1). Orthodox (2). Catholic (3).protestant (4).Muslim (5).others	867 30 567 1362 52	209 11 127 296 13	658 19 440 1066 39	75.89 63.33 77.60 78.27 75.00
Place of residence	(0). Urban (1). Rural	462 2416	89 567	373 1849	80.74 76.53
region	(1). Tigray (2). Affar (3). Amhara (4). Oromiya (5). Somali (6). Ben-Gumuz (7). SNNP (8). Gambela (9). Harari (10). Dire-dawa (11). Addis Ababa	278 252 303 466 266 266 422 207 162 94 162	64 60 80 96 54 72 98 56 38 11 27	214 192 223 370 212 194 324 151 124 83 135	76.98 76.19 73.60 79.40 79.70 72.93 76.78 72.95 76.54 88.30 83.33

Sanitation variables:	Category	total	Event/death	censored	Percent censored
source of drinking water	(1).piped water (2). otherwise	189 2689	27 629	162 2060	85.71 76.61
Types of toilet facility	(1). Flush toilet (2). otherwise	71 2807	14 642	57 2165	80.28 77.13

4.3. Descriptive analysis

Before proceeding to more complicated models, we make a descriptive analysis that will use as initiation to our subsequent findings. Here we start with the test of whether the observed differences given in the data summary among different factors are statistically significant or not with the help of log-rank test and Kaplan-Meier survival estimates. The log rank test is performed to test if there are statistically significant differences among the survival experience of the different groups of each of the covariates at 5% level of significance. The null hypothesis to be tested is that there is no difference between the probabilities of an event occurring at any time point for each covariate. The SAS results have been summarized in Table 4.2 below.

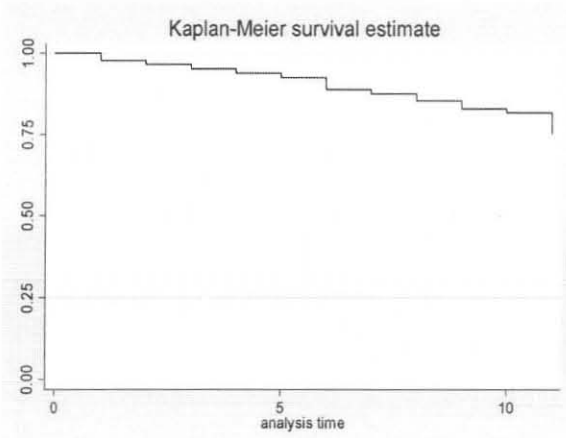
Table 4.2: Log rank test for equality of survival experience among the different groups of covariates

Test of equality over strata			
Variable	Chi-square	DF	Pr > chi-square
Region	21.0650	10	0.0206
Place of residence	5.0176	1	0.0251
Sores of drinking water	8.1096	1	0.0044
Types of toilet facility	0.3228	1	0.5699
Wealth index	2.1619	2	0.3393
Maternal age	4.5141	2	0.1047
Mothers educational level	12.1917	2	0.0023
infants birth order	4.9830	2	0.0828
Fathers educational level	7.5182	2	0.0233
Religion	4.4223	4	0.3519
Type of birth	194.2766	1	<.0001
Sex	8.2442	1	0.0041

Table 4.2 shows that there is a significant difference of survival experience among groups of the levels of sex, region, sources of drinking water, Place of residence, mothers educational level, fathers educational level and types of birth. On the other hand, there are statistically no significant differences in survival experienc among groups of the remaining categorical covariates Wealth index, types of toilet facility, maternal age, birth order and religion.

The log-Rank test results suggests that level of sex, region, sources of drinking water, Place of residence, mothers educational level, fathers educational level and types of birth are significant covariates whose different levels have an impact on the survival time of infants; while Wealth index, types of toilet facility, maternal age, childes birth order and religion are covariates whose different levels do not have an impact in the survival time of infants. Plots of different groups of categories are given bellow.

Figure 4.1: The plot of the overall estimate of Kaplan-Meier survivor function of infants in Ethiopia.



The Kaplan-Meier estimator survival curve gives the estimate of survivor function among different strata or groups of covariates to make comparisons. Separate graphs of the estimates of the Kaplan-Meier survivor functions are constructed for different categorical covariates. In general, the pattern that one survivorship function lying above another means the group defined by the upper curve has a longer survival than the group defined by the lower curve. From the graph there are clear differences among the various groups of level of sex, region, sources of drinking water, place of residence, mothers educational level, fathers educational level and types of birth. However, the difference is not clear among Wealth index, types of toilet facility, maternal age, childes birth order and religion. For detail see Appendix D (Fig1 (a-k)).

Figure 4.2: survival curves of infant by types of birth

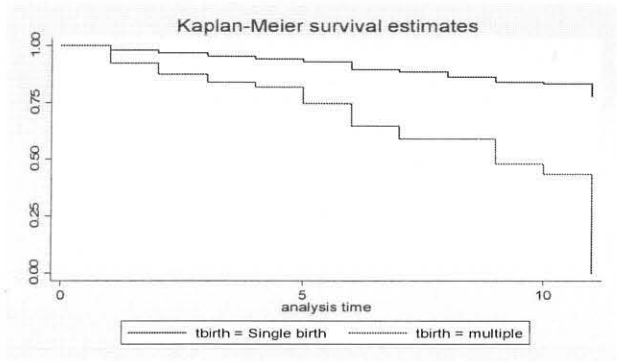


Figure 4.2 shows that there is a difference in survival functions between categories and further showed that; infant with single birth survive better than those with multiple births.

Figure 4.3: survival curves of infant by region

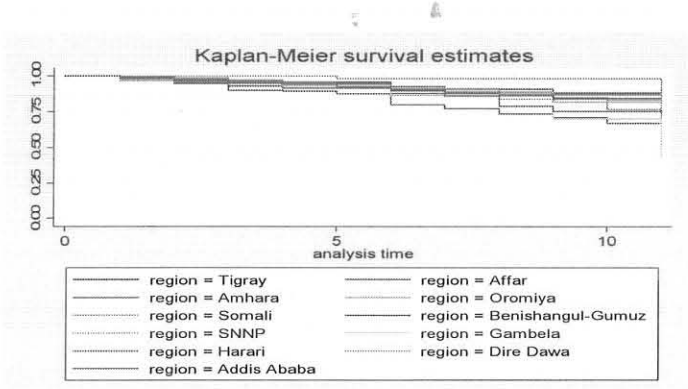


Figure 4.3supports that there is a difference in survival functions between categories. In general infant living in Dire- Dawa survive longer than the other 10 regions where as infants living in Benishangule- Gumuz have the lowest chance to survive.

Figure 4.4: survival curves of infant by source of drinking water

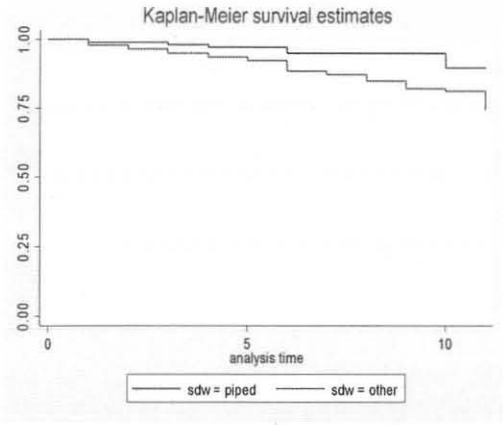


Figure 4.4 shows that there is a difference in survival functions between categories and further showed that; infant who used piped water survive longer than those who used other sources of water.

Figure 4.5: survival curves of infant by place of residence

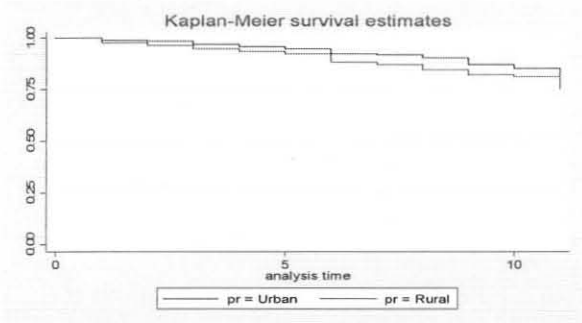


Figure 4.5 also shows that there is a difference in survival functions between categories and further showed that; infants who live in urban area have a better survival than infants who live in rural area.

Figure 4.6: survival curves of infant by mothers and fathers educational level

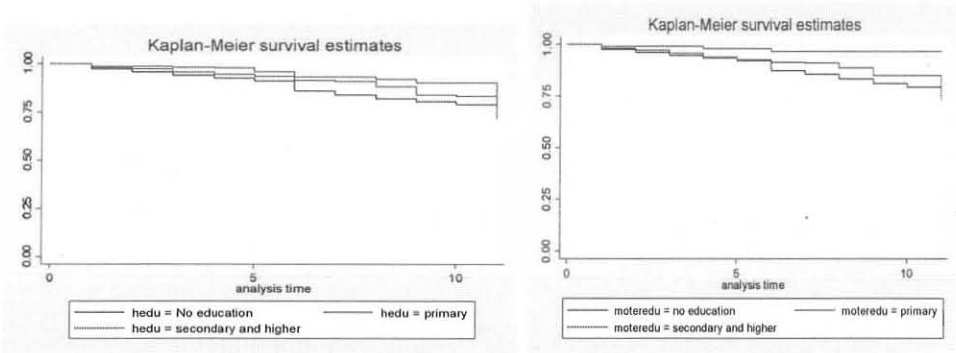


Figure 4.6 show's that there is a difference in survival functions between categories. Infants whose fathers and mothers educational level had secondary and above level of education survived longer than the other two where as infants whose families had no education had the lowest chance to survive. Infants born to mothers with secondary or higher education had greatly improved chances of surviving.

Figure 4.7: survival curves of infant by sex

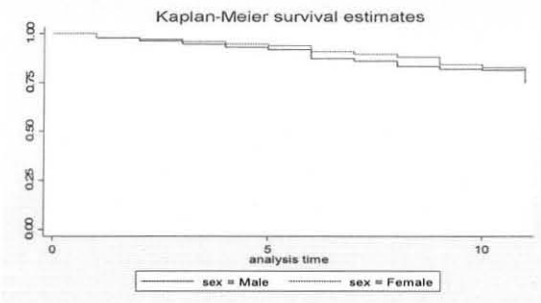


Figure 4.7 show that females have better survival than males.

4.4. Results of Cox-proportional hazards model

Assessing the relationship between the outcome and explanatory variables is necessary to develop the model.

Univariate Analysis

The first step in model development process is to select explanatory variables that have the potential to be included in the proportional hazards model.

The model will be constructed by first identifying factors which are significant at 25% level of significance in univariate Cox proportional hazard analysis. For each covariate we will use a univariate Cox proportional hazards model analysis that contains a single independent variable in order to have an idea about each covariate. Likelihood ratio chi-square test and wald test are used to test the significance of univariable relationship collett (2003).

Thus, using the Wald chi-square test, the predictors that are found to be significant, and will be considered for the next multivariable analysis at 25%level of significance are Place of residence, source of drinking water, wealth index, mothers educational level, fathers educational level, types of birth, birth order, maternal age and sex. Table 4.3 bellow summarizes the findings of univariable analysis.

Table 4.3: Univariable analysis result for each covariate

Variable	category	d. f.	Parameter Estimate	Standard Error	Chi-Square	Pr> Chi-Square	Hazard Ratio	-2 LOG L
Region	SNNP	1	0.33993	0.21738	2.4453	0.1179	1.405	10012.366
	Affar	1	0.42757	0.23179	3.4027	0.0651	1.534	
	Amhara	1	0.50794	0.22266	5.2043	0.0225	1.662	
	Benishangul Gumuz	1	0.54084	0.22573	5.7405	0.0166	1.717	
	Dire Dawa	1	-0.37720	0.35770	1.1120	0.2916	0.686	
	Gambela	1	0.56028	0.23435	5.7160	0.0168	1.751	
	Harari	1	0.35704	0.25171	2.0120	0.1561	1.429	
	Oromiya	1	0.24047	0.21786	1.2183	0.2697	1.272	
	Somali	1	0.23029	0.23572	0.9545	0.3286	1.259	
	Tigray	1	0.30945	0.22956	1.8171	0.1777	1.363	
Place of residence	Rural	1	0.24231	0.11404	4.5143	0.0336	1.274	10027.913
Source of drinking water		1	0.52642	0.19655	7.1732	0.0074	1.693	10024.241
Types of toilet facility		1	0.14573	0.27017	0.2909	0.5896	1.157	10032.399
Wealth index	Medium	1	-0.06540	0.11015	0.3524	0.5527	0.937	10030.739
	Rich	1	-0.12274	0.08875	1.9124	0.1667	0.884	
Maternal age	20-29	1	-0.16882	0.08374	4.0643	0.0438	0.845	10028.564
	>=30	1	-0.06652	0.31993	0.0432	0.8353	0.936	
Mothers educational level	Primary	1	-0.12504	0.08973	1.9420	0.1634	0.882	10019.810
	Secondary and above	1	-0.73800	0.24018	9.4417	0.0021	0.478	
Birth order	2-4	1	-0.21217	0.10027	4.4771	0.0344	0.809	10028.356
	5+	1	-0.14671	0.11386	1.6601	0.1976	0.864	
Fathers education level	Primary	1	-0.13461	0.08372	2.5848	0.1079	0.874	10025.692
	Secondary and above	1	-0.32451	0.13674	5.6323	0.0176	0.723	
Religion	protestant	1	0.09883	0.28588	0.1195	0.7296	1.104	10029.217
	Catholic	1	0.46681	0.30935	2.2771	0.1313	1.595	
	Muslim	1	-0.06981	0.11254	0.3848	0.5351	0.933	
	Other	1	-0.07671	0.09043	0.7196	0.3963	0.926	
Types of birth	Multiple birth	1	1.38547	0.11393	147.8950	<.0001	3.997	9926.173
Sex	Female	1	-0.21438	0.07872	7.4164	0.0065	0.807	10025.235

Table 4.4: Testing Global Null Hypothesis BETA=0

variable	DF	Likelihood ratio		wald	
		Chi-square	Pr>chisq	Chi-square	Pr>chisq
region	10	20.3377	0.0262	18.5655	0.3461
Place of residence	1	4.7908	0.0286	4.5143	0.0331
Source of drinking water	1	8.4629	0.0036	7.1732	0.0074
Types of toilet facility	1	0.3051	0.5807	0.2909	0.5896
Wealth index	2	1.9529	0.3766	1.9508	0.1770
Maternal age	2	4.1394	0.1262	4.0654	0.1310
Mothers educational level	2	12.8941	0.0016	10.6402	0.0049
Birth order	2	4.3482	0.1137	4.4793	0.1065
Fathers educational level	2	7.01223	0.03	6.7447	0.0343
religion	4	3.4864	0.4800	3.9175	0.4173
Types of birth	1	106.530	<.0001	147.8950	<.0001
sex	1	7.4689	0.0063	7.4164	0.0065

Multivariable Analyses

The most appropriate subset of covariates to be included in the multivariable model will be selected based on their contribution to reduce the value of -2log L of the null model considerably on the basis of p-values; otherwise they are not candidates for inclusion. For the null model the value of -2log L is 10032.704. From Tables 4.3 and 4.4 we observe that among the explanatory variables, types of birth leads to the highest reduction in the value for the null/empty model, from10032.704 to 9926.173, the difference is 106.531 and it is statistically significant (p-value <0.0001). The next highest change is obtained by region (20.338), followed by mothers educational level (12.894) and source of drinking water (8.463).

Proceeding in this manner covariates which increase -2LL ($\hat{\beta}$) will be eliminated. Thus, using the Wald chi-square test, the predictors that are found to be significant, and will be considered for the next multivariable analysis at 25%level of significance are Place of residence, maternal age, source of drinking water, wealth index, mothers educational level, fathers educational level, types of birth, birth order and sex . We proceed fitting the

initial multiple Cox proportional model by including these covariates that are significant at univariate analysis. Then, another multivariable Cox proportional regression will be fitted by eliminating those covariates that are not significant at 5% significance level. Following the same procedure until, we get covariates that are all significant at 5%level of significance [See Appendix A Tables: 1-6].

The next step is assessing the importance of the variables which were not significant in the univariable analysis as predictors or useful confounder for survival time of infant and child and their effects. The effect of those variables not significant in multivariable analysis is also investigated. These variables are added one at a time in the model containing the variables birth order, sex, types of birth and mother's educational level which are significant at 5% significance level.

The result of the analysis reveals that, none of those variables was found to be significant and therefore they cannot be retained in the model. Thus, variables that were neither significant at univariable analysis nor at multivariable analysis are not confounders of the main factors in the preliminary model [see Appendix B, Tables: 1-8].

Partial likelihood ratio test for the contribution of the interaction effect

The final step in model development strategy is consideration of interaction terms that may be useful in the improvement of the model. Moreover, we need to assess some realistic situations to see if two interaction effects can increase or decrease the survival time of infant's. The partial likelihood ratio test is used to identify the significance of some reasonable and possible interactions. The hypothesis to be tested is

H_0: The model with only main effect fits equally well as the model having

the main effect and their interaction as predictors.

H_1: H_0 is not true

Decision: Reject *Ho* at $\alpha = 0.05$ level of significance if -2 LOG L2- (-2 LOG L1) ≥

$x^2{}_1(\alpha = 0.05) = 3.84$, otherwise do not reject H_o. This means we need to include the corresponding interaction in the multivariate analysis.

Table 4.5: Partial likelihood ratio test for checking interaction terms

variables	-2 LOG L2 with main effect only	-2 LOG L1 with main effect and interaction	-2 LOG L2-(-2 LOG L1)	Sign
Sex*types of birth	9891.464	9890.236	1.228	Do not reject
Sex*mothers educational level	9891.464	9891.452	0.012	Do not reject
Sex,*birth order	9891.464	9889.554	1.91	Do not reject
Types of birth*mothers educational level	9891.464	9889.317	2.147	Do not reject
Types of birth *birth order	9891.464	9888.856	2.608	Do not reject
Mothers education* birth order	9891.464	9891.014	0.45	Do not reject

In Table 4.5 the interaction of each variable is assessed. Accordingly, none of the variables significantly interact with the other variables. Therefore, the final model will be the one which contains only the main effects sex, types of birth, mother's educational level and birth order given in Table 4.6.

Table 4.6: Estimated values of the coefficients, hazard ratios, 95% CI for the hazard ratio and P-values of the explanatory variables on fitting the proportional hazards model.

Analysis of Maximum Likelihood Estimates								
variables	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% CI for Hazard Ratio	
							LCL	UCL
sex female	1	-0.21035	0.07878	7.1295	0.0076	0.810	0.694	0.946

ref(male)								
Types of birth multiple birth Ref(single birth)	1	1.44362	0.11571	155.6543	<.0001	4.236	3.376	5.314
Mothers educational level primary Secondary and above Ref(no education)	1 1	-0.15568 -0.92759	0.09356 0.24461	2.7689 14.3801	0.0961 0.0001	0.856 0.396	0.712 0.245	1.028 0.639
birth order 2-4 5+ Ref(first birth)	1 1	-0.37559 -0.42350	0.10390 0.12128	13.0687 12.1929	0.0003 0.0005	0.687 0.655	0.560 0.516	0.842 0.83

Type 3 Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
sex	1	7.1295	0.0076
Types of birth	1	155.6543	<.0001
Mothers educational level	2	15.6645	0.0004
Birth order	2	15.2648	0.0005

Table 4.6 presents computer output of the result of the fitted hazard model. Based on the result we observe that all the covariates namely sex, types of birth, mothers educational level and birth order are significant at 5% level of significance. Since there is no continuous covariate, we can not check the linearity of covariates in the model. So, we consider this model as a preliminary final model and it could be the final model after we check proportionality assumptions.

4.5. Assessment of Model Adequacy

Having identified the preliminary final model the next step and most important in the statistical analysis is to diagnose the fit of the model.

After a model has been fitted to an observed set of survival data the adequacy of the fitted model needs to be assessed. The use of diagnostic procedures for model checking is an

essential part of the model in process. In our survival regression analysis assessment of model adequacy we must.

i) Test the assumption of proportional hazards.

ii) Check influence and poorly fit subjects and

iii) Overall summary measures of goodness of fit.

4.5.1. Assessment of the proportional hazards assumption

The proportional hazard assumption is one of the very important assumptions in the Cox model. The proportional hazards assumption, which asserts that the hazard ratios are constant overtime, is vital to the interpretation and use of a fitted proportional hazards model. That means, the risk of failure must be the same no matter how long subjects have been followed. In order to test this assumption, graphical diagnoses of scaled Schoenfeld residuals and likelihood-based tests, like the Wald test can be employed to assess the proportional hazard assumption to covariates that are significant in the multivariate analysis.

Under the assumption of proportionality of the proportional hazards model, the distribution of residuals over time must be random and LOWESS smoothing line should be a straight line around zero.

One of the statistical tests for proportional hazards assumption is to generate time varying covariates by creating interactions of the predictors and a function of survival times, usually covariate time's log of time, and including these in the model. If any of the covariates is significant then those predictors do not show a proportional effect over the study period. That is the proportional hazard assumption fails to hold.

Table 4.7: SAS result of the assumption of proportionality test

Analysis of Maximum Likelihood Estimates								
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Wald Chi-Square	Pr > ChiSq	Hazard Ratio
sex female	1	-0.25773	0.24370	1.1184	0.2903	1.1184	0.2903	0.773
birth order								
2-4	1	-0.46246	0.22985	4.0484	0.0442	4.4622	0.1074	0.630
5+	1	-0.47533	0.36654	1.6817	0.1947			0.622
Mothers educational level								
primary	1	-0.26247	0.27486	0.9119	0.3396	5.4354	0.0660	0.769
secondary and above	1	-1.69460	0.73448	5.3232	0.0210			0.184
Types of birth multiple birth	1	1.25245	0.42650	8.6235	0.0033	8.6235	0.1033	3.499
Sext	1	0.04596	0.16256	0.0800	0.7774	0.0800	0.7774	1.047
birthordert	1	0.07444	0.12420	0.3592	0.5489	0.3592	0.5489	1.077
moteredut	1	-0.02512	0.18000	0.0195	0.8890	0.0195	0.8890	0.975
Tbirtht	1	0.14204	0.27724	0.2625	0.6084	0.2625	0.6084	1.153

Linear Hypotheses Testing Results			
Label	Wald Chi-Square	DF	Pr > ChiSq
test proportionality	0.9537	4	0.9167

Table 4.7 shows the Wald chi-square value and corresponding p-values for each covariate. The result shows that, the p-value of the Wald test is greater than 0.05 for all covariates, implying that the proportionality assumption is satisfied. On the other hand, there are no covariates which show a trend/pattern with the time, which indicates the hazard ratios, will be constant over the study time.

Furthermore, plotting the scaled Schoenfeld residuals of each covariate against log time will be used to check whether the assumption of proportional hazards is violated or not.

The graphical display shows plots of the scaled Schonfield residuals against the survival time for each covariate namely types of birth, sex, birth order and mothers educational level. In Appendix D Figure 2 (a-d) plots of scaled Schoenfeld residuals show randomness. Moreover, the smoothed curve is an approximate horizontal line; so this also suggests that for the above four covariates the assumption of proportional hazards is satisfied.

4.5.2. Checking influential and poorly fit observations

The next step we follow is regression diagnostic with the purpose to determine whether any particular observation, if any, has an undue impact (leverage) on inferences made on the basis of model fitted to an observed set of survival data. It is, therefore, of particular interest to examine the influence of each particular observation on these estimates. This is done by examining the extent to which the estimated parameters and the maximized likelihood in the fitted model are affected by omitting in turn the data record for each individual in the study. The DFBETA statistic is used to examine the untoward effect of each observation on the j^{th} parameter estimate and the maximized log partial likelihood, respectively in the fitted Cox regression model Collett(2003). The five largest changes in the parameter estimates are presented in Table 2(Appendix C).

The largest difference for sex occurs for observation 2862. The change in the parameter estimate on omitting the data for this observation is 0.004060177. Therefore, omission of this observation increases the hazard of death relative to the baseline hazard. The

standard error of the parameter estimate for sex in the full data set is 0.07878 and so the maximum amount by which this estimate changed when one observation is deleted is about 5.15% of the standard error (less than one standard error). Thus, the change in sex effect by deleting this observation can be considered as insignificant.

The largest difference for birth order and types of birth occurs for observations 108 and 175 respectively. The change in the parameter estimate on omitting the data for each observation are 0.012142 (10.01% of the standard error), and 0.049829 (1.306% of the standard error), respectively. Both of them are within one standard error of the estimates. The effect of deleting these observations is increasing the relative hazard of death relative to the baseline hazard.

Omitting the data from observation 36 from the dataset brought the largest changes in the parameter estimates for the variable mother's educational level. The maximum change in the parameter estimates when this observation is omitted in turn is 0.006881956(7.35% of the standard error) within one standard error of the estimates. The effect of deleting these observations is increasing the relative hazard of death. Thus, the change in mother's educational level effect by deleting this observation can be considered as insignificant. We can conclude that neither the estimates for each of the parameters nor the set of parameter estimates are affected by any of the observations in the dataset.

4.5.3 Overall Goodness of Fit

After fitting the Cox model, its accuracy for predicting the survival of a given subject and the extent to which the fitted model provides an appropriate description of the observed data should be assessed. Thus, the Cox-Snell residuals are used to asses the overall

goodness of fit of the model. The plot of the cumulative hazard function of the Cox-Snell residual against the Cox-Snell residuals is shown in Figure3 (Appendix D).The plotted points in this Figure are fairly close to the straight line through the origin, which has unit slope. This suggests that the model fitted to the data is satisfactory.

4.6. Interpretation and discussion of the results

4.6.1. Interpretation of the results

The interpretation of the results from the fitted final model is based on the hazard ratios. The coefficient of the categorical covariates is interpreted as the logarithm of the ratio of the hazard of death to the baseline (reference group) hazard. That is, they are interpreted by comparing the reference group with others.

The estimated relative risk (hazard ratio) of dying for females compared to males is 0.810 (95% CI: 0.694-0.946, p=0.0076) implying that the risk of dying for females is 19% lower than the risk of dying for males (reference group) controlling for other covariates in the model.

The estimated relative risks (hazard ratio)of death for infant whose mothers had primary and secondary or above level of education compared to those infants whose mothers had no education (reference group) were 0.856 (95% CI: 0.712-1.028) and 0.396 (95% CI: 0.245 -0.639), respectively. These results reveal that the risk of death of infants to mothers with primary and secondary or above educational level are 14.4% and 60.4% lower than that of infants whose mothers have no education respectively controlling for other covariates in the model.

The estimated relative risks (hazard ratio) of death for infants with birth order 2-4 and 5+ compared to those infants with first birth (reference group) were 0.687 (95% CI: 0.560-0.842) and 0.655 (95% CI:0.516-0.830) respectively. These results reveal that the risk of death of infants with birth order 2-4 and 5+ are 31.3% and 34.5% lower than that of infants who had first birth order controlling for other covariates in the model.

The estimated relative risks (hazard ratio) of death for infants with multiple birth compared to those infants with single birth (reference group) were 4.236 (95% CI: 3.376-5.314) .These results indicates that infants with multiple birth order were 4.236 times more likely to die than those infants with single birth order controlling for other covariates in the model.

4.6.2. Discussion of the results

The results of the Cox proportional hazards regression analysis showed that mother's educational level, birthorder; types of birth and sex were significantly associated with infant mortality.

The impact of mother's educational level on survival rate of infant and child has been assessed by several studies indicating that high risk of death of infant and child was associated with low level of education of mothers. Aguirre (1995) identified that the mother's education to be the most important factor that directly affects infant mortality. A similar study in Malakal town, southern Sudan, Mahfouz et al (2009) found that mother's education, had a significant influence on infant and under-five mortality. A study by Caldwell (1979) found that infant and child mortality was highly associated with mother's education. A study conducted in Ethiopia by Ezra and Gurum (2002) and by

Kumar and Gemechis (2010) also found that the mortality risk of children born to non educated mothers as higher compared to those born to educated mothers. Another study by Desai and Alva (1998) examined the effect of maternal education on infant and child mortality and found that there was a consistent negative relationship between maternal education and the probability of infant death. Children of mothers who attended primary school were less likely to die than were children born to mothers with no education. The finding of the current study also agrees with the above sited findings and showing that infants who born with none educated mother experience higher risk of mortality than infants born to mothers with primary and higher education.

The current study showed that birth order was a risk factor of infant mortality. This study suggests that first born infants experienced higher risk of death than infants whose birth order was two to four and five and more; infants with birth order two to four had a higher risk of dying than infants whose birth order was five and above. A study conducted in Ethiopia by Kumar and Gemechis (2010) and by Desta (2011) also found that birth order was one of the determinants of child mortality showing that a first born child was exposed to a high risk of mortality. Another study by Balk et al (2003) indicated that first births were less likely to survive than higher order births.

The current study found out that infant mortality risk associated with multiple births was 4.236 times higher relative to singleton births. Kombo and Ginneken (2009) in Zimbabwe found that multiple births tend to increase infant and child mortality. Another study by Balk et al (2003) also found that multiple births were experienced much higher risk of death.

The current study identified that sex of child as a risk factor of infant mortality. This study suggested that the risk of dying for females was lower than for males. Mustafa and Odimegwu (2008) in a study in Kenya found that sex of child is important determinant for the risk of infant and child mortality.

CHAPTER FIVE

5. CONCLUSIONS AND RECOMMENDATIONS

5.1. Conclusions

In this study we identified the factors that are associated with high risk of infant mortality in Ethiopia using the methods of survival analysis. The Kaplan-Meier method was used to estimate the survival time of infants. A death proportion seems lower for females (20.6%) than for males (24.85%). Using Cox proportional hazard model covariates that significantly influence the survival of infants are identified. The study suggests that mother's educational level, birth order, sex and types of birth had statistically significant effects on the survival time of infants. The result of this study also indicated that Infant with multiple birth, first birth order, non educated mothers and male children are less likely to survive.

5.2. Recommendations

One of the goals of millennium development goals is reduction of infant and child mortality. The government of Ethiopia has implemented health oriented interventions to reduce child mortality. In order to reduce the rate of infant mortality this study suggests the following:

- Increase their education levels of mothers up to at least secondary levels. Education will change the parental behaviour toward children and educate the girls at least secondary level, so that they could take care of their children in respect to their health.

- We also saw that in Ethiopia multiple births are strongly negatively associated with infant and child survival. These results suggest that improving maternal and child health services, screening for high-risk pregnancies and making referral services for high-risk pregnancies more accessible. That means maternal and child health services should focus and identify such cases and provide those good health care services and guidance

REFERENCES

1. Aguirre, P.G. (1995). Child mortality and reproductive patters in urban Bolivia. *CDE working paper* No. 95-28, Center for Demography and Ecology, University of Wisconsin-Madison.
2. Baker, R. (1999). Differential in child mortality in Malawi. *Social networks project working Papers*, No. 3. University of Pennsylvania.
3. Balk, D., Tom, P., Adam, S., Fern, G. and Melissa, N. (2003). Spatial Analysis of Childhood Mortality in West Africa. Calverton, Maryland, USA: ORC Macro and Center for International Earth Science Information Network (CIESIN), Columbia University.
4. Caldwell, J. C. (1979). Education as a factor in mortality decline: an examination of *Nigerian data* , Population studies, **33(3)**,395-413.
5. Central Statistical Agency (The 1990 National Family and Fertility Survey Report), Addis Ababa, Ethiopia: Central Statistical agency.
6. Child Health in Ethiopia (2004) Background Document for the National Child Survival Conference, April 22-24, 2004, Addis Ababa, Ethiopia.
7. Collett, D. (2003). *Modeling survival data in medical research*, Second Edition. Chapman and Hall/CRC, London.
8. Cox, D.R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B (Methodological)*, **34(2)**, 187-220.

9. Desta, M. (2011). Infant and Child Mortality in Ethiopia: The role of socioeconomic, demographic and biological factors in the previous five years period of 2000 and 2005. Lund University.
10. Desai, S. and Alva, S. (1998). Maternal education and child health: Is there a strong causal relationship?', Demography, **35(1),**71-81
11. Ezra, M. and Gurum, E. (2002). Breastfeeding, birth intervals and child survival: analysis of the 1997 community and family survey data in southern Ethiopia.
12. Federal Ministry of Health. Essential Health Service Package for Ethiopia. Addis Ababa; 2005 August 2005.
13. Grambsch, P. and Therneau, T. (1994). Proportional hazards tests and diagnosis based on weighted residuals, *Biometrika* **81**, 515-526.
14. Hosmer, D.W. and Lemeshow, S. (1999). *Applied Survival Analysis*. John Wiley and Sons, Inc., New York.
15. Klaauw, V.B. and Wang, L. (2003). Child Mortality in Rural India, World Bank Working Paper, Washington DC: World Bank.
16. Kombo, J. and Ginneken, V. (2009). Determinants of infant and child mortality in Zimbabwe: Result of multivariate hazard analysis.
17. Kumar, P. and Gemechis, F. (2010). Infant and child mortality in Ethiopia: As statistical analysis approach. *Ethiopian Journal of Science and Education*, **5(2)**: 51-57.
18. Mahfouz, M.S., Dr. Surur, A.A., Ajak, D.A.A. and Eldawi, E.A. (2009). Level and Determinants of Infant and Child Mortality in Malakal Town. Southern Sudan, *Sudanese journal of public health*, **4 (2),** 250-255.

19. Manda, S.O.M. (1999). Birth intervals, breastfeeding and determinants of childhood mortality in Malawi. Social Science and Medicine, **48(3)**: 301-312.
20. Ministry of Finance and Economic Development (MoFED), Millennium Development Goals Report on Ethiopia, 2010.
21. Ministry of Finance and Economic Development (2008). Ethiopia: Progress toward achieving the MDGs.
22. Mturi, A. J. and Curtis, L. S. (1995). The determinants of infant and child mortality in Tanzani. Health Policy Plan, **10**,384-94.
23. Mustafa, E. and Odimegwu, C. (2008). Socioeconomic determinants of infant mortality in Kenya. Analysis of Kenya DHS 2003.
24. National Strategy for Child Survival in Ethiopia (2005) Family health department: Federal ministry of health, Addis Ababa, Ethiopia.
25. Schoenfeld, D. (1982). Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika*, **7,** 145-153.
26. UNICEF (2006). The State of the World's Children. New York.
27. UNICEF: Monitoring the situation of children and women, 2010. http://www.childinfo.org/mortality.html. posted on September 2011.
28. UNICEF (1999). The State of the World's Children. New York.
29. UNICEF (2000). The State of the World's Children. New York.
30. UNICEF (2008). Child mortality rate in Ethiopia falls by 40 percent.http://www.medindia.net/news/Child-Mortality-Rate-in-Ethiopia-Falls-by-40-Percent-UNICEF-32194-1.htm. Seen on 9/10/2011.

31. UNICEF Report (2010). Levels and Trends in child mortality, Estimates Developed by the UN Inter-agency Group for Child Mortality Estimation (IGME).

32. Uddin, J. (2009). Child mortality in a developing country: A statistical analysis. *Journal of Applied Quantitative Methods*, **4**, 270-283.

33. Wang, L. (2003). Environmental Determinants of Child Mortality: Empirical Results from the 2000 Ethiopia DHS. World Bank, Washington D.C.

34. Zimbabwe Central Statistical Office/ Macro International Inc. (2007), Zimbabwe Demographic and Health Survey Country Report. Harare: Central Statistical Office

APPENDIXES

APPENDIX A: Results of the multivariable proportional hazards Cox regression model

Table 1: Results of the multivariable proportional hazards Cox regression model containing the variables significant at 20 - 25% level in the univariable proportional hazards Cox regression model.

Analysis of Maximum Likelihood Estimates						
variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard ratio
sex		-0.20224	0.07888	6.5741	0.0103	0.817
female	1					
Types of birth						
multiple birth	1	1.44345	0.11609	154.5985	<.0001	4.235
Source of drinking water						
Other	1	0.23318	0.23886	0.9530	0.3290	1.263
Wealth index						
medium	1	-0.00805	0.11105	0.0053	0.9422	0.992
rich	1	0.05532	0.10382	0.2839	0.5942	1.057
Mothers educational level						
primary	1	-0.14704	0.08506	2.9886	0.0839	0.863
secondary and above	1	-0.02802	0.32278	0.0075	0.9308	0.972
Fathers educational level						
primary	1	-0.14596	0.09005	2.6276	0.1050	0.864
secondary and above	1	-0.14149	0.16367	0.7473	0.3873	0.868
Birth order 2-4	1	-0.39487	0.104410	14.3024	0.0002	0.674
5+	1	-0.46276	.12261	14.2449	0.0002	0.630
Maternal age 20-29	1	-0.14704	0.08506	2.9886	0.0839	0.863
>=30	1	-0.02802	0.32278	0.0075	0.9308	0.972
Place residence rural	1	0.00295	0.14960	0.0004	0.9842	1.003

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	148.6423	14	<.0001
Score	213.2024	14	<.0001
Wald	187.3600	14	<.0001

Wald			
Effect	DF	Chi-Square	Pr > ChiSq
sex	1	6.5741	0.0103
Types of birth	1	154.5985	<.0001
Source of drinking water	1	0.9530	0.3290
Wealth index	2	0.3317	0.8472
Mothers educational level	2	6.9490	0.0310
fathers educational level	2	2.7445	0.2535
Birth order	2	17.2132	0.0002
Maternal age	2	2.9969	0.2235
Place of residence	1	0.0004	0.9842

Table 2: Results of the multivariable proportional hazards Cox regression model after eliminating the variable place of residence from the multivariable proportional hazards Cox regression model in Table 1.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	148.6419	13	<.0001
Score	213.1991	13	<.0001
Wald	187.3577	13	<.0001

Type 3 Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
sex	1	6.5744	0.0103
Types of birth	1	154.6605	<.0001
Source of drinking water	1	1.1230	0.2893
Wealth index	2	0.3748	0.8291
Mothers educational level	2	6.9626	0.0308
fathers educational level	2	2.7715	0.2501
Birth order	2	17.2242	0.0002
Maternal age	2	3.0016	0.2229

Table 3: Results of the multivariable proportional hazards Cox regression model after eliminating the variable wealth index from the multivariable proportional hazards Cox regression model in Table 2.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	148.2698	11	<.0001
Score	212.7925	11	<.0001
Wald	186.8937	11	<.0001

Type 3 Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
sex	1	6.6333	0.0100
Types of birth	1	154.8220	<.0001
Source of drinking water	1	0.8985	0.3432
Mothers educational level	2	6.7056	0.0350
Fathers educational level	2	2.5317	0.2820
Birth order	2	17.3073	0.0002
Maternal age	2	3.0089	0.2221

Table 4: Results of the multivariable proportional hazards Cox regression model after eliminating the variable source of drinking water from the multivariable proportional hazards Cox regression model in Table 3.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	147.3287	10	<.0001
Score	212.2217	10	<.0001
Wald	186.4674	10	<.0001

Type 3 Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
sex	1	6.6634	0.0098
Types of birth	1	156.5447	<.0001
Mothers educational level	2	8.9825	0.0112
Fathers educational level	2	2.6748	0.2625
Birth order	2	17.1471	0.0002
Maternal age	2	3.2271	0.1992

Table 5: Results of the multivariable proportional hazards Cox regression model after eliminating the variable father's educational level from the multivariable proportional hazards Cox regression model in Table 4.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	144.6399	8	<.0001

Score	209.2038	8	<.0001
Wald	183.3521	8	<.0001

		Type 3 Tests	
Effect	DF	Wald Chi-Square	Pr > ChiSq
sex	1	6.9604	0.0083
Types of birth	1	154.9558	<.0001
Mothers educational level	2	14.6369	0.0007
Birth order	2	16.5287	0.0003
Maternal age	2	3.3432	0.1879

Table 6: Results of the multivariable proportional hazards Cox regression model after eliminating the variable Maternal age from the multivariable proportional hazards Cox regression model in Table 5.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	141.2402	6	<.0001
Score	205.2832	6	<.0001
Wald	179.2359	6	<.0001

Type 3 Tests			
Effect	DF	WaldChi-Square	Pr > ChiSq
sex	1	7.1295	0.0076
Types of birth	1	155.6543	<.0001
Mothers educational level	2	15.6645	0.0004
Birth order	2	15.2648	0.0005

APPENDIX B: The result of multivariable Cox hazard model those variables not significant in both the univariable analysis and in multivariable analysis fitted by included in the model containing those variable significant in multivariate analysis one at a time.

I. Result of multivariate Cox proportional hazard model containing variables in Table 6(Appendix A) including those not significant in univariable analysis one at a time.

Table 1: When types of toilet facility is included

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	141.8454	7	<.0001
Score	206.0394	7	<.0001
Wald	179.9549	7	<.0001

Type 3 Tests			
		Wald	
Effect	DF	Chi-Square	Pr > ChiSq
sex	1	7.2403	0.0071
Types of birth	1	156.2073	<.0001
Mothers educational level	2	16.2185	0.0003
Birth order	2	15.5510	0.0004
types of toilet facility	1	0.6461	0.4215

Table 2: When region is included

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	154.8540	16	<.0001
Score	217.2724	16	<.0001
Wald	190.6032	16	<.0001

Type 3 Tests			
		Wald	
Effect	DF	Chi-Square	Pr > ChiSq
sex	1	6.6470	0.0099
Types of birth	1	147.9792	<.0001
Mothers educational level	2	12.5259	0.0019
Birth order	2	15.9244	0.0003
region	10	12.7913	0.2356

Table 3: When religion is included

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	146.4502	10	<.0001
Score	210.7409	10	<.0001
Wald	184.2632	10	<.0001

Type 3 Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
sex	1	7.4024	0.0065
Types of birth	1	156.4813	<.0001
Mothers educational level	2	16.6954	0.0002
Birth order	2	15.2695	0.0005
religion	4	5.9547	0.2026

II. When those factors not significant in multivariate model are included to the Multivariate model of covariate in Table 6(Appendix A) one at a time

Table 4: when place of residence is included

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	141.6028	7	<.0001
Score	205.5412	7	<.0001
Wald	179.4590	7	<.0001

Type 3 Tests			
		Wald	
Effect	DF	Chi-Square	Pr > ChiSq
sex	1	7.1750	0.0074
Types of birth	1	154.9345	<.0001
Mothers educational level	2	12.5247	0.0019
Birth order	2	15.4088	0.0005
place of residence	1	0.3570	0.5502

Table 5: when Wealth index is included

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	141.2467	8	<.0001
Score	205.3005	8	<.0001
Wald	179.2470	8	<.0001

Type 3 Tests			
		Wald	
Effect	DF	Chi-Square	Pr > ChiSq
sex	1	7.1220	0.0076
Types of birth	1	155.3237	<.0001
Mothers educational level	2	14.5347	0.0007
Birth order	2	15.2169	0.0005
Wealth index	2	0.0065	0.9968

Table 6: when Source of drinking water is included

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	142.5865	7	<.0001
Score	206.1776	7	<.0001
Wald	180.0016	7	<.0001

Type 3 Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
sex	1	7.0940	0.0077
Types of birth	1	153.8009	<.0001
Mothers educational level	2	10.4500	0.0054
Birth order	2	15.6351	0.0004
Source of drinking water	1	1.2706	0.2597

Table 7: when Fathers educational level is included

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	144.0483	8	<.0001
Score	208.2560	8	<.0001
Wald	182.3148	8	<.0001

Type 3 Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
sex	1	6.8183	0.0090
Types of birth	1	157.4425	<.0001
Mothers educational level	2	9.6263	0.0081
Birth order	2	15.9273	0.0003
Fathers educational evel	2	2.7931	0.2475

Table 8: when maternal age is included.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	144.6399	8	<.0001
Score	209.2038	8	<.0001
Wald	183.3521	8	<.0001

Type 3 Tests			
		Wald	
Effect	DF	Chi-Square	Pr > ChiSq
sex	1	6.9604	0.0083
Types of birth	1	154.9558	<.0001
Mothers educational level	2	14.6369	0.0007
Birth order	2	16.5287	0.0003
Maternal age	2	3.3432	0.1879

APPENDIX C: Results of Model diagnostics

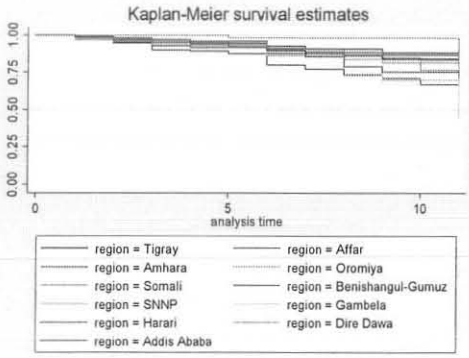
Table 1: The five highest differences in the parameter estimates of the variables included in the model in Table 4.6 when the data value for each patient is in turn deleted from the model.

Covariates	D e leted Observation (i)	$\Delta_{j(-i)} = \bar{\beta}_j - \bar{\beta}_{j(-i)}$	$\lvert\Delta_{j(-i)} = \bar{\beta}_j - \bar{\beta}_{j(-i)}\rvert$
Sex	36	-.003280137	.003280137
	251	0.003248421	0.003248421
	2862	- 0.004060177	0.004060177
	1297	0.003089101	0.003089101
	2512	0.003016100	0.003016100
Mothers education	36	-.006881956	0.006881956
	251	0.006809629	0.006809629
	484	0.005729522	0.005729522
	2512	-.003597072	0.003597072
	2870	0.003832860	0.003832860
Types of birth	175	0.049829	0.049829
	1263	-0.005008	0.005008
	1512	-0.004799	0.004799
	101	-.004546891	0.004546891
	10	-.004357086	0.004357086

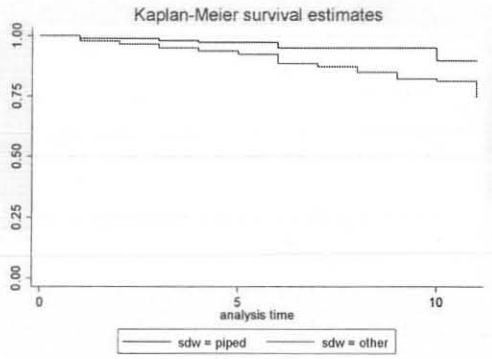
Birth order	108	-0.012142	0.012142
	2862	-0.012002	0.012002
	251	-.002522464	0.002522464
	175	-.002054738	0.002054738
	1297	-.001993450	0.001993450

APPENDIX D: Figures

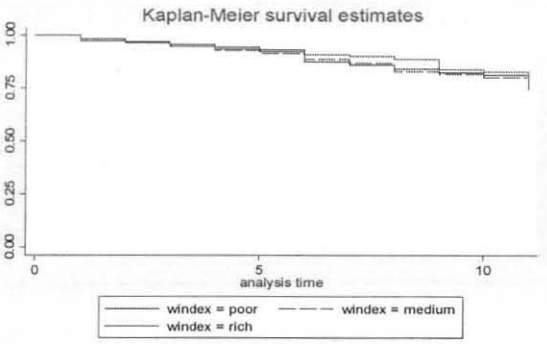
Figure 1 (a – k): Plots of Kaplan-Meier survivor functions different categories or group



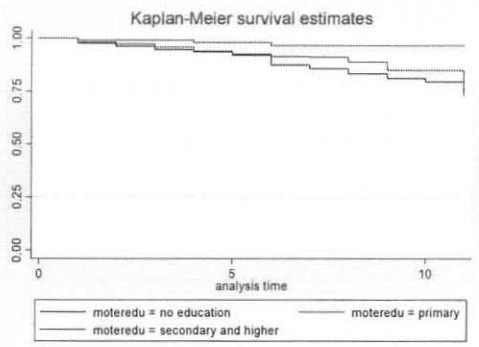
a) Survival curves of infant by region



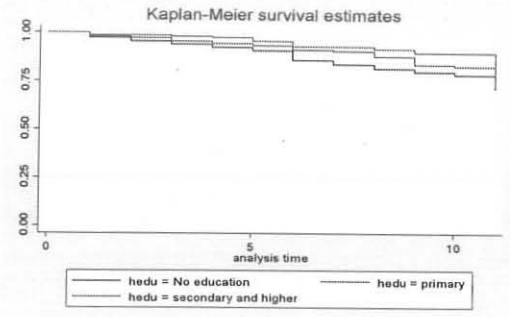
b) Survival curves of infant by source of drinking water



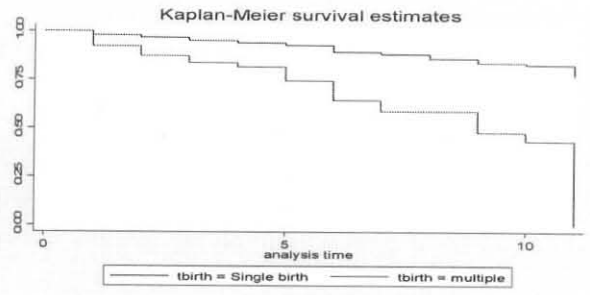
c) Survival curves of infants by wealth index



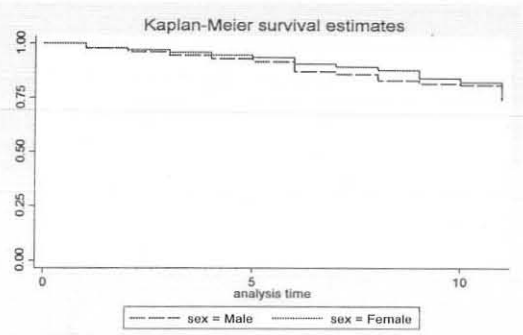
d) survival curves of infants by mother's education



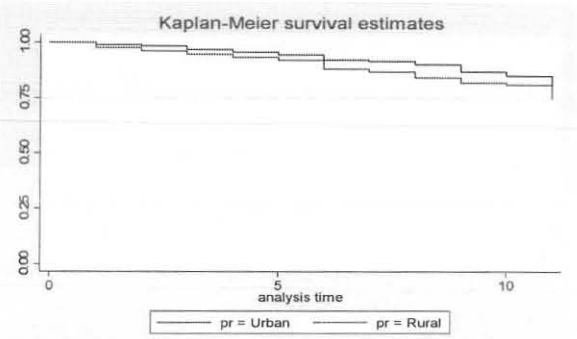
e) Survival curves of infants by fathers education



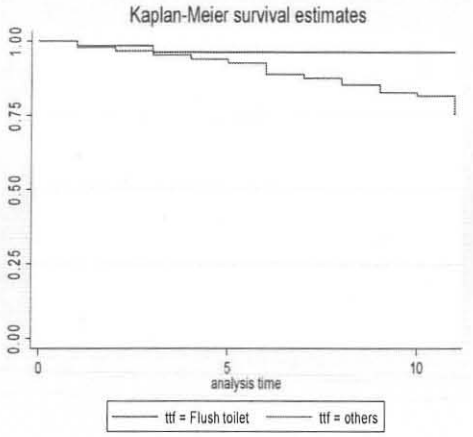
f) survival curves of infant by type of birth



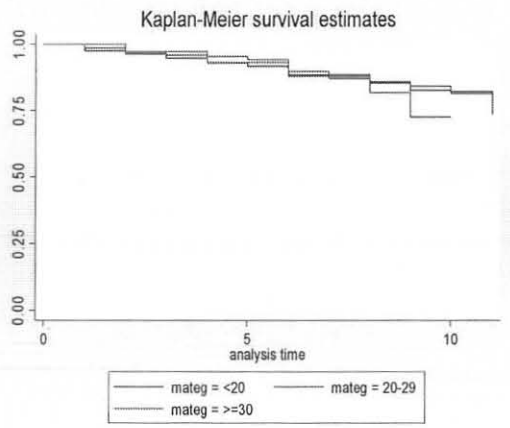
g) Survival curves of infant s by sex



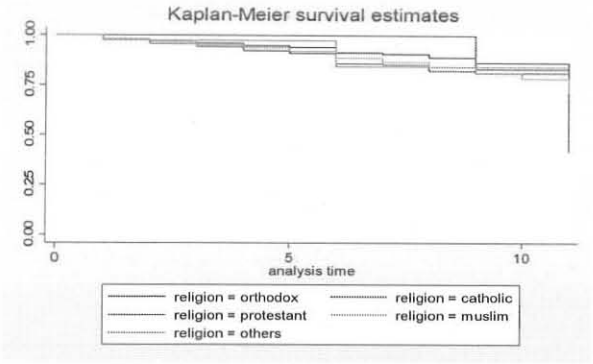
h) survival curves of infant by place of residence



i) Survival curves of infants by type of toilet

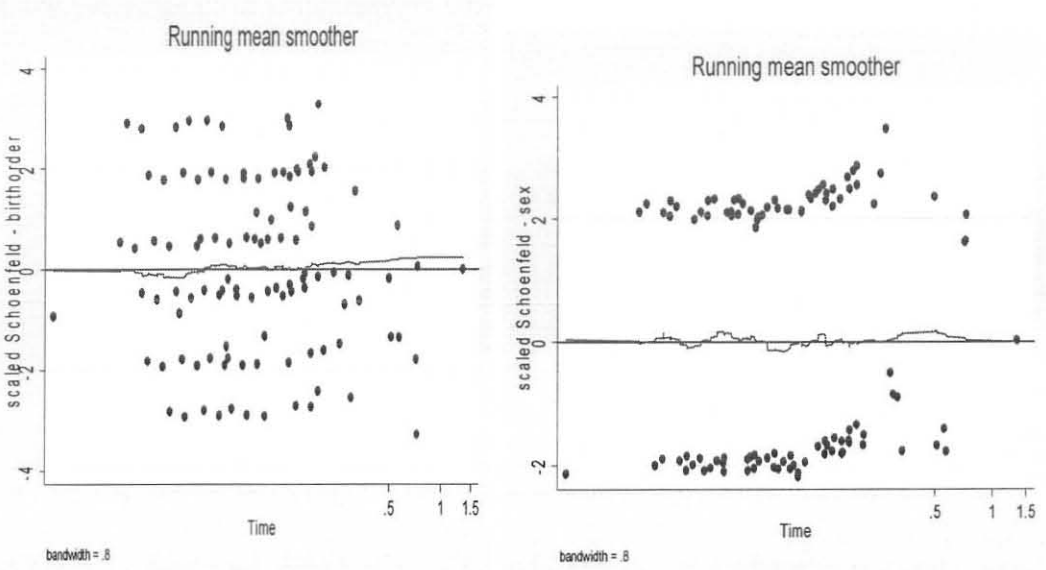


j) survival curves of infants by maternal age

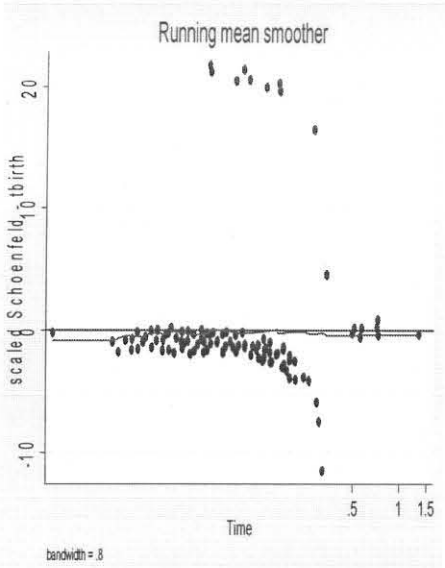


k) Survival curves of infants by religion

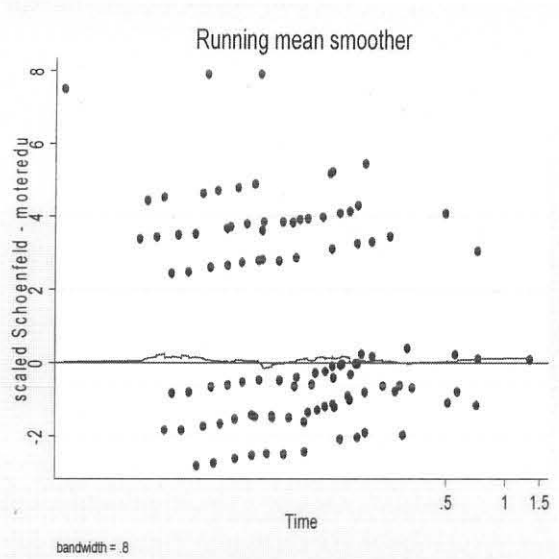
Figure 2: (a – d): Graphs of the scaled Schoenfeld residuals and their LOESS smooth curves obtained from the model in Table 4.6 for the covariates sex, types of birth , mothers educational level, birth order and birth order by types of birth interaction. The line that passes through zero is the reference line.



a) Scaled Schoenfeld Residuals by birth order b) Scaled Schoenfeld Residuals for sex



c) Scaled Schoenfeld Residuals for birth order



d) Scaled Schoenfeld Residuals for Mothers' education

Figure 3: Cumulative hazard plot of the Cox-Snell residuals of the proportional hazards Cox regression model obtained from the model in Table 4.6. The 45^0-striaght line through the origin is drawn for reference.