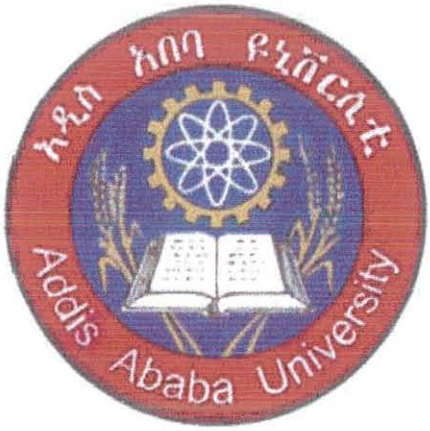


ADDIS ABABA UNIVERSITY

GRADUATE STUDIES PROGRAMME

DEPARTMENT OF STATISTICS

Determination of Factors Associated with High Risk of Infant Mortality in Ethiopia

By

Samuel Muluye

A thesis submitted to the school of Graduate Studies of Addis Ababa University in partial fulfillment of the requirements for the Degree of Masters of Science in Statistics (Biostatistics)

June, 2011

ADDIS ABABA UNIVERSITY

GRADUATE STUDIES PROGRAMME

DEPARTMENT OF STATISTICS

Determination of Factors Associated with High Risk of Infant Mortality in Ethiopia

By

Samuel Muluye

Approved by the Board of Examiners:

Department Head	**Signature**
Butte Gotu	
Examiner	**Signature**
Mekonnen Tadesse	
Examiner	**Signature**

TABLE OF CONTENTS

List of tables

List of figures

ACRONYMS

AIDS	Acquired Immunodeficiency Syndrome
ANOVA	Analysis of Variance
CI	Confidence Interval
EDHS	Ethiopian Demographic Health Survey
EMMUS	Morbidity and Services Utilization Survey
HIV	Human Immunodeficiency Virus
HR	Hazard Ratio
IMR	Infant Mortality Rate
KM	Kaplan-Meier
KNFS	Korean National Fertility Survey
LB	Live Births
LR	Likelihood Ratio
MDGs	Millennium Development Goals
MLE	Maximum Likelihood Estimator
PEM	Protein-energy Malnutrition
SE	Standard Error
SSA	Sub-Saharan Africa
STD	Sexually Transmitted Diseases
U-5 MR	Under-Five Mortality Rate
UN	United Nations
UNECA	United Nations Economic Commission for Africa
UNICEF	United Nations International Children's Emergency Fund
WHO	World Health Organization

ACKNOWLEDGMENTS

I am extremely grateful to my advisor Prof. Eshetu Wencheko for his help, guidance, consultation and encouragement up to the completion of this work.

My sincere and heartfelt thanks go to my mother Yeshialem Kumlachew, my sister Mahi and my Alish, who were the source of especial strength towards the successful completion of the study.

I would like to acknowledge my friends and colleagues Ermias Desse, Abdulkerim Kedir, Birtukan Tsehaineh and Zebene Ayele for their invaluable support during my period of study.

Finally, my heartfelt gratitude goes to my brother, Yalemzewd Molla for his generous co-operation in computer facilities and other related works throughout the study.

ABSTRACT

Ethiopia has the highest rate of infant deaths in Eastern and Southern Africa. This study addresses important issues concerning infant mortality in Ethiopia. The objective of the paper is to determine the impact of socioeconomic, demographic and environmental variables on infant mortality. The data for this study were obtained from the demographic and health survey (DHS) conducted in Ethiopia 2005. The results of Kaplan-Meier estimation show that most of the deaths occurred in the earlier month from birth to one month and then after death declined in the later months. The Cox proportional regression model was fitted to select the significant factors affecting infant mortality in Ethiopia. The model considered provided good fit for the data. Based on the result of the Cox proportional regression model, infant mortality was significantly associated with breast status, mother age at birth, mother's education, birth order, source of drinking water and sex ($p < 0.05$). This study supports health policy initiatives to stimulate use of family planning methods to increase birth spacing. It is hoped that, the results could be used by policy makers and programme managers in the child health sector to formulate appropriate strategies to improve the situation of infants in Ethiopia.

CHAPTER ONE
INTRODUCTION

1.1 Background of the study

A newborn is an infant who is within hours, days, or up to a few weeks after birth. In medical contexts, newborn or neonate (from Latin, neonatus, newborn) refers to an infant in the first 28 days of life (Merriam-Webster online dictionary, 2007). The term "newborn" includes premature infants, post mature infants and full term newborns. The term infant is derived from the Latin word infans, meaning "unable to speak" or "speechless." It is typically applied to children between the ages of 1 month and 12 months; however, definitions vary between birth and 3 years of age. "Infant" is also a legal term referring to any child under the age of legal adulthood. (Merriam-Webster online dictionary, 2007).

Children in the third world, especially in sub-Saharan Africa, usually suffer from more than one disease at a time. There appears to be a 'synergism of infection' whereby children tend to suffer from several diseases at the same time on top of protein-calorie malnutrition, which appears to be largely responsible for infant and child mortality (UNECA, 1979). In most countries of sub-Saharan Africa, the main causes of infant deaths are more or less the same.

Ethiopia is a sub-Saharan Africa country with a land area of 1.14 million square kilometers. The size of the country and its location has accorded it with diverse topography, geographic and climatic zones and resources. With a projected population of 75.1 million in 2006, Ethiopia is the second most populous country in Sub-Saharan Africa (SSA). About 85% of the population resides in rural areas while the rest live in urban areas. Females constitute about 50% of the population.

The rate of infant mortality is an important indicator of a nations' socioeconomic welfare. Tremendous reduction in infant mortality rate took place since 1900s. Despite this reduction, infant mortality rate (IMR) – the probability, expressed as a rate per 1,000 live births, of a child born in a specified year dying before reaching the age of one – is still high especially in less developed countries. According to the data of Millennium Development Goals Indicators collected by the United Nations, IMR in 2007 at world level is 47; this rate is 5 for developed regions whereas 51 for the developing countries (Millennium Development Goals Indicators

2009). Infant mortality rates are divided into two periods: neonatal and postneonatal. Neonatal mortality is death occurring in the first month of life and is typically associated with events surrounding the neonatal period and the infant's delivery. The highest risk for infant death is in the neonatal period. The primary direct causes of neonatal death worldwide are preterm birth (28 percent), severe infections (26 percent), and asphyxia (23percent).

According to UNICEF Plan of Operation in Cooperation with the Ethiopian Government 1994-1999, Ethiopia is the country with the highest rates of child and maternal deaths in all of Eastern and Southern Africa. The same applies to protein-energy malnutrition (PEM) and to long-term development retardation of survivors. The problem is deep-rooted in areas of prolonged civil war and drought, where social rehabilitation becomes much more complex due to disruption of social support networks, displacement of populations and environmental degradation. The situation is illustrated by IMR 101.1 per 1000 live births (LB), an Under-Five Mortality Rate (U-5 MR) of 152 per 1000 live births, and Maternal Mortality Rate of 700 per 100,000 live births. The underlying causes of children and women deaths can be attributed to household food insecurity, inadequate environmental sanitation and safe water supply, inadequate access to health services and inadequate care of children and women. Inadequate care results in improper feeding practices for children and poor dietary habits for women during pregnancy. In Ethiopia, this is usually the result of an excessive maternal work burden and lack of mother's time to take care of the child and herself.

Poverty is one of the most important factors affecting the infant mortality rate in Africa. Ethiopia is one of the poorest African countries with, according to UNICEF (2009) report, with a Gross National Income per capita of about $220 in 2007. The mortality rate for infants is 96.8 per 1000 live births (EDHS 2000). According to UNICEF (2009) report the estimated infant mortality rate was 122 deaths per 1000 live births in 1990 and 75 deaths per 1000 live births in 2007.

In recent years, Ethiopia has made progress in improving health care for children, reducing the under-five mortality rate by 42% since 1990, but the rapidly growing population means that the number of children dying is now almost constant. Infant and child mortality rates remain high, with most deaths being caused by easily preventable diseases, such as malaria, pneumonia and diarrhoea. (UNICEF, 2009)

1.2 Statement of the problem

The study of infant mortality becomes one of the most important researches in developing countries because there is high level of infant mortality. There is little research on the patterns of determinants of infant mortality, by analyzing how infant mortality is differently affected by demographic, socioeconomic and environmental variables.

Most paper not provides an in-depth use of Demographic and Health Survey data. The data used in this paper were obtained from the Demographic and Health Survey conducted in Ethiopia in 2005. The overall purpose of the paper is to determine the relative importance of various demographic, socioeconomic and environmental variables on infant mortality in Ethiopia. In particular, the study will focus on the relationship between infant mortality and birth order, mother age at birth, child's sex, , breastfeeding status, family size, marital status, mother's education, father's education, wealth index, area of residence, source of drinking water.

1.3 Objectives of the study

The general objective of the study is to identify the determinant factors of infant mortality in Ethiopia by using survival analysis using non-parametric and semi-parametric methods.
The specific objectives of the study are:

- To determine the survival time of infants.
- To determine the factors and/or covariates that affects the survival of infants.
- To compare survival time among the different groups of infants.

1.4 Significance of the study

- It is hoped that the study would provide an in-depth use of Demographic and Health Survey data. It is expected to improve the understanding of the mortality situation of infant under the age of one in Ethiopia.

- The results could be of interest to other studies related to infant mortality risks in Ethiopia.
- The result of this study could provide information to government and other concerned bodies in setting policies, strategies, and further investigation for reducing infant mortality.

1.5 Limitation of the study

- This paper is restricted to the period 2000-2005, covariates refer to the time of the survey, but not the covariates indicate the exact time of exposure of the children.
- The study is based on only the set of data for which complete information on survival times are available because of missing value.
- The study used a cross sectional data from Ethiopian demographic and health survey, which does not give the exact survival time. i.e. The survival time of children calculated from their birth calendar which does not consider the days.

CHAPTER TWO
LITERATURE REVIEW

2.1 Theoretical literature

Theoretical frameworks are often presented as health production functions, which capture the structural relation between health outcomes and the household's behavioral variables, like nutrition, breastfeeding, child spacing, etc. (see Schultz, 1984). In the framework of a health production function, child mortality risks depend on both observed health inputs and unobserved biological endowment or frailty. Not properly taking into account of these unobserved characteristics or the relation between children within a family may lead to inconsistent and inefficient estimators (for example, see Ridder and Tunali, 1999).

There are a number of different analytical frameworks through which the effects of different determinants on children mortality could be viewed. Demographic research by Mosley and Chen (1984) and by Schultz (1984) made the distinction between variables considered to be exogenous or socioeconomic (i.e. cultural, social, economic, community, and regional factors) and endogenous or biomedical factors (i.e. breastfeeding patterns, hygiene, sanitary measures, and nutrition). The effects of the exogenous variables are considered indirect because they operate through the endogenous biomedical factors. Likewise, bio-medical factors are called intermediate variables or proximate determinants because they constitute the middle step between the exogenous variables and child mortality (Jain, 1988; Mosley and Chen, 1984; Schultz, 1984; UN, 1985).

Mosley and Chen (1984) were among the first to study the intermediate biomedical factors affecting child mortality, labeled 'proximate determinants' They distinguished fourteen proximate determinants and categorized them into four groups: maternal [fertility] factors, environmental sanitation factors, availability of nutrients to the foetus and infant, injuries, and personal illness control factors.

Demographers view declines in mortality and fertility as components of a single "demographic transition". There are many competing theories about why fertility declines. One theory favoured by demographers is that fertility decline is due to the mortality decline, i.e., it is the response to the improved survival chances of the offspring. An alternative theory

proposed by Becker (1981) suggests that the demographic transition occurs since at high levels of income, the adverse effect of the opportunity cost of children on child rearing dominates the positive income effect. This theory, however, is inconsistent with the simultaneous occurrence of the fertility transition in countries that markedly differed in their levels of income (Kalemli-Ozcan, 2002). Proponents of modern population theories (Houndroyiannis and Papaetrou, 2002) argued that there might be a case that fertility may be induced by mortality. In general, mortality has a direct relationship with fertility. Since, higher fertility can be a signal for narrow birth spacing parents may not allot enough time to look after their children. Above all, a larger family size with meagre resources forces the family to malnourishment, lower health care utilization, use of improper sanitation system and water supply. For this reason, there is a high risk of infant and child death.

And further the WHO Commission established that reducing child mortality is a key to economic growth, for a variety of reasons. Societies with high rates of infant and child mortality have higher rates of fertility, and large numbers of children reduce the ability of poor families to invest in health and education, resulting in an under-trained, under-skilled productive work force.

In addition to its effect on fertility, child mortality is also important for the human capital investment decision of parents. Lower mortality implies a higher rate of return to education, and thus declining child and youth mortality provides an important incentive to increase investment in the education of each child (Kalemli-Ozcan, 2002). Heckman (2000) also argues that the return to human capital investment is highest before age five. Secondly, lowering infant mortality rates tends to lower, not raise, population growth over the long run, as people adjust to having smaller families (WHO, 2001).

2.2 Empirical literature

Casterline et al. (1989) examine the effects of income on infant and early childhood mortality at the household level in Egypt. They also incorporate socioeconomic and demographic variables in their logistic regression equation, where this type of model does not account for censored data. The main conclusions of Casterline et al. concerning income are: (1) household income does not affect survival through infancy but the effects are pronounced

during early childhood; and (2) the data used suggests that the impact of income is somewhat greater for educated mothers, when the father is of higher socioeconomic status and where the household receives piped water.

Jacoby and Wang (2003) examine the linkages between child mortality and morbidity, and the quality of the household and community environment in rural China using a competing risks approach. The key findings are that (1) the use of unclean cooking fuels (wood and coal) significantly reduces the neonatal survival probability in rural areas; (2) access to safe water or sanitation reduces child mortality risks by about 34% in rural areas; (3) a higher maternal education level reduces child mortality and that female education has strong health externalities (4) access to safe water/sanitation, and immunization reduce diarrhea incidence in rural areas, while access to modern sanitation facilities (flush toilets) reduces diarrhea prevalence in rural areas; (5) significant linkages between Acute Respiratory Infections (ARI) incidence and use of unclean cooking fuels are found using the city level data constructed from the survey.

Bicego (1990) employed a three-step procedure using proportional hazards regression to examine the determinants of childhood mortality in Haiti. He used the data from the 1987 Mortality, Morbidity and Services Utilization Survey (EMMUS) in Haiti. Maternal education and low age at birth were found to have marked effects on neonatal survivorship but little effect thereafter. Indices that reflect community-level access to child health services were shown to be important especially during childhood.

A study based on the 2003 Kenyan Demographic and Health Survey by Hisham and Clifford (2008) showed that for the post-neonatal mortality, the significant factors are maternal education, place of residence, ethnicity, and age at first birth, sex of the child and breastfeeding status. In rural areas, the significant predictors of post-neonatal mortality are ethnicity, breastfeeding status, birth order, birth interval and mother's educational level. In the urban areas, survival of the child is determined by maternal awareness of child care and level of education. The most important determinants of infant mortality are breastfeeding status followed by ethnicity, then fertility factors (birth order and intervals) and the least is the gender of the child. Once the child has survived the first month, ethnicity becomes the most important determinant of mortality in both urban and rural settings, then, followed in sequence by breastfeeding status, gender of the child, fertility factors, and the least significant

ones are the mother's occupation and her highest level of education. The effect of ethnicity on infant mortality could be explained by the socioeconomic inequality between the ethnic groups. In general, the study showed that biological and demographic variables are more important determinants of infant and post neonatal mortality.

Kim (2004) also examines the determinants of infant and child mortality in Korea. The study identifies the major factors which were associated with infant and child mortality in Korea using data from the 1974 Korean national fertility survey (KNFS). In urban and rural areas, mother's education, maternal age, number of rooms in household home, previous and successive birth interval were the most important determinant of infant mortality and child mortality. Infant mortality was also significantly affected by sex of child in urban areas and by birth order in rural areas. And also demographic factors are more important determinate of infant mortality in rural areas, where as socioeconomic factor play a major role for infant mortality in urban areas.

Woldemichael (1988) examined the effect of some environmental and socioeconomic factors that determine childhood diarrhea in Eritrea. He applied logistic regression by using data from the 1995 Eritrea Demographic and Health Survey. The results show that type of floor material, household economic status and place of residence are significant predictors of diarrhea.

White (2006) applied a Cox proportional hazards model to examine the determinants of infant and child mortality in Andhra Pradesh (where the Young Lives project was taking place) and Kerala. Infant mortality is found to depend on biological factors, including mother's age and birth order, and also factors related to health service provision such as tetanus injection and use of antenatal services. Economic wellbeing is not significant once these factors are taken into account. By contrast, economic well-being was found to be a significant determinant of child mortality, but substantially outweighed in importance by other factors such as maternal education and knowledge of health practices (ORS) and access to safe water. The result also show gender discrimination in Andhra Pradesh, notably toward girls with only female siblings, which is absent in Kerala. The study concluded that raising service levels across India toward the levels found in Kerala is a necessary step towards meeting the MDGs, and that the success of these efforts is reinforced by female empowerment.

Zerai (1996) employed Cox regression to examined socio-economic and demographic variables in a multi-level framework to determine conditions influencing infant survival in Zimbabwe. He used data from 1988 Zimbabwe Demographic and Health Survey data to study socioeconomic determinants of infant mortality. The unique finding was that women's average educational level in their community exerts a great influence on infant survival. This result supports assertions that child survival is strongly impacted by mass education (Cleland and van Ginneken 1988)

Wang (2003), using the results from the 2000 Ethiopia Demographic and Health Survey considered environmental determinants of child mortality by constructing three hazard models (the Weibull, the Piece-wise Weibull and the Cox model) to examine three age-specific mortality rates: neonatal, infant, and under-five mortality by location (rural/rural), female education attainment, religious affiliation, income status, and access to basic environmental services (water, sanitation and electricity). The estimation results show a strong statistical association between child mortality rates and poor environmental conditions.

Hailemariam and Tesfaye (1997) conducted a study in a small urban community in Sebeta, a town 25 km west of Addis Ababa, Ethiopia. They showed that higher birth order, early pregnancy and late pregnancy do have significant negative impact on the livelihood of infants. Their finding using Cox's Proportional Hazard Model show that maternal education, occupation of the father, household income, source of drinking water, availability of latrine and survival status of older offspring has direct effect on infant mortality.

A Tanzanian study had shown lack of infant and child mortality differentials by such socioeconomic factors as maternal education, partner's education, urban/rural residence, and presence of radio in the household. But demographic factors such as short birth interval (less than 2 years), teenage pregnancies (< 20 years) and previous child death were all significantly associated with increased infant and child mortality. There is lack of infant-child mortality differentials by economic status (wealth index), ethnicity and sex of the child (Mturi and Curtis, 1995).

Schellenberg et al. (2002) examined the risk factors for child mortality in rural Tanzania. They conducted a community based Nested Case Control Study of Post-neonatal death in

children less than five years. They investigated the effects of demographic, socio-economic, health seeking behavior and household environment on enhancing or impeding infant or child mortality. The results have shown that maternal education, socioeconomic status and breastfeeding have significant impact on infant and child mortality.

In Kenya, Hill et al. (2001) reported an inverse relationship between mother's educational level and economic status (wealth index) and child mortality. While for the relationship between urban/rural residence and child mortality, urban areas showed higher mortality risks than rural, but when adjusted for HIV prevalence, child mortality was lower in urban areas.

In Kenya, Mutunga (2004) found that child survival was found better for those who were of birth order 2-3, birth interval more than 2 years, not outcomes of multiple births, living in wealthier households, had an access to drinking water and sanitation facilities. But maternal age, maternal education and gender of the child had no significant association with child mortality.

Baker (1999) use indirect methods to estimate levels and trends of mortality in Malawi. Although the results from a previous study indicate that owning a pit latrine does not have a significant effect on child mortality (which is explained by the argument that just because a household has sanitation facilities does not mean that it will be used hygienically or by all members of the household). The results by Baker (1999) indicate that source of drinking water and sanitation facilities are strong predictors of infant mortality.

Hala (2002) applied duration modelling to assess the impact of water and sanitation on child mortality in Egypt. Results show that access to municipal water decreases sanitary risks.

Manda (1999) used data from the 1992 Demographic and Health Survey 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 children mortality. The study employed the proportional hazards model. The results show that birth interval and maternal age effects are largely limited to the period of infancy.

Ali (2002) employed a three part model: a probit model specification for the neonatal case, nonparametric, semi-parametric, and parametric duration model to examine the effects of water and sanitation on child mortality risk in Egypt. He used data from the Demographic and Health Survey conducted in Egypt between November 1995 and January 1996. The results of the study showed that access to municipal water and improved sanitation facilities had significant positive impact on child mortality. Moreover, the study indicated that mother education as an important factor to reduce child mortality. The study recommended that improving the knowledge about health care and hygiene in the society is crucial to reduce risk of child death.

Lee et al (1997) examined the effects of improved nutrition, sanitation and water quality on child health in high mortality population. They paid particular attention on non-random allocation of household resources to children and to the selectivity effects of health interventions via their effects on child survival. Unlike the previous studies, they employed a simultaneous equation model with selectivity. The results show that child mortality was affected by source drinking water, sanitation facilities, a child specific nutritional intake and mothers' education. In contrast to other studies, they concluded that variation in water sources and improvement in sanitation facilities do not have significant impact on child mortality, but wealth and parental schooling levels were significantly and positively associated with higher survival. The rationale behind such deviated result was that they focused on the reduced allocation of household income on children's health care with better facilities and not to survival selectivity.

Regarding the association between socioeconomic status and infant and child mortality, Hobcraft (1993) explained that education can contribute to child survival by making women more likely to marry and enter motherhood later and have fewer children, utilize prenatal care and immunize their children.

Asefa and Tessema (1997) showed that infant mortality rate was higher for males 103.7 per 1000 live births compared with females 86.6 per 1000 live births for singletons and 477.8 per 1000 live births for males compared with 417.9 per 1000 live births for females in the case of multiple births. Male children in general experience higher mortality than female children. The gender difference is especially pronounced for infant mortality, where 1 in 11 boys dies before his first birthday, compared with 1 in 14 girls in Ethiopia

Childhood mortality in general and infant mortality in particular is often used as broad indicators of social development or as specific indicators of health status. This is because more than any other age-group of a population, infant's survival depends on the socioeconomic conditions of their environment (Madise, 2003). Hence its description is very vital for evaluation and planning of the public health strategies. One of the most important items in the Millennium Development Goals (MDG) is to reduce infant and child mortality by two-thirds between 1990 and 2015 (UNICEF, 2006).

In general, maternal (and related) factors, household socio-economic status, and environmental characteristics have significant effects on child and infant mortality. This is true for studies which employ both direct and indirect techniques to estimate infant and child mortality.

CHAPTER THREE
DATA AND METHODOLOGY

3.1 Data

The data were obtained from the Demographic and Health Survey conducted in Ethiopia in 2005, which is a second comprehensive and nationally representative population and health survey. The data set give in-depth information on demographic and health aspects of households. Information regarding fertility and family planning behaviour, child mortality, nutritional status of children, utilization of maternal and children health services and knowledge of HIV/AIDS and sexually transmitted diseases (STD) is available from the data set.

The 2005 Ethiopia Demographic and Health Survey was designed to provide estimates for the health and demographic variables of interest for the following domains: Ethiopia as a whole; urban and rural areas (each as a separate domain); and 11 geographic areas (9 regions and 2 city administrations).

In general, the Ethiopia Demographic and Health Survey sample is stratified, clustered and selected in two stages. In the 2005 Ethiopia Demographic and Health Survey a representative sample of approximately 14,500 households from 540 clusters was selected. In the first stage, 540 clusters (145 urban and 395 rural) were selected from the list of enumeration areas based on the 1994 population and housing census sample frame.

The number of children at this level was 9,861 representing the number of live births born to the interviewed mothers in the period of five years preceding the date of the survey. Then, after a certain rearrangement and reorganization of the data 7,118 children with complete information were used as the data for this study.

3.2 Variables to be included in the study

The Response (dependent) Variable

Infant mortality is analysed in age period: mortality from birth to the age of 11 months, which will be referred to as "infant mortality".

The outcome variable is the survival time of infant, the length of time from birth date until the date of death/censor measured in months.

Predictor (independent) Variables

The explanatory variables/factors considered in this study are categorized as Socioeconomic, Demographic and Environmental variables/factors.

A. Demographic variables/factors

- Child's birth order(1, 2-4, >4)
- Mother's age(15-20, 20-34, >=35)
- Child's sex(Male, Female)
- Breastfeeding status(Yes, No)
- Family size(1-3, 4-6, >=7)
- Marital status(Currently Married, Currently not Married)

B. Socioeconomic variables/factors

- Mother's education(No Education, Primary, Secondary and Higher)
- Father's education(No Education, Primary, Secondary and Higher)
- Wealth index(Poor, Medium, Rich)
- Area of residence(Urban, Rural)

C. Environmental variables/factors

- Source of drinking water(Pipe, protected well/ spring, unprotected source)

3.3 Methodology

Survival Analysis

Survival Analysis involves the modelling and analysis of data that have a principal end point, the time until an event occurs (time-to-event data). Generally, survival analysis is a collection of statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs. Such types of data frequently arise from medicine, public health, demography, etc where the analysis is usually referred to as Survival data analysis and industrial studies in engineering fields as Reliability analysis.

Survival data analysis involves a dependent variable, time-to-an event, which is always non-negative and has a positively skewed distribution. It considers conditional information on the remaining time of subject's survival given current survival time. Moreover, there are certain aspects of survival analysis data, such as censoring and non-normality that generate great difficulty when trying to analyze the data using traditional statistical models such as multiple linear regressions. Therefore, survival data are not in general amenable to standard statistical procedures, such as mean, standard deviation and ANOVA, used in statistical analysis.

One of the most important differences between the outcome variables modeled via linear and logistic regression analyses and the time variable in the survival data is the fact that we may only observe the survival time partially. The variable time actually records two different things. For those subjects who died, it is the outcome variable of interest, the actual survival time. However, for subjects who were alive at the end of the study, or for subjects who were lost to follow-up, time indicates the length of follow-up (which is a partial or incomplete observation of survival time). These incomplete observations are referred to as being censored. There are four different types of censoring possibility: right truncation, left truncation, right censoring and left censoring.

This paper focused exclusively on right censoring. When an observation is right censored it means that the information is incomplete because the subject did not have an event during the time that the subject was part of the study. The point of survival analysis is to follow subjects over time and observe at which point in time they experience the event of interest.

3.3.1 Descriptive methods for survival data

Descriptive analysis for survival data is to present numerical or graphical summaries of the survival times in a particular group. In general, a statistical analysis should begin with a thoughtful and thorough univariate description of the data. Survival data are conveniently summarized through the estimates of survivor function and hazard function. These methods of estimation are said to be non-parametric or distribution-free, since they do not require specific assumptions to be made about the underlying distribution of the survival times.

The survivor function S (t)

The survival function denotes the probability that an individual survives up to a particular time t. The function is obtained from what is known in the survival analysis literature as the failure function. Which is the distribution of T or the cumulative distribution function of T is the probability that an individual will die before time t.

$$F(T)=P(T<t)=\int_0^t f(u)du, t\geq 0 \qquad [1]$$

The survivor function, $S(t)$, is defined to be the probability that the survival time of a randomly selected subject is greater than or equal to some specified time t and so

$$S(t) = p(T\geq t) = 1 - F(t),\ t\geq 0 \qquad [2]$$

The survival function is the probability that an individual will survive at time t or beyond t, and then the probability density function f (t) will be:

$$f(t) = \frac{d(1-S(t))}{dt} = -\frac{d(S(t))}{dt} \qquad [3]$$

The hazard function h (t)

The hazard function is widely used to express the risk of hazard of death at some time t, and is obtained from the probability that an individual dies at time t, conditional on he or she having survived to that time. It therefore, represents the instantaneous failure rate for an

individual surviving to time t. For $h(t) \geq 0$, the hazard function h (t) is given by the following:

$$h(t) = \lim_{\Delta t \to 0} \frac{p\{\text{an individual fails in the time interval } (t, t+\Delta t) \backslash \text{it survives until time } t\}}{\Delta t}$$

$$h(t) = \lim_{\Delta t \to 0} \frac{p\{t \leq T \leq t+\Delta t \backslash T \geq t\}}{\Delta t}$$

By applying the theory of conditional probability and the relationship in equation [3], the hazard function can be expressed in terms of the underlying probability density function and the survivor function as follows

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d \ln s(t)}{dt} \qquad [4]$$

A related quantity is the cumulative hazard function H(t) defined by

$$H(t) = \int_0^t h(u)du = -\ln s(t) \qquad [5]$$

Thus $S(t) = \exp(-H(t))$ consequently $f(t) = h(t) \exp(-H(t))$ [6]

There are different estimators of the survival and hazard functions. The most common used methods are Kaplan-Meier analysis, Nelson-Aalen and Life Tables.

The Kaplan-Meier estimator of the survivorship function (Kaplan and Meier, 1958) is also called the product limit estimator. The Kaplan-Meier estimator is used to estimate survival time of infants and construct survival curve to compare survival experience of infant between different categorical variables. Moreover, Kaplan-Meier estimator of the survival function is based on individual observations. 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.

The Kaplan-Meier estimator of the survival function is to order the survival times as $t_{(1)} < t_{(2)} < \ldots . < t_{(n)}$. Assume that among the n observations there are $m \leq n$ failures occurred at distinct m times.

Then the Kaplan-Meier estimator of the survival function at time t is obtained from the equation.

$$\hat{S}(t)=\prod_{i-1}^{k}\left(\frac{n_i-d_i}{n_i}\right) \quad [7]$$

for $t_{(k)}\leq t<t_{(k+1)}$ k=1,2,.....,m with the convention that $\hat{S}(t)=1$ if $t < t_{(1)}$.

In equation [7]:

n_i= the number at risk of dying or failure at $t_{(i)}$

d_i= the number of failures at $t_{(i)}$

The standard error of the KM survival estimator which is also known as the Greenwood's formula is (Collett, 2003)

$$\text{Se}\{\hat{S}(t)\}=\hat{S}(t)\left\{\sum_{j=1}^{k}\frac{n_j}{n_j(n_j-d_j)}\right\}^{1/2}, \text{ for } t_{(k)}\leq t < t_{(k+1)}. \quad [8]$$

The Nelson-Aalen estimate of the survivor function, which is based on the individual event times, is the Nelson-Alen estimate, is given by

$$\hat{S}(t)=\prod_{j=1}^{k}\exp\left(-\frac{d_j}{n_j}\right) \quad [9]$$

This estimate can be obtained from an estimate of the cumulative hazard function (see in Collett, 2003). Moreover, the KM estimate is regarded as an approximation to the Nelson-Aalen estimate. The Nelson-Aalen estimate of the survivor function has been shown to perform better than the KM estimate in small samples.

Comparison of Survivorship Functions

After providing a description of the overall survival experience in the study, we turn our attention to a comparison of the survivorship experience in key subjects in the data. The simplest way of comparing the survival times obtained from two or more groups is to plot the Kaplan-Meier curves for these groups on the same graph. However, this graph does not allow us to say, with any confidence, whether or not there is a real difference between the groups. The observed difference may be a true difference, but equally, it could also be due merely to

chance variation. Assessing whether or not there is a real difference between groups can only be done, with any degree of confidence, by utilizing statistical tests.

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 lies above another, meaning that he group defined by the upper curve lived longer, or had a more favourable 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. A number of statistical tests have been proposed to answer this question such as Log-rank, Generalized Wilcoxen, Tarone-Ware test and so on.

The calculation of each test is based on a contingency table of groups by status at each observed survival time. The general form of these test statistics for the comparison of survival functions between two groups can be defined as follows:

$$Q=\frac{[\sum_{i=1}^{m} w_i(d_{1i}-\hat{e}_{1i})]^2}{\sum_{i=1}^{m} w_i^2 \hat{v}_{1i}} \qquad [10]$$

where:

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

n_{1i} is the number of individuals at risk in group 1 just prior to failure time t_i

n_{2i} is the number of individuals at risk in group 2 just prior to failure time t_i

n_i is the number of individuals at risk in both groups 1 and 2 just prior to failure time t_i

d_{1i} is the observed number of failure (death) in group 1 at failure time t_i

$\hat{e}_{1i}=\frac{n_{1i}\times d_i}{n_i}$ is the expected number of failures corresponding in group 1 at time t_i

$\hat{v}_{1i}=\frac{n_{1i}n_{2i}d_i(n_i-d_i)}{n_i^2(n_i-1)}$ is the variance of the number of failures in group 1 at time t_i

w_i is the weight for censor adjustment at failure time t_i .

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, then the significance of Q may

be tested using the chi-square distribution with one degree of freedom. We can also use the above test to compare more than two groups (see in Collett, 2003)

The most frequently used test, the log rank test, sometimes called the Mantel-Haenszel test, is the most well known and widely used test, this test is based on weights equal to one, i.e. $w_i = 1$. The log rank test is a non-parametric test for comparing two or more independent survival curves. Since it is a non-parametric test, no assumptions about the distributional form of the data need to be made. This test is however most powerful when used for non-overlapping survival curves. The test can be generalized to accommodate other tests that are equally used sometime in practice such as Generalized Wilcoxon test, Tarone-Ware test, and Peto-Peto-Prentice test. Each of these tests uses different weight to adjust for censoring that is often encountered in survival data.

3.3.2 Modelling of survival data

Through a modelling approach to the analysis of survival data, we can explore how the survival experience of a group of individuals depends on the values of one or more explanatory variables, whose values have been recorded for each individual at the time origin. In most medical studies that give rise to survival data, supplementary information will also be collected on each individual so that the relationship between the survival experience of individuals and various explanatory variables may be investigated.

The hazard function is modelled directly in survival analysis. There are two broad reasons to model survival data. One objective of the modelling process is to determine which combinations of potential explanatory variables affect the form of the hazard function. In particular, the effect that the treatment has on the hazard of failure can be studied, as can the extent to which other explanatory variables affect the hazard function. Another reason for modelling the hazard function is to obtain an estimate of the hazard function itself for an individual.

A variety of models and methods have been developed for doing this sort of survival analysis using either parametric or semi-parametric approaches. Semi-parametric models are models that parametrically specify the functional relationship between the lifetime of an individual and his characteristics (demographic, socio-economic, etc.) but leave the actual distribution

of lifetimes arbitrary. The most popular of the semi-parametric models is the Proportional hazards model. It has the property that the ratio of the hazards depends on the values of their explanatory variables, say, X_1, X2, ... but does not depend on time t.

Cox-proportional Hazards Model

This model was proposed by Cox (1972) and has also come to be known as the Cox regression model. Although the model is based on the assumption of proportional hazards, no particular form of probability distribution is assumed for the survival times.

The set of values of the explanatory variables in the proportional hazards model will be represented by vector X. Let $h_0(t)$ be the hazard function for an individual for whom the values of all explanatory variables that make up the vector X are zero. The function $h_0(t)$ is called the baseline hazard function. The hazard function for the individual can then be written as

$$h_i(t)=h_0(t)\exp(\beta'X) \qquad [11]$$

where, where β is a p x l vector of regression coefficients.

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 times dependent, this assumption may not hold. This can be verified by considering the hazard ratios of different individuals. The logarithm of the hazard ratio for two individuals having two distinct covariate values x_j and x_i can be expressed as

$$\ln\left[\frac{h(t,x_j,\beta)}{h(t,x_i,\beta)}\right]=\ln\left[\frac{h_0(t)\exp(\beta' x_j)}{h_0(t)\exp(\beta' x_i)}\right]=\beta'(x_j-x_i) \qquad [12]$$

Clearly the above ratio is independent of time which means that the log hazard ratio is constant at any given time.

The Cox proportional hazards model can equally be regarded as linear model, as a linear combination of the covariates for the logarithm transformation of the hazard ratio given by:

$$\ln\left[\frac{h(t,x)}{h_0(t)}\right]=\beta' X \qquad [13]$$

The hazard function in the Cox model is called semi-parametric function since it does not explicitly describe the baseline hazard function, $h_o(t)$. The survival function is given by:

$$S(t, x, \beta) = e^{-H(t,x,\beta)} \qquad [14]$$

where, H (t, x, β) is the cumulative hazard function at time *t* for a subject with covariate x. Since we have assumed that survival time is absolutely continuous, the value of the cumulative hazard function is expressed as:

$$H(t, x, \beta) = H_0(t) \exp(\beta' x) \qquad [15]$$

Consequently, from the proportional hazards function, we obtained the survivor function given by:

$$S(t, x, \beta) = [S_0(t)]^{\exp(\beta' x)} \qquad [16]$$

where $S_0(t)$ is the baseline survival function.

Fitting the Proportional Hazards Model

Fitting the proportional hazards model to observed survival data entails estimating the unknown regression coefficients. Also, the baseline hazard function must be estimated. It turns out that these two components of the model can be estimated separately. The coefficients should be estimated first and the estimates are then used to construct an estimate of the baseline hazard function. The regression coefficients in the proportional hazards Cox model, which are the unknown parameters in the model, can be estimated using the method of maximum likelihood.

Suppose the survival data based on *n* independent observations are denoted by the triplet (t_i, x_i, c_i) for i= 1,2,3.....,n among whom there are *r* distinct death times and *n-r* right censored survival times. There is one individual dies at each death time, so that there are no ties in the data. Therefore, the data consist of *n* observed survival times, denoted by $t_1, t_2, \ldots\ldots, t_n$, and

that c_i is an event indicator, which is zero if the i^{th} survival time is right censored, and unity otherwise.

The likelihood function which holds for any censored survival data with generalized hazard function $h(t_i, \beta, x_i)$, which may not assume proportional hazards, will be

$$L(\beta)=\prod_{i=1}^{n} f(t_i,\ \beta,\ x_i)^{c_i} S(t_i,\ \beta,\ x_i)^{1-c_i} = \prod_{i=1}^{n} h(t_i,\ \beta,\ x_i)^{c_i} S(t_i,\ \beta,\ x_i) \qquad [17]$$

Cox showed that the relevant likelihood function which considers the baseline hazard rate as a nuisance parameter; he called it a partial likelihood function, for the proportional hazards model assuming no tied survival times is given by

$$L_p(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp(\beta' x_i)}{\sum_{j \in R(t_i)} \exp(\beta' x_j)} \right\}^{c_i} \qquad [18]$$

where, $R(t_i)$ represents the risk set just prior to time t_i. The corresponding log-partial likelihood function is given by

$$\log L(\beta) = \sum_{i=1}^{n} c_i \left\{ \beta' x_i - \log \sum_{j \in R(t_i)} \exp(\beta' x_j) \right\} \qquad [19]$$

Estimation of the regression parameters

The regression coefficients in the proportional hazards Cox model, which are the unknown parameters in the model, can be estimated using the method of maximum likelihood. The maximum likelihood estimates of the regression parameters in the proportional hazards model can be found by maximizing the log-likelihood function using numerical methods. This maximization is accomplished using the Newton-Raphson procedure (see in Collett, 2003)

By using Newton-Raphson procedure to maximize the partial likelihood function, and let $u(\beta)$ be the $p \times 1$ vectors of first derivatives of the log- likelihood function with respect to the β-parameters. This quantity is known as the vector of efficient scores. Also, let $I(\beta)$ be the $p \times p$ matrix of negative second derivatives of the log-likelihood, so that the (j, k)th element of $I(\beta)$ is

$$-\frac{\partial^2 \log L(\beta)}{\partial \beta_j \partial \beta_k} \qquad [20]$$

The matrix I(β) is known as the observed information matrix.

An estimate of the vector of β-parameters at the *(s+1)th* cycle of iterative procedure, $\hat{\beta}_{s+1}$, is

$\hat{\beta}_{s+1} = \hat{\beta}_s + I^{-1}(\hat{\beta}_s)U(\hat{\beta}_s)$ for $s = 0,1,\ldots\ldots,$

where $U(\hat{\beta}_s)$ is the vector of efficient scores and $I^{-1}(\hat{\beta}_s)$ is the inverse of the observed information matrix, both evaluated at $\hat{\beta}_s$. The process can be started by taking $\hat{\beta}_0=0$ and continue until the change in the likelihood function is sufficiently small.

Thus, after getting the MLE, $\hat{\beta}$, the covariance matrix of $\hat{\beta}$ can be approximated by the inverse of the information matrix, evaluated at $\hat{\beta}$, that is

$$\text{Var}(\hat{\beta}) = I^{-1}(\hat{\beta}) \qquad [21]$$

There are two approaches that incorporate for tied survival times. These are the Breslow and the Efron approximations. In many applied settings there will be little or no practical difference between the estimators obtained from the two approximations. Because of this, and since the Breslow approximation is more commonly available, unless stated otherwise, analysis presented in this study will be based on it.

Variable Selection Procedures

The methods available to select a subset of the covariates to include in a proportional hazards regression model are essentially the same as those used in the other regression models, like purposeful selection, stepwise (forward selection and backward elimination) and best subsets selection.

When the number of variables is relatively large, it can be computationally expensive to fit all possible models. In this situation, automatic routines for variable selection that are available in many software packages might seem an attractive prospect. These routines are based on forward selection, backward elimination or a combination of the two known as the stepwise procedure. These automatic routines have a number of disadvantages. Typically, they lead to the identification of one particular subset, rather than a set of equally good ones. The subsets found by these routines often depend on the variable selection process that has been used, that

is, whether it is forward selection, backward elimination or the stepwise procedure, and generally tend not to take any account of the hierarchic principle. They also depend on the stopping rule that is used to determine whether a term should be included in or excluded from a model.

Thus, instead of using automatic variable selection procedures, the following general strategy for model selection is recommended by Collet (2003).

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.
2. The variables that appear to be important from step 1 are then fitted. In the presence of certain variables others may cease to be important. Consequently, those variables that do not significantly increase the value of $-2\log\hat{L}$ when they are omitted from the model can now be discarded. We therefore compute the change in the value of $-2\log\hat{L}$ when each variable on its own is omitted from the set. Only those that lead to a significant increase in the value of $-2\log\hat{L}$ are retained in 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, one at a time, and any that reduce $-2\log\hat{L}$ significantly are retained in the model. 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 no term in the model can be omitted without significantly increasing the value of $-2\log\hat{L}$, and that no term not included significantly reduces $-2\log\hat{L}$.

When using this selection procedure, rigid application of a particular significance level should be avoided. In order to guide decisions on whether to include or omit a term, the significance level should not be too small. A level of around 20% - 25% is recommended.

3.3.3 Assessment of Model Adequacy

Model-based inferences depend completely on the fitted statistical model. For these inferences to be valid in any sense of the word, the fitted model must provide an adequate summary of the data upon which it is based. Some of the methods for the assessment of a fitted proportional hazards model can be equally used for parametric regression models. The fit of a regression model involves assessment of the regression coefficients and the formation of confidence intervals for the parameters and related quantities. Under the assumption of proportional hazards, there are three different tests for model assessment (the significance of the coefficients): the partial likelihood ratio test, the Wald test and the score test. These tests are presented below as discussed in Hosmer and Lemeshow (1998).

The Partial Likelihood Ratio Test (LR) is the best of the three tests for testing the significance of a subset of q explanatory variables from p explanatory variables, and fit both the unrestricted and the restricted models. We shall obtain the value of the log-partial likelihood function $LL_p\ (\hat{\beta}_{p-q})$ in the unrestricted model and $LL_p\ (\hat{\beta}_p)$ when the model imposes the restrictions under H_0.

$$Q_{LR} = -2LL_P(\hat{\beta}_{p-q}) - (-2LL_p(\hat{\beta}_p)) \qquad [22]$$

The test statistic for H_0 is based on the difference of the log-likelihood values. Under H_0, the statistic is asymptotically distributed as chi-squared with q number of degrees of freedom at a significance level α.

The Wald test: this requires fitting the unrestricted model, and is based on the partial likelihood estimator $\hat{\beta}$. The test statistic is

$$Q_W = \hat{\beta}' I_{q\times q}^{-1}(\hat{\beta})\ \hat{\beta} \sim X_{(q)}^2 \qquad [23]$$

The quadratic form of the above equation requires the inverse of the variance-covariance estimates corresponding to the q parameters in H_0 matrix Iq×q and, under H_0 is asymptotically distributed as chi-squared with a number q degrees of freedom.

The Score Test: The score test statistic, to test H_0 : $\beta q = (0, 0, 0....0)'$ is defined as

$$Q_S = U'(\beta_q, \hat{\beta}_{p-q})I^{-1}(\beta_q, \hat{\beta}_{p-q})U(\beta_q, \hat{\beta}_{p-q}) \quad [24]$$

where $U(\beta_q, \hat{\beta}_{p-q})$ and $I^{-1}(\beta_q, \hat{\beta}_{p-q})$ are the score vectors and inverse of the observed information matrix evaluated at the hypothesized value of β_q and the restricted partial maximum likelihood estimator of β_{p-q}. Under null hypothesis and for large sample Q_s is asymptotically distributed as chi-squared with a number q degrees of freedom.

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 behaviour is known, at least approximately, when the fitted model is satisfactory. The following residuals have been proposed for use by different authors.

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

$$rc_i = \exp(\hat{\beta}' x_i)\hat{H}_0(t_i) = \hat{H}_i(t_i) = -\log \hat{S}_i(t_i) \quad [25]$$

where $\hat{H}_0(t_i)$ is an estimate of the baseline cumulative hazard function at time t_i, the observed survival time that individual, $\hat{H}_i(t_i)$ 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.

In the argument, if the model fitted to the observed data is satisfactory, then the model-based estimate of the survivor function for the i^{th} individual at t_i, the survival time of that individual, will be close to the corresponding true values $S_i(t_i)$.if the observed survival time for an individual is right-censored, the corresponding value of the residual is also right-censored. The residual will therefore be a censored sample from the unit exponential distribution, and a test of this assumption provides a test of model adequacy.

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

$$r_{Mi} = c_i - r_i \quad [26]$$

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 liner 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, reference line.,This residual, differ from those considered previously in one other important respect .this is that there is no a single value of the residual for each individual, but a set of values, one for each explanatory variable include in the fitted cox regression model.

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

$$r_{pji} = c_i\{x_{ji} - \hat{a}_{ji}\}, \quad [27]$$

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

$\hat{a}_{ji}=\frac{\sum_{l\in R(t_i)} x_{jl}\exp(\hat{\beta}'x_l)}{\sum_{l\in R(t_i)}\exp(\hat{\beta}'x_l)}$ and $R(t_i)$ is the set of all individuals at risk at time t_i. [28]

The above residuals have the disadvantages that they depend heavily on the observed survival time and require an estimate of the cumulative hazard function. Schoenfeld proposed residuals that overcome these disadvantages.

I. Testing for 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

and the smoothed plot is a horizontal straight line. This indicates that the covariate is linear in the model.

II. Subject-wise diagnostic measures

It may happen that the structure of the fitted model is particularly sensitive to one or more observations in the data set. Such observations which are referred to as influential observations can be detected using diagnostics that are designed to highlight observations that influence the complete set of parameter estimates in the linear predictor. In other words, it may happen that the structure of the fitted model is particularly sensitive to one or more observations in the data set. Another important aspect of model evaluation is a thorough examination of regression diagnostic statistics to identify which subjects have an unusual configuration of covariates exert an undue influence on the estimates of the parameters and on the fit of the model. In many occasions, the influence that each observation has on the estimated hazard function will be of interest, and it will then be important to identify observations that influence the complete set of parameter estimates in the model.

Suppose that we wish to determine whether any particular observation has an untoward effect on $\hat{\beta}_j$, the j^{th} parameter estimate, j=1,2,.....,p, in a fitted Cox regression model. One way of doing this would be to fit the model all n observations, $\hat{\beta}_j$ is the j^{th} parameter estimate, 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. Suppose that the value of the j^{th} parameter estimate on omitted the i^{th} observation is denoted by $\hat{\beta}_{j(i)}$. Then, the statistic $\Delta_{j(i)}=\hat{\beta}_j - \hat{\beta}_{j(i)}$, 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. An index plot, or a plot of the likelihood displacements against the rank order of the survival times, provides information visual summary of the values of the diagnostic. Observations that have relatively large values of the diagnostic are influential.

Moreover, examining scaled score residuals is helpful in identifying outliers by observing how large the deviation is. The larger the deviation the more distant the residual is to the mean. The plot of the score residuals looks like a basic hourglass shape, fanning out from its narrowest point at approximately the mean of the covariate. The effect of outliers on the

regression model may be easily checked by dropping these points and refitting the regression equation.

III. Methods for testing the assumption of proportional hazards

The proportional hazards assumption is vital to the interpretation and use of a fitted proportional hazards model. This is an assessment of to what extent the two curves are equidistant over time. If hazards are not proportional, this means that the linear component of the fitted model varies with time in some manner. As can be seen in equation [12], the hazard ratio for two individuals in the proportional hazards model is independent of time and is constant over time. Thus the plot of the logarithm of the Kaplan-Meier cumulative hazards function based on different factors may help in assessing the proportional hazards assumption before fitting a Cox model. If this assumption is met, then the plots will be more or less parallel. But, this method will not give any clue if the plots for different categories of covariates cross each other. The other method, which could be used after the fit of the model, is extending the proportional hazards model by defining several product terms involving each time independent variable with some function of time. That is, if the j^{th} time-independent variable is denoted as x_j, then we can define the j^{th} product term as $x_j \times g_j(t)$ where $g_j(t)$ is some function of time for the j^{th} variable. The extended Cox model that simultaneously considers all time-independent variables of interest can be expressed as:

$$h(t,x,\beta) = h_0(t)\exp\left(\sum_{j=1}^{p} \beta_j x_j + \sum_{j=1}^{p} \delta_j x_j g_j(t)\right) \qquad [29]$$

To check the proportional hazards assumption using a statistical test, we consider the null hypothesis that all the δ terms, which are coefficients of the $x_j \times g_j(t)$ product terms in the model, are zero. Usually the function $g_j(t)$ is chosen to be the logarithm of survival time i.e. $g_j(t)=ln(t)$. Under the null hypothesis that all the δ terms are zero, the model reduces to the proportional hazards model(Kleinbaum(2005),Hosmer and Lemeshow (1998)).

Moreover, for greater diagnostic power the scaled Schoenfeld residual is preferred. If the plot of scaled Schoenfeld residuals versus the logarithm of time is a random, smooth, straight line about zero the proportional hazards assumption will be satisfied. Furthermore, the assumption of proportional hazards would be fulfilled if the interaction of logarithm of time with the covariate is

found to be insignificant (meaning the regression coefficients are not time varying). This means that the study covariates have values that remained fixed over the follow-up period.

IV. Overall Goodness of Fit

One method of checking goodness of fit of the model is to use R^2. In proportional hazards regression model as in all regression analyses there is no single, simple method of calculating and interpreting R^2, because in Cox proportional hazards model, R^2 depends on the proportion of the censored observations in the data. A perfectly adequate model may have what, at face value, seems like a terribly low R^2 due to high percent of censored data (Hosmer and Lemeshow, 1998). The measure of goodness of fit R_p^2 based on partial likelihood is given by:-

$$R_p^2 = 1-\left\{\exp\left(\frac{2}{n}\left[l_0 - l_p\right]\right)\right\} \qquad [30]$$

where l_p is log of partial likelihood for the fitted model with p covariates, l_0 is the log partial likelihood for empty/null model, the model with no covariates and n number of subjects. 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 $\hat{H}_r(r_i)$, 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. In addition, results of the Likelihood Ratio, Score and Wald tests use for checking model goodness of fit.

CHAPTER FOUR

STATISTICAL DATA ANALYSIS AND DISCUSSION

4.1. Descriptive survival analyses

The study included 7118 children, who were born during the five years preceding the date of the survey. Summary results for socio-demographic and environmental variables included in this study are presented in Table 1A (Appendix). Of the total of 7118 children included, 3466 were females, 6107 were born in rural part of Ethiopia, and 5127 were breastfed. Among the infant's mothers, 6715 were currently married. The table shows that 926 infant mothers were 15-20 years old, 4987 infant mothers were between 20-34 years old and the remaining 1205 were 35-48 years old when they gave birth. There were 684 households of size 1-3 members, 3660 had household size 4-6 and 2774 of households had more than 6 members. With regard to educational attainment, about 5457 of the mothers and 4255 of the fathers had no education while 1197 of the mothers and 2041of the fathers had primary education and the remaining 482 mothers and 822 fathers had attained secondary education. About 3210 of the households were classified as poor while 2730 had medium income and 1178 were rich. In addition, 1560 households had piped water while 2516 used water from protected source and the remaining 3042 used unprotected source of water.

The graph of the estimate of overall Kaplan-Meier survivor function Table 2A (Appendix) and Figure 4.1 shows that most deaths occurred in the first month and it declined in the later months of follow up. Separate graphs of the estimates of the Kaplan-Meier survivor functions, for different factors, have also been constructed in order to assess whether there is difference in survival experience between different groups of individuals. Most of the graphs did show differences between different categories. The graphs of Kaplan-Meier survival estimates based on different categories of factors are presented in the displays of Figure 1A (Appendix). Individually, relatively significant differences are observed in covariates such as breastfeeding status and mother's education between secondary education and no education. Thus, the breastfeeding status show that the upper curve of the survival functions is for infants, who were breastfed, indicating greater survival experience as compared to infants

who were not breastfed. Moreover, the graphs of the survival functions for infants show infants mother level that have secondary education have greatest survival experience.

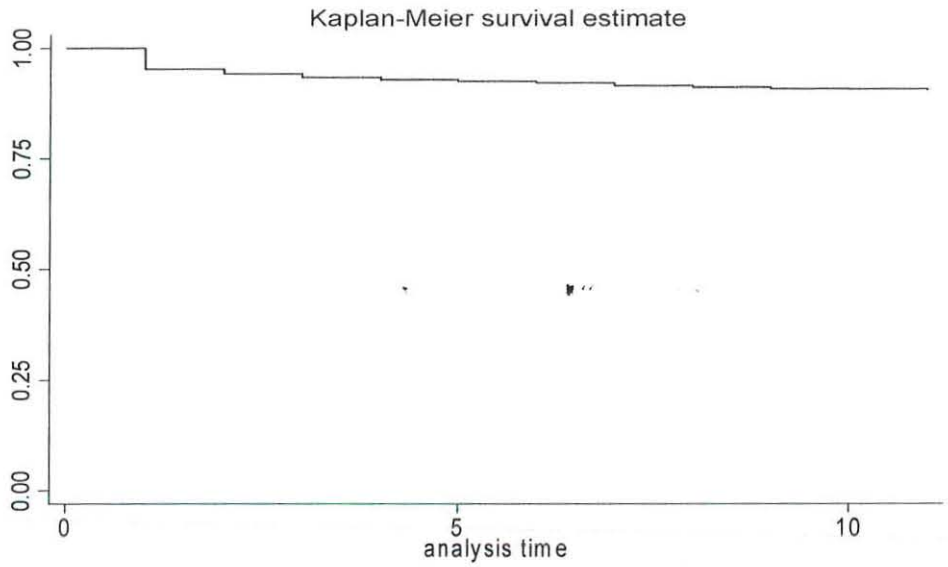


Figure 4.1: The plot of the overall estimate of Kaplan-Meier survivor function

The Log-rank test was performed to investigate the significance of the observed difference in the Kaplan-Meier estimates of the survivor functions among different categories of the factors. This means there is difference in survival experience between two or more levels of the factors. The result obtained from Table 4.1 the p-value of the Log-rank test points out that all factors except marital status have differences in the levels of their survivorship function.

Table 4.1: Results of the Log-rank test for the categorical variables

Covariate / factor	DF	Chi-square	P-Value
Residence	1	13.987	.000
Mother Education	2	33.113	.000
Father Education	2	19.503	.000
Breastfeeding status	1	437.384	.000
Marital status	1	1.048	.306
Child's birth order	2	18.161	.000
Family size	2	15.365	.000
Wealth index	2	12.872	.002
Source of drinking water	2	16.867	.000
Mother's age	2	53.888	.000

Child's sex	1	15.795	.000

4.2. Results of the Cox proportional hazards model

The first step in the model development process is to select explanatory variables which are important for the study. When the number of variables is relatively large, it can be computationally expensive to fit all possible models. Therefore, fit models using appropriate selection method for fitting a multivariable model containing variables that are significant at a modest level of significance in a univariable analysis. Table 3A (Appendix) show that eleven univariable Cox proportional hazards models were fitted.

The result shows that not all explanatory variables are important to fit multivariate Cox proportional hazards model. Thus, the most appropriate covariates will be selected based on their contribution to the maximized log partial likelihood of the model (-2LL ($\hat{\beta}$)). The value of -2LL ($\hat{\beta}$) for the null or empty model is 12623.946. Therefore, inclusion of covariates will be based on the amount of reduction of this value. The bigger the reduction the better the fit. Hence, based on the amount of reduction the log partial likelihood breast status has high contribution to the maximized log partial likelihood of the model (-2LL ($\hat{\beta}$)) because it shows highest reduction in (-2LL ($\hat{\beta}$)). It reduces the value from 12623.946 to 12254.895. This reduction of 369.051 is highly significant (p-value < 0.0001) when compared with percentage points of the x^2distribution on 1 degree of freedom. The second highest reduction in (-2LL ($\hat{\beta}$)) is obtained from mother age, which is included to the null model in its continuous form, is shows significant change. It reduces the value from 12623.946 to 12574.831, which reduced the statistic by 49.115. The next reduction in (-2LL ($\hat{\beta}$)) on including mother education to the null model, this reduction of 41.615 is significant. Then, the reduction in (-2LL ($\hat{\beta}$)) due to inclusion of father education, birth order, source of drinking water, family size, sex, residence, wealth index to the null model successively one at a time are 19.511, 18.531, 17.524, 15.523, 15.454, 15.168 and 13.623 respectively. All of them are significant at 5% level.

At the next step of fitting the Cox proportional hazard model using the variables breast status, mother age, mother education, father education, birth order, source of drinking water, family size, sex, residence, wealth index, and the value of (-2LL ($\hat{\beta}$)) will be 12112.045. Then, omitting variables from the model will be based on the increasing in (-2LL ($\hat{\beta}$)) and p-value.

Table 4A (Appendix) shows that the increasing in (-2LL ($\hat{\beta}$)) and p-value due to omission of variables from the model. The variable residence does not have significant increase in (-2LL ($\hat{\beta}$)) and the p-value is 0.682 when residence is removed from the model. Continuing the fitting processes by eliminating the variable residence, the model consisted of the remaining nine variables is fitted and assess the effect of eliminating variables from the model. Table 5A (Appendix) shows increase in (-2LL ($\hat{\beta}$)) and p-values when omitting variables from the model. The variable family size does not have significant increase in (-2LL ($\hat{\beta}$)) and the p-value is 0.243 when family size is removed from the model. Therefore, the variable family size excluded from the model.

We then fit the remaining eight variables and examine the effect of omitted variables from the model. Table 6A (Appendix) shows that the minimum insignificant increase in -2LL ($\hat{\beta}$) (p-value 0.101) is when wealth index is removed from the model. As a result, wealth index can be excluded from the model which contains the eight variables and another model containing the remaining seven variables will be fitted.

Table 7A (Appendix) shows that when father education is eliminated from the model the increase in (-2LL ($\hat{\beta}$)) is insignificant (p-value 0.080). Thus, the variable father education will not be included in the next model.

Table 8A (Appendix) presents that the comparison of the six variables, which should be included in the model. All the six variables breast status, mother age, mother education, birth order, source of drinking water and sex of parametric estimation are presented in Table 4.2.

Table 4.2: 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

Covariates/ Factors	B	SE	Wald	df	Sig.	Exp(B)	95% CI for Exp(B)	
							Lower	upper
Breastfeeding status	1.473	.077	361.239	1	.000	4.362	3.747	5.077
Mother's age			34.541	2	.000			
15-20	.280	.119	5.538	1	.019	1.323	1.048	1.670
20-34	-.284	.093	9.357	1	.002	.753	.627	.903
Mother education			28.845	2	.000			
No education	1.343	.266	25.579	1	.000	3.832	2.277	6.449
primary	1.099	.279	15.537	1	.000	3.000	1.737	5.181
Birth order			14.477	2	.001			
2-4	.292	.121	5.783	1	.016	1.339	1.056	1.699
>=5	.450	.121	13.807	1	.000	1.568	1.237	1.989
Water			12.332	2	.002			
Protected source	.314	.114	7.584	1	.006	1.368	1.095	1.711
unprotected source	.382	.109	12.233	1	.000	1.465	1.183	1.814
sex	-.221	.076	8.420	1	.004	.802	.691	.931

After fitting the reduced model we assess whether or not the removal of the covariate has produced an important change in the coefficients of the variables remaining in the model. A value of 20% change is generally considered as an important change in a coefficient (Hosmer and Lemeshow (1999)). Thus, the variables marital status, residence, family size, wealth index, father education were included one at a time; the change in the coefficients of the significant variables was examined in Table 9A (Appendix). The maximum change in the coefficient of the variable remaining in the model is 18.89 percent, which is less than 20 percent, for birth order which is judged not to be an important change to warrant inclusion of the design variables of family size in the model. Thus, those non significant variables include those which have less than 20 percent change in the coefficient of the variables. Therefore, the appropriate main effects model contains the variables breast status, mother age, mother education, birth order, source of drinking water and sex.

The plots of the martingale residuals are used to demonstrate the linearity of continuous covariates after excluding the covariate for which we are checking the assumption of

linearity. The scatter plots or the smoothed curve using lowess can be used for this purpose. The plot of martingale residuals for the continuous covariate" ungrouped mother age" in Figure 2A (Appendix) is random showing no systematic pattern and the lowess smoothed curve is approximately a horizontal line. As a result the continuous covariate ungrouped mother age is linear in the model. Since the remaining five covariates are not continuous there is no need for checking linearity.

The final step in model development strategy is consideration of interaction terms that may be useful in the improvement of the model. The significance of each separate interaction is assessed by adding interaction terms to the main effects model one at a time and using the Wald test. Then, examining the p-values of the Wald statistic in Table 10A (Appendix) the interactions between them were not found to be significant. Thus, the last model will be the one which contains only the main effects in Table 8A (Appendix). The parameter estimates and hazard ratios of the covariates are shown in Table 4.2 but the interpretation based on this model should not be made until the important assumptions associated with the proportional hazards Cox regression model has been checked.

4.3. Model diagnostics

The next important step in statistical analysis is model diagnostics, which is assessing the adequacy of the model should be done in order to evaluate how well the fitted regression describes the data set. There are a series of regression diagnostics for the final proportional hazards model. Requirements for model assessment include testing the assumption of proportional hazards, checking for the presence of leverages and influential observations and measuring the overall goodness of fit of the model.

4.3.1. Checking the proportionality of covariates in the model

The proportional hazards assumption should be checked for the interpretation and use of a fitted proportional hazards model. The basic assumption of the Proportional Hazards Model is that the hazard ratios are constant overtime. That means the risk of failure is the same no matter how long subjects have been followed. The adequacy of the preliminary final model is checked for the validity of proportional hazards assumption using test based on the

interaction between variables in the model with logarithm of survival time and assess their significance using the Wald test. Also the plot of the scaled Schoenfeld residuals is also used to provide any additional insight into any departure from proportionality.

Table 4.3 shows that the time-dependent covariates (interaction of covariates with logarithm of time) were not significant and the global fit test also shows that all the covariates were not significant which justifies the proportional hazard assumption holds at 5% level of significance. Therefore, there is no evidence against the null hypothesis that the coefficients of the time varying variables (interaction terms) are zero ascertaining the validity of the proportional hazards assumption for the data. Figures 3A-8A (Appendix) shows the plots of the scaled Schoenfeld for each covariate against log time. All curves seem to approximate the horizontal line through zero. The residuals look random showing no trend with time. This implies that the proportional hazard model fulfills the proportionality assumption.

Table 4.3: Results of the multivariable proportional hazards Cox regression model containing the variables in Table 8A (Appendix) and their interaction with log time.

Covariates/ Factors	DF	Parameter estimate	Error	Chi-square	Pr>chisq	Hazard ratio
Breastfeeding status	1	1.38195	0.19390	50.7936	<.0001	3.983
Mother's age(15-20)	1	-0.01727	0.30582	0.0032	0.9550	0.983
Mother's age(20-34)	1	-0.52467	0.22383	5.4947	0.0191	0.592
Mother Education (no education)	1	1.69011	0.84938	3.9594	0.0466	5.420
Mother Education (primary)	1	1.19995	0.88958	1.8195	0.1774	3.320
Birth order (2-4)	1	0.40104	0.31775	1.5926	0.2069	1.493
Birth order (>4)	1	0.61853	0.31625	3.8253	0.0505	1.856
Source of drinking water(protected)	1	0.84210	0.33888	6.1750	0.0130	2.321
Source of drinking water(unprotected)	1	0.95250	0.32850	8.4073	0.0037	2.592

sex	1	-0.38656	0.19559	3.9060	0.0481	0.679
Breastfeeding status*log time	1	-0.03609	0.12464	0.0838	0.7722	0.965
Mother's age(15-20)*log time	1	0.09861	0.20031	0.2423	0.6225	1.104
Mother's age(20-34)*log time	1	0.15419	0.14732	1.0954	0.2953	1.167
Mother Education(no education)*log time	1	-0.05732	0.50965	0.0126	0.9105	0.944
Mother Education(primary)* log time	1	-0.07466	0.53642	0.0194	0.8893	0.928
Birth order(2-4)*log time	1	-0.11611	0.19549	0.3528	0.5526	0.890
Birth order(>4)*log time	1	-0.20299	0.19606	1.0719	0.3005	0.816
Source of drinking water(protected)*log time	1	-0.21795	0.20685	1.1102	0.2920	0.804
Source of drinking water(unprotected)* log time	1	-0.22471	0.19995	1.2630	0.2611	0.799
Sex*log time	1	0.16579	0.12448	1.7739	0.1829	1.180

Linear Hypotheses Testing Results			
label	Wald chi-square	DF	Pr>chisq
Proportionality_ test	5.7065	10	0.8393

4.3.2. Checking for influential and outlier observations

A thorough evaluation of regression diagnostic statistic to identify, if any, subjects that have undue influence on the estimates of the Cox regression parameters, or have an unusual configuration of the covariates, or have an unexpected influence on the fit of the model is carried out using DFBETA statistic which 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 Collet (2003). The five largest changes in the parameter estimates are presented in Table 11A.

The largest DFBETA for breastfeeding status occurs for child 699. The change in the parameter estimate (DFBETA) on omitting the data for this child is 0.0045. Therefore, omission of this child decreases the hazard of death relative to the baseline hazard. The standard error of the parameter estimate for breastfeeding in the full data set is 0.077, and so the maximum amount by which this estimate changed when one observation is deleted is about 6% of the standard error (less than one standard error). Thus, the change in breastfeeding status effect by deleting this child can be considered as insignificant.

Omitting the data from infant 485 and infant 5655 from the dataset brought the largest changes in the parameter estimates for the other two levels of the baseline mother's age (15-20) and baseline mother's age (20-34), respectively. The maximum change in the parameter estimates for mother's age (15-20) and mother's age (20-34) when each observation is omitted is -0.00811(7% of the standard error) and 0.00742 (8% of the standard error), respectively; both of them are within one standard error of the estimates. The effect of deleting the observation 485 and observation 5655 are increasing and decreasing relative hazard of death, respectively; but again these decreases or increase are not large. Therefore, the changes in mother's age by deleting these observations are insignificant.

The highest change in the parameter estimates (DFBETA) of mother education (no education) and mother education (primary) on omitting the data for 2593 and 1441 children are 0.06967(26% of the standard error) and 0.06866(25% of the standard error), respectively. Thus, the change in mother education effect by deleting these observations can be considered as insignificant.

The differences in the parameter estimates for the levels of the other categorical variables were also assessed by omitting the highest change in the parameter estimates (DFBETA) observations 4215, 5780,4215, 4385, 3784 from the variables birth order(2-4), birth order(>=5), source of drinking water(protected), source of drinking water(unprotected source),sex, respectively. Therefore, omission of the data about these children decreases the hazard of death relative to the baseline hazard, but again these decreases are not great. The differences in the parameter estimates for the levels of the categorical variables were also assessed. However, the largest differences are less than a quarter of the standard error of the corresponding estimate.

In addition, the plots of Score residuals (partial leverage residuals) which have the linear regression leverage property for continuous covariates were considered. The score residuals for the ungrouped mother age in Figure 9A (Appendix) display the fan shape. If the plot fans out from the narrow point that is approximately from the mean of the covariate then it suggests that none of the observations is terribly influential in the study. In the plot there is one point in the top right that falls a bit away from the rest of the points. However, the distance between these points and the others is not striking. The elder mother's age have score residuals that are well within the observed range of values. Therefore, neither the estimates for each of the parameters nor the set of parameter estimates are affected by any of the observations in the dataset.

4.3.3. Assessment of overall goodness of fit

The final step in the model assessment is to measure the overall goodness of fit. All measures depend on the proportion of values that are censored. A perfectly adequate model may have low R^2 due to high percent of censored data. The model in Table 8A (Appendix) presented the value of the -2Log-Likelihood with covariates which is equal to 12124.677 and the -2Log-Likelihood for the null or empty model equals 12623.946. The resulting goodness of fit is calculated as:

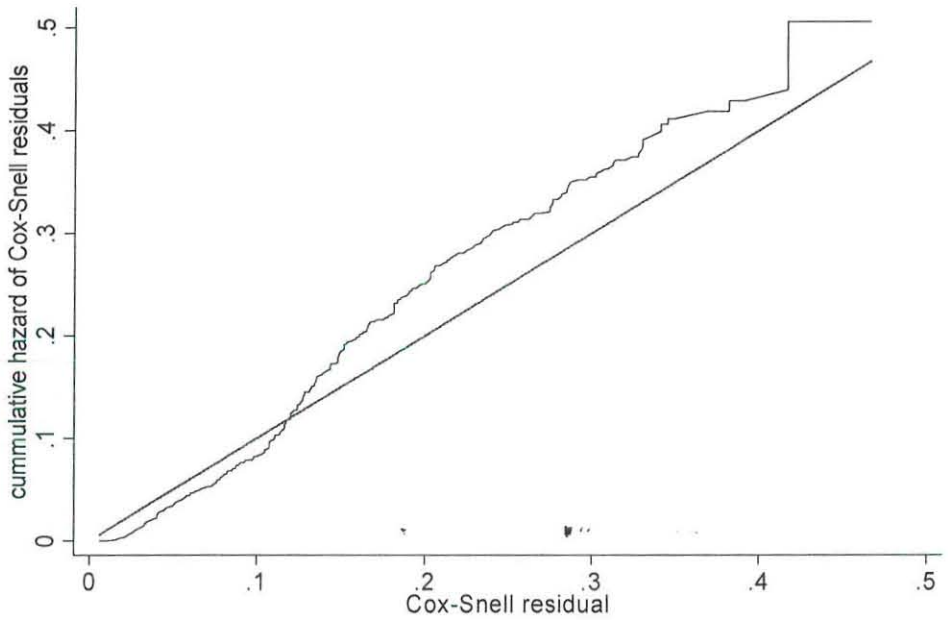
$$R_p^2 = 1-\left\{\exp\left(\frac{2}{n}[l_0 \quad l_p]\right\}\right. = 1-\left\{\exp\left([\frac{-12623.946+12124.677}{7118}]\right\}\right. = 0.0677$$

Also results of the Likelihood Ratio, Score and Wald tests for model goodness of fit displayed in Table 4.4, suggests that model is good fit, i.e. significant at 5% level of significance.

Table 4.4: The Likelihood Ratio, Score and Wald tests for overall measures of goodness of fit of the final model: BETA=0

Test	Chi-square	Df	Pr>chsq
Likelihood ratio	499.2695	10	<.0001
Score	550.2074	10	<.0001
wald	475.9465	10	<.0001

Moreover, Cox-Snell residuals are used to assess the overall goodness of fit of the model. The plot in Figure 4.2 of the cumulative hazard function of the Cox-Snell residual against the Cox-Snell residuals are fairly close to the 45° straight line through the origin. This suggests that the model fit to the data is satisfactory.



The 45°-striaght line through the origin is drawn for reference line.

Figure 4.2: Cumulative hazard plot of the Cox-Snell residuals of the proportional hazards Cox regression model in Table 8A (Appendix).

4.4. Interpretation and Discussion of the results

4.4.1. Interpretation

When the proportional hazards model is used in the analysis of survival data, the coefficients of the explanatory variables in the model can be interpreted as logarithms of the ratio of the hazard of death to the reference group hazard. This means that estimates of this hazard ratio, and corresponding confidence intervals, can easily be found from the fitted model. The interpretation of parameters corresponding to different variables which are found significant in the final model is described in the following section.

Breast status, grouped mother age, mother education, birth order, source of drinking water, sex are the six categorical variables that are found to be significantly associated with the survival of infants in the fitted Cox regression model.

The estimated relative risk (hazard ratio) of dying of infants who were not breastfed as compared to those who were breastfed is 4.362 (95% CI: 3.747-5.077). This means infants who were not breastfed are dying at a rate which is about 4.362 times higher than infants who were breastfed. The 95 % confidence interval also suggests that the risk of death for infants who were not breastfed is 3.747 times as low and 5.077 times as large as compared to those who were breastfed.

The reference category for the mother age group is age >=35. The estimated hazard ratio for the covariate mother age between 15 and 20 is 1.323. This implies that infants who are born to mothers of age group of 15-20 are dying at a rate 32% higher than those who are in the elder age group (age >=35). The confidence interval suggests that the hazard ratios are as low as 1.048 and as high as 1.670. The hazard ratio for the covariate mother age between 20 and 34 is 0.753. The interpretation of this is that infants who are born to mothers of age between 20 and 34 are dying at a rate 25% lower than those who are in the elder age group (age >=35). The 95% confidence interval implies that the rate could be as low as 0.627 and as high as 0.903.

The estimated risks of death for infants whose mothers have no education and primary education compared to those infants whose mothers have secondary education are 3.832 (95% CI: 2.277-6.449) and 3.000 (95% CI: 1.737-5.181), respectively. This means that the hazard rate of death for infants whose mothers have no education and primary education is 3.832 times and 3 times higher respectively than infants whose mothers have secondary education (reference group).

Higher birth orders (>=5) have the highest mortality risk. Infants with these characteristics are 57% more likely to die in infancy relative to the reference group births of order one (HR=1.568, 95% CI: 1.237 to 1.989). Infants of order two through four are dying at a rate 34% higher than infants of order one (HR=1.339, 95% CI: 1.056 to 1.699). The confidence

intervals for higher birth order and birth order two through four indicate that the rate could actually be as high as 1.989 and 1.699 and as low as 1.237 and 1.056, respectively.

The risk of dying for infants born in households with access to unprotected drinking water is higher by 47% relative to those born in households with access to piped drinking water. The estimated risk of death for infants born in households with access to protected source of drinking water compared to those born in households with access to piped drinking water is 1.368 (95% CI: 1.095-1.711, p=<.006). Since the p-value is significant and the confidence interval does not contain 1, an infant born in a household with access to protected source of drinking water has a significantly higher hazard rate, at any given time, than infants born in households with access to piped drinking water (reference group). Thus, infants born in households with access to protected source of drinking water are 37% higher to die in infancy relative to infants born in households with access to piped drinking water.

The hazard ratio for female is 0.802. Thus, female infants have a 20% lower risk rate of death than male infants. The confidence interval indicates that the risk of death for female infants could be as low as 0.691 and as high as 0.931.

4.4.2. Discussion of the results

The results of the analysis presented in this paper, identified variables/factors that are significantly associated with high risk of infant mortality. The paper is important to reducing infant mortality by ensuring the effective implementation of a limited number of significant variables/factors.

In this study the covariates residence, breastfeeding status, mother education, father education, mother age, birth order, family size, wealth index, source of drinking water and sex are significantly associated with increased risk of infant mortality in univariate analysis but in multivariable analysis only breastfeeding status, mother education, mother age, birth order, source of drinking water and sex significantly affect the survival of infants.

This study shows that the risk of infant death is higher among infants who are not breastfed than those breastfed. Demographic research by Mosley and Chen (1984) and by Schultz (1984) also found that the risks of mortally of infants are affected by biomedical factors (i.e.

breastfeeding patterns, hygiene, sanitary measures, and nutrition). A similar study in Kenya by Hisham and Clifford (2008) also found that the most important determinants of infant mortality are breastfeeding status; this study also showed that infants who are not breastfed have high risk of mortality. A study in Malawi by Manda (1999) also investigated the direct and indirect (through its relationship with birth intervals) effects of breastfeeding on children mortality. This result suggests that breastfeeding provides the ideal food for healthy growth and development of infants.

We expect that infant born to young mothers (age between 15 and 20) and those born to older mothers (age above 35 years) would have higher mortality than born to mothers aged 20-34 years. The lower risks of infant death among children those born to mothers aged 20-34 years found in this paper is as the expected mortality pattern. Bicego (1990) in Haiti also showed that low age at birth was found to have marked effects on infant survivorship. The study in Kerala by White (2006) also showed that mother's age has significant effect on infant mortality. Hailemariam and Tesfaye (1997) study in a small urban community in Sebeta also showed that early pregnancy and late pregnancy have a significant negative impact on the livelihood of infants. A Tanzanian study also showed that a teenage pregnancy (15-20 years) was significantly associated with increased infant mortality. Thus, as expected, in this study infants born to mothers aged 20-34 years are at a lower risk of mortality.

There is higher mortality in children whose mothers were not educated or had primary education than children whose mothers were attending secondary education in this paper. The study in rural china by Jacoby and Wang (2003) showed that a higher maternal education level reduces child mortality and that female education has strong health externalities. In Zimbabwe a similar study showed that women's average educational level in their community exerts a great influence on infant survival (Zerai (1996)). A study In Ethiopia by Wang (2003) also showed that female education attainment has significant effect on reducing infant mortality. Therefore, improving the knowledge of females and mothers in the societies are crucial to reduce risk of infant death.

The findings of this study suggest that birth order is a dominant determinant of infant mortality. The higher birth order (>=5) shows a higher risk of dying than birth order one. Hailemariam and Tesfaye (1997) also found that higher birth order have a significant negative impact on the livelihood of infants. In Kenya, Mutunga (2004) found that infant

survival was found better for those who were of birth order 2-3. A similar study in Kerala showed that Infant mortality is found to depend on birth order (White (2006)). In Korea, Kim (2004) study identified birth order is major factor which were associated with infant and child mortality.

The result of this study also shows that infants whose parents use unprotected drinking water have less survival chance than those who use piped drinking water. A study in China showed that access to safe water or sanitation reduces child mortality risks by about 34% in rural areas, which means access to safe water/sanitation, and immunization reduce diarrhea incidence in rural areas (Jacoby and Wang (2003)). A similar study in a small urban community in Sebeta also found that source of drinking water has direct effect on infant mortality (Hailemariam and Tesfaye (1997)). In Kenya, Mutunga (2004) found that child survival was found better for those who had access to safe drinking water and sanitation facilities. Baker (1999) also indicated that source of drinking water and sanitation facilities are strong predictors of infant mortality. A study in Egypt by Hala (2002) showed that access to municipal water decreases sanitary risks. Access to municipal water and improved sanitation facilities had significant positive impact on children mortality (Ali (2002)). Therefore, higher mortality rates are experienced in households that have access to unprotected drinking water.

The lower probability of dying in infancy period for females compared to males found in this study is consistent with many studies all over the world. Likewise, more boys die before their first birth day than girls in Kenya (Hill et al. (2001)). A similar study in Kenya by Hisham and Clifford (2008) also showed that the sex of infants is the significant factors of infant mortality. A study in Ethiopia, Asefa and Tessema (1997) showed that infant mortality rate was higher for males than females.

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

5.1. Conclusions

The Kaplan-Meier results showed that most of the deaths occurred in the earlier month, that is, from birth to one month and it declined in the later months of follow up. About 47.9 % and 58.4% of the deaths occurred in the first and second months of follow up period, respectively. This study has also examined the socioeconomic, demographic and environmental determinants of infant mortality in Ethiopia. Results based on the Cox proportional regression model show that socio-demographic and environmental variables are more important determinants of infant mortality. The findings of the study demonstrate that different factors such as breast feeding status, mother's education, mother age at birth, sex of infant, birth order and source of drinking water have statistically significant impacts on the survival experience of infant. The findings further suggest the following: breastfeeding status and mother's education number have a significant effect on the survival of infants, that is, infants who are breast fed have less risk of death and infants born to mothers who have secondary education are less likely to face the risk of death. Also infants born to mothers of age group 20-34 have lower risk of death. Birth order and sex were also found to have contribution to infant mortality. Infants that are in a higher birth order number have higher risk of infant mortality and infants who are female have less risk of death. The study also shows, that source of drinking water has a significant effect on infant mortality. An infant that comes from a household with access to protected water and unprotected water are associated with higher risk of mortality than an infant that comes from a household with access to piped water.

5.2. Recommendations

The Government of Ethiopia has given high priority for infant survival interventions. This decision has been taken in a context which strongly supports such action. Not only is there powerful international support, but also recent developments in the health and health-related sectors in Ethiopia can provide the practical means for implementing a successful Child Survival Strategy. Moreover, identifying the important of the socioeconomic, demographic

and environmental factors that affect infant mortality, and acting on them is mandatory. Based on our findings, we make the following recommendations:

- Achieving the Millennium Development Goals for child survival in Ethiopia demands focused and coordinated action to improve nutrition, to strengthen health service systems, and to reduce inequities in access to effective interventions against the diseases which kill young children. Thus, breastfeeding has an important role for reduction of infant mortality, women continue to breastfeed for an extended period.

- The policies and efforts have to be put in place to improve women education. Since women are the primary caretakers of infant, they should be empowered through education, so that the health and survival of their infant will be enhanced.

- The government should work closely with both the private sector and civil society to ensure that households have universal access to safe water as this will to a great extend reduce the number of infant deaths. The study shows that access to adequate and safe water can influence infant mortality and, therefore, these major determinants must be addressed in developing sustainable preventive interventions.

References

- Ali, SM. (2001), "Poverty and Child Mortality in Pakistan" Micro Impact of Macroeconomic Adjustment Policies (MIMAP), Technical Paper Series No. 6.
- Asefa M. and Tessema F. (1997). *Infant survivorship and occurrence of multiple births*. A longitudinal community based study. Ethiop. J. Health. Dev.; 11(3): 283-288.
- Baker, R. (1999) "Differential in Child Mortality in Malawi", Social Networks Project Working Papers, No. 3, University of Pennsylvania, USA.
- Becker, G. S. (1981), "Altruism in the Family and Selfishness in the Market Place," Economica 48,1-15.
- Bicego, G. (1990). Trends, age patterns and determinants of childhood mortality in Haiti. [PhD dissertation]. Baltimore: The Johns Hopkins University.
- Casterline J.B., E.C. Cooksey and A.F. Ismail (1989). 'Household income and child survival in Egypt', *Demography,* **26**: 15-35.
- Cleland, J.G. and Van Ginneken, J.K. (1988). Maternal education and child survival in developing countries: The search for pathways of influence. *Social Science and Medicine* 27(12):1357-1368.,doi:10.1016/0277-9536(88)90201-8.
- Collett, D. (2003). Modelling survival data in medical research (Second edition). Chapman and Hall/CRC, London.
- CSA (2005), Ethiopian Demographic and Health survey 2005, Addis Ababa, Ethiopia and Calverton, Maryland USA: Ethiopia Central Statistical Agency and ORC macro
- DHS/WB (2005), DHS Dimensions: A Semi-Annual Newsletter of the Demographic and Health Related Surveys Project; 4(2).
- Ethiopia Demographic and Health Survey, Central Statistical Authority, Addis Ababa, May 2001
- Federal ministry of health (FMOH, 2005).National Strategy for Child Survival in Ethiopia.
- Hailemariam, A and Tesfaye, M(1997) "Determinants of Infant and Early Childhood Mortality in Small Urban Community of Ethiopia using Hazard Model Analysis" Ethiopian Journal of Health Development, 11(3):189-200.
- Hala A. (2002) "The effect of water and sanitation on child mortality in Egypt", Environmental Economics Unit, Department of Economics, Gothenburg University, Sweden

- Heckman, J. J. (2000), "Policies to Foster Human Capital," Research in Economics 54, 3-56.
- Hill, K., G. Bicego and M. Mahy. (2001). Childhood Mortality in Kenya: An Examination of Trends and Determinants in the Late 1980s to Mid 1990s. Accessed from http://www.jhsph.edu/popcenter/publications/pdf/WP01-01.pdf
- Hisham E. M. and Clifford O. (2008). *Socioeconomic Determinants of Infant Mortality in Kenya*: Analysis of Kenya DHS 2003. Journal of Humanities and Social Sciences
- Hobcraft J. (1993). *Women's education, child welfare and child survival: a review of the evidence*. Health Transition Review 3(2):159-173.
- Hondroyiannis, G. and E. Papaetrou (2002), "Demographic Transition in Europe" Economics Bulletin; 10(3): 1-8. ISSN 1934-7227 Volume 2, Issue 2: 1-16.
- Hosmer, D.W. and Lemeshow S. (1998). Applied Survival Analysis. John Wiley and Sons, Inc., New York.
- Jacoby, H. and L. Wang (2003) "Environmental Determinants of Child Mortality in Rural China: A Competing Risks Approach" World Bank, Washington D.C
- Jain, A. (1988) "Determinants of Regional Variations in Infant Mortality in Rural India" In A. Jain and L. Visaria (Eds.), *Infant Mortality in India: Differentials and Determinants,* Sage Publications, Bombay, India.
- Kalemli-Ozcan, Sbnem (2002), "Does the Mortality Decline Promote Growth?" Journal of Economic Growth, July (2002).
- Kim TH (2004), "determinants of Child and Infant Mortality in Korea 1955-1973".
- Kleinbaum, D.G. and Klein, W. (2005). Survival Analysis a self learning text, second edition. Springer Science+Business Media, Inc., New York.
- Lee L.F., M.R. Rosenzweig and M.M. Pitt (1997), "The Effects of Improved Nutrition, Sanitation, and Water Quality on Child Health in High Mortality Populations". Journal of Applied Business Research; 13(1).
- Madise, N. (2003). Infant mortality in Zambia: Socioeconomic and demographic correlates. Social Biology.
- Manda, S.O.M. (1999). Birth intervals, breastfeeding and determinants of childhood mortality in Malawi. *Social Science and Medicine* 48(3): 301-312. doi:10.1016/S0277-9536(98)00359-1.
- *Merriam-Webster online dictionary*. Merriam-Webster. http://www.m-w.com/cgi-bin/dictionary?book=Dictionary&va=infant. Retrieved 2007-03-27.

- *Merriam-Webster online dictionary*. Merriam-Webster. http://www.m-w.com/cgi-bin/dictionary?book=Dictionary&va=neonate. Retrieved 2007-03-27.
- *Millennium Development Goals Indicators*. United Nations 2009 [cited. Available from

  http://unstats.un.org/unsd/mdg/Metadata.aspx?IndicatorId=0&SeriesId=562.
- Mosley, W. and L. Chen. (1984) "An Analytical Framework for the Study of Child Survival in Developing Countries" , *Population and Development Review* 10: 25-45
- Mturi A.J. and Curtis S.L. (1995).*The determinants of infant and child mortality in Tanzania*. Health Policy Plan 10:384-94.
- Mutunga C. (2004). *Environmental Determinants of Child Mortality in Kenya*. Kenya Institute for Public Policy Research and Analysis (KIPPRA), Nairobi, Kenya.
- Ridder G. and I. Tunali (1999) "Stratified partial likelihood estimation", *Journal of Econometrics*, 92: 193-232
- Schellenberg, J.A., R. Nethan, S.Abdulla, O. Mukasa, T.J. Marchant, M. Tanner and C. Langeler (2002), "Risk Factors for Child Mortality in Rural Tanzania" Tropical Medicine and International Health; 7(6): 506-511.
- Schultz, T. (1984) "Studying the Impact of Household Economic and Community Variables on Child Mortality", *Population and Development Review* 10: 215-235
- UNICEF (2009). The state of world children 2009.
- UNICEF. (2006): State of World's Children 2006.
- United Nations Economic Commission for Africa, 1979:22; Newman, 1979.
- Wang, L. (2003) "Environmental Determinants of Child Mortality: Empirical Results from the 2000 Ethiopia DHS", World Bank, Washington D.C
- White, H. (2006), "determinates of infant and child mortality in Andre Pradesh", University library of Munich, German
- WHO (2001), Macroeconomics and Health: Investing in Health for Economic Development. Report of the Commission on Macroeconomics and Health.
- Woldemicael, G. (1988) "Diarrhoeal morbidity among young children in Eritrea: Environmental and socio-economic determinants", Department of Statistics and Demography, University of Asmara, Asmara, Eritrea
- Zerai, A. (1996). Preventive health strategies and infant survival in Zimbabwe. *African Population Studies* 11(1): 29-62.

APPENDIX

Table 1A: Distribution of important socio-demographic and environment characteristics of infants in Ethiopia.

Covariate / factor	Category	Censored	Dead (%)	Total
Residence	Urban	943	68(6.7)	1011
	Rural	5457	650(10.6)	6107
Mother Education	No Education	4855	602(11)	5457
	Primary	1179	101(8.6)	1280
	Secondary and Higher	482	15(3.1)	497
Father Education	No Education	3770	485(11.4)	4255
	Primary	1872	169(8.3)	2041
	Secondary and Higher	758	64(10.1)	822
Breastfeeding status	No	1546	445(22.4)	1991
	Yes	4854	273(5.3)	5127
Marital status	Currently Married	6044	671(10)	6715
	Currently not Married	356	47(11.7)	403
Child's birth order	1	1111	87(7.3)	1198
	2-4	2848	312(9.9)	3160
	>4	2441	319(10.1)	2760
Family size	1-3	636	48(7.0)	684
	4-6	3310	350(9.6)	3660
	>=7	2454	320(11.5)	2774
Wealth index	Poor	2859	351(10.9)	3210
	Medium	2448	282(10.3)	2730
	Rich	1093	85(7.2)	1178
Source of drinking water	Pipe	1445	115(7.3)	1560
	protected source	2256	260(10.3)	2516
	unprotected source	2699	343(11.3)	3042
Mother's age	15-20	793	133(14.4)	926
	20-34	4569	418(8.4)	4987
	>=35	1038	167(3.9)	1205
Child's sex	Female	3166	300(8.7)	3466
	Male	3234	418(11.4)	3652

Table 2A: Results of the Kaplan-Meier Estimates of infant survival function.

Time	Beg.Total	Fail	Net Lost	Survivor Function	Std.Error	[95% conf. Int.]
0	7118	344	22	0.9517	0.0025	0.9464 0.9564
1	6752	75	55	0.9411	0.0028	0.9354 0.9463
2	6622	54	112	0.9334	0.0030	0.9274 0.9390
3	6456	39	129	0.9278	0.0031	0.9215 0.9336
4	6288	23	130	0.9244	0.0031	0.9180 0.9303
5	6135	23	146	0.9209	0.0032	0.9144 0.9270
6	5966	42	125	0.9144	0.0033	0.9076 0.9208
7	5799	19	116	0.9114	0.0034	0.9045 0.9179
8	5664	21	105	0.9081	0.0035	0.9010 0.9146
9	5538	9	121	0.9066	0.0035	0.8995 0.9132
10	5408	12	147	0.9046	0.0035	0.8974 0.9113
11	5249	57	71	0.8948	0.0037	0.8872 0.9018

Table 3A: Results of the univariable proportional hazards Cox regression model

Covariates/ Factors	B	SE	Wald X^2	df	Sig.	Exp(B)	LR Sig.	-2logL
Residence	-.466	.127	13.369	1	.000	.627	.000	12608.778
Mother Education			28.865	2	.000		.000	12582.331
No Education	1.296	.261	24.584	1	.000	3.655		
Primary	1.042	.277	14.194	1	.000	2.836		
Father Education			18.789	2	.000		.000	12604.435
No Education	.384	.133	8.353	1	.004	1.469		
Primary	.055	.147	.142	1	.706	1.057		
Breastfeeding status	1.455	.077	357.930	1	.000	4.284	.000	12254.895
Marital status	-.152	.151	1.018	1	.313	.859	.313	12622.972
Birth order			17.439	2	.000		.000	12605.418
2-4	.319	.121	6.939	1	.008	1.376		
>=5	.493	.121	16.622	1	.000	1.637		
Family size			14.774	2	.001		.001	12608.423
1	-.524	.155	11.461	1	.001	.592		
4-6	-.204	.077	6.976	1	.008	.815		
Wealth index			12.364	2	.002		.002	12610.323
Poor	.422	.121	12.163	1	.000	1.524		
Medium	.374	.124	9.118	1	.003	1.453		
Source of drinking water			16.202	2	.000		.000	12606.422
protected	.349	.112	9.701	1	.002	1.417		
unprotected	.432	.108	16.098	1	.000	1.541		
Mother's age			51.168	2	.000		.000	12574.831
15-20	.027	.116	.056	1	.813	1.028		
20-34	-.529	.092	33.365	1	.000	.589		
sex	-.296	.076	15.260	1	.000	.744	.000	12608.492

Table 4A: Results of the multivariable proportional hazards Cox regression model containing the variables significant at 20% level in the univariable proportional hazards Cox regression model

Covariates/ Factors	df	Wald X^2	Sig.	LR X^2	Sig.
Breastfeeding status	1	362.489	.000	373.252	.000
Mother's age	2	34.729	.000	32.412	.000
15-20	1	6.766	.009		
20-34	1	7.758	.005		
Mother education	2	21.220	.000	28.364	.000
No education	1	20.375	.000		
primary	1	13.971	.000		
Father education	2	5.271	.072	5.433	.066
No education	1	.000	.999		
primary	1	1.879	.170		
Birth order	2	5.219	.074	5.451	.066
2-4	1	3.202	.074		
>=5	1	5.210	.022		
Water	2	8.223	.016	8.599	.014
Protected source	1	4.576	.032		
Unprotected source	1	8.221	.004		
Family size	2	2.757	.252	2.833	.243
1	1	2.642	.104		
4-6	1	.905	.341		
sex	1	8.336	.004	8.421	.004
Residence	1	.167	.683	.168	.682
Wealth index	2	4.201	.122	4.205	.122
poor	1	.640	.424		
medium	1	2.949	.086		

N.B The value of -2log L for the null model is 12623.946

Table 5A: Results of the multivariable proportional hazards Cox regression model after eliminating the variable Residence from the multivariable proportional hazards Cox regression model in Table 4A

Covariates/ Factors	df	Wald X^2	Sig.	LR X^2	Sig.
Breastfeeding status	1	362.300	.000	373.126	.000
Mother's age	2	34.635	.000	32.332	.000
15-20	1	6.689	.010		
20-34	1	7.806	.005		
Mother education	2	22.901	.000	31.796	.000
No education	1	21.926	.000		
primary	1	14.816	.000		
Father education	2	5.465	.065	5.640	.060
No education	1	.001	.982		
primary	1	1.879	.170		
Birth order	2	5.223	.073	5.456	.065
2-4	1	3.202	.074		
>=5	1	5.215	.022		
Water	2	8.785	.012	9.219	.010
Protected source	1	5.191	.023		
Unprotected source	1	8.778	.003		
Family size	2	2.749	.253	2.826	.243
1	1	2.635	.105		
4-6	1	.902	.342		
sex	1	8.766	.003	8.855	.003
Wealth index	2	4.689	.096	4.711	.095
poor	1	.907	.341		
medium	1	3.614	.057		

Table 5A: Results of the multivariable proportional hazards Cox regression model after eliminating the variable Residence from the multivariable proportional hazards Cox regression model in Table 4A

Covariates/ Factors	df	Wald X^2	Sig.	LR X^2	Sig.
Breastfeeding status	1	362.300	.000	373.126	.000
Mother's age	2	34.635	.000	32.332	.000
15-20	1	6.689	.010		
20-34	1	7.806	.005		
Mother education	2	22.901	.000	31.796	.000
No education	1	21.926	.000		
primary	1	14.816	.000		
Father education	2	5.465	.065	5.640	.060
No education	1	.001	.982		
primary	1	1.879	.170		
Birth order	2	5.223	.073	5.456	.065
2-4	1	3.202	.074		
>=5	1	5.215	.022		
Water	2	8.785	.012	9.219	.010
Protected source	1	5.191	.023		
Unprotected source	1	8.778	.003		
Family size	2	2.749	.253	2.826	.243
1	1	2.635	.105		
4-6	1	.902	.342		
sex	1	8.766	.003	8.855	.003
Wealth index	2	4.689	.096	4.711	.095
poor	1	.907	.341		
medium	1	3.614	.057		

Table 6A: Results of the multivariable proportional hazards Cox regression model after eliminating the variable Family size from the multivariable proportional hazards Cox regression model in Table 5A

Covariates/ Factors	df	Wald X^2	Sig.	LR X^2	Sig.
Breastfeeding status	1	362.692	.000	373.536	.000
Mother's age	2	34.012	.000	31.820	.000
15-20	1	6.246	.012		
20-34	1	8.050	.005		
Mother education	2	24.023	.000	33.760	.000
No education	1	23.022	.000		
primary	1	15.582	.000		
Father education	2	5.485	.064	5.659	.059
No education	1	.001	.970		
primary	1	2.025	.155		
Birth order	2	14.070	.001	14.911	.001
2-4	1	5.933	.015		
>=5	1	13.548	.000		
Water	2	8.729	.013	9.162	.010
Protected source	1	5.175	.023		
Unprotected source	1	8.722	.003		
sex	1	8.892	.003	8.984	.003
Wealth index	2	4.559	.102	4.577	.101
Poor	1	.840	.359		
Medium	1	3.473	.062		

Table 7A: 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 6A

Covariates/ Factors	df	Wald X^2	Sig.	LR X^2	Sig.
Breastfeeding status	1	359.527	.000	370.322	.000
Mother's age	2	33.987	.000	31.789	.000
15-20	1	6.291	.012		
20-34	1	8.046	.005		
Mother education	2	25.558	.000	36.448	.000
No education	1	24.538	.000		
primary	1	16.704	.000		
Father education	2	4.903	.086	5.061	.080
No education	1	.049	.825		
primary	1	1.294	.255		
Birth order	2	14.210	.001	15.017	.001
2-4	1	5.725	.017		
>=5	1	13.574	.000		
Water	2	11.315	.003	12.079	.002
Protected source	1	7.348	.007		
unprotected source	1	11.163	.001		
sex	1	8.653	.003	8.741	.003

Table 8A: Results of the multivariable proportional hazards Cox regression model after eliminating the variable Fathedu from the multivariable proportional hazards Cox regression model in Table 7A.

Covariates/ Factors	df	WaldX^2	Sig.	LR X^2	Sig.
Breastfeeding status	1	361.239	.000	372.053	.000
Mother's age	2	34.541	.000	32.443	.000
15-20	1	5.538	.019		
20-34	1	9.357	.002		
Mother education	2	28.845	.000	40.960	.000
No education	1	25.579	.000		
primary	1	15.537	.000		
Birth order	2	14.477	.001	15.299	.000
2-4	1	5.783	.016		
>=5	1	13.807	.000		
Water	2	12.332	.002	13.194	.001
Protected source	1	7.584	.006		
unprotected source	1	12.233	.000		
sex	1	8.420	.004	8.505	.004

Table 9A: Percentage changes in the coefficients of the variables included in Table 8A, when the variables those were not significant in the univariable and multivariable proportional hazards Cox regression model are added one at a time

Covariates/ Factors	Marital status	Residence	Family size	Wealth index	Father education
Breastfeeding status	0.31	-0.16	0.04	-0.48	0.21
Mother's age					
15-20	0.37	-1.71	-3.39	-0.05	-6.82
20-34	0.57	0.78	1.04	-0.17	7.16
Mother education					
No education	-3.08	4.99	1.74	4.07	0.75
primary	-3.02	4.32	2.14	3.70	-3.86
Birth order					
2-4	-0.68	-0.34	16.12	-1.68	0.34
>=5	0.22	0	18.89	0	0.89
Water					
Protected source	-2.79	14.18	-0.63	17.83	0.64
unprotected source	-1.80	8.52	-0.78	12.68	3.24
sex	1.93	4.65	0.66	-0.72	-1.42

Table 10A: Values of Wald statistic and corresponding p-values of possible interaction terms, added one at a time, to the main effects variables included in the model in Table 8A.

Interaction between Covariates/ Factors		DF	Wald	P-value
Breastfeeding status	Mother's age	2	3.555	.169
	Mother Education	2	1.768	.413
	Birth order	2	2.805	.246
	Source of drinking water	2	2.775	.250
	Sex	1	.590	.442
Mother's age	Mother Education	4	2.313	.678
	Birth order	4	9.323	.053
	Source of drinking water	4	7.093	.131
	sex	2	4.738	.094
Mother Education	Birth order	4	1.505	.826
	Source of drinking water	4	7.884	.096
	Sex	2	4.053	.132
Birth order	Source of drinking water	4	1.114	.892
	Sex	2	.735	.692
Source of drinking water	Sex	2	1.858	.395

Table 11A: The five highest differences of the parameter estimates of the variables included in the model in Table 8A when the data value for each infant is deleted from the model

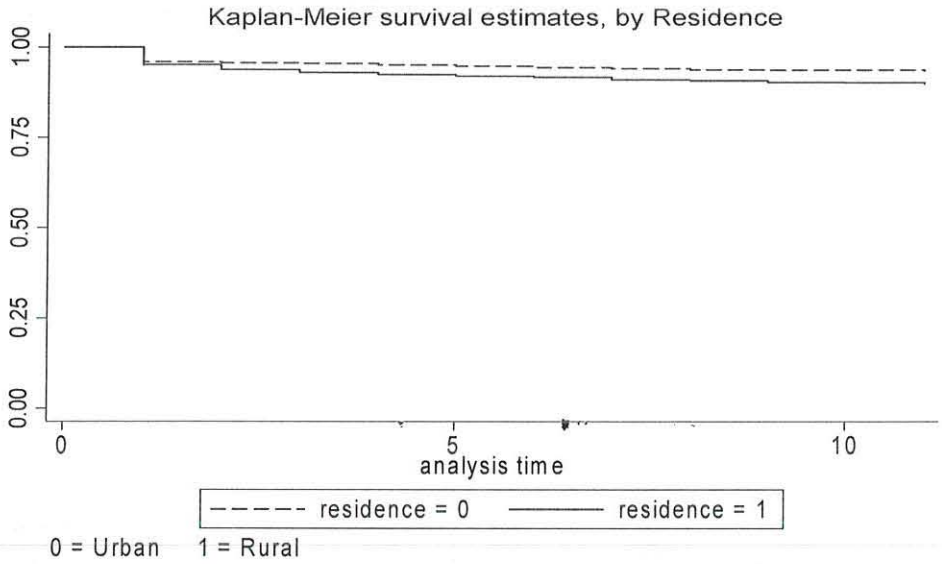
Covariate / factor	Deleted observation (*i*)	$\Delta_{j(i)}=\hat{\beta}_j - \hat{\beta}_{j(i)}$	$\lvert\Delta_{j(i)}=\hat{\beta}_j - \hat{\beta}_{j(i)}\rvert$
Breastfeeding status (No)	699	0.0045	0.0045
	704	0.00442	0.00442
	701	0.00441	0.00441
	20	0.0044	0.0044
	706	0.00426	0.00426
Mother's age(15-20)	485	-0.00811	0.00811
	33	-0.00811	0.00811
	299	-0.00808	0.00808
	29	0.0079	0.0079
	419	-0.0079	0.0079
Mother's age(20-34)	5655	0.00742	0.00742
	768	0.00732	0.00732
	3903	0.00663	0.00663
	7100	0.0066	0.0066
	817	0.00655	0.00655
Mother Education (no education)	2593	0.06967	0.06967
	1980	0.0693	0.0693
	2464	0.06856	0.06856
	5050	0.06634	0.06634
	1796	0.06601	0.06601
Mother Education (primary)	1441	0.06866	0.06866
	5706	0.06864	0.06864
	4274	0.06767	0.06767
	434	0.06611	0.06611
	2053	0.0656	0.0656
Birth order (2-4)	4215	0.01154	0.01154
	2187	0.01148	0.01148
	2046	0.01142	0.01142
	1554	0.01139	0.01139
	3109	0.01139	0.01139
Birth order(>=5)	5780	0.01155	0.01155
	7041	0.01148	0.01148
	1154	0.01143	0.01143
	5329	0.01138	0.01138
	1554	0.01136	0.01136
Source of drinking water (protected source)	4215	0.00981	0.00981
	2187	0.00965	0.00965
	2046	0.00961	0.00961
	3226	0.00958	0.00958
	2269	0.00958	0.00958
Source of drinking water (unprotected source)	4385	0.00948	0.00948
	1643	0.00935	0.00935
	3074	0.00935	0.00935

	3546	0.00931	0.00931
	4581	0.00929	0.00929
Sex(female)	3784	0.00416	0.00416
	556	0.00411	0.00411
	6826	0.00381	0.00381
	4908	0.0038	0.0038
	3446	0.00377	0.00377

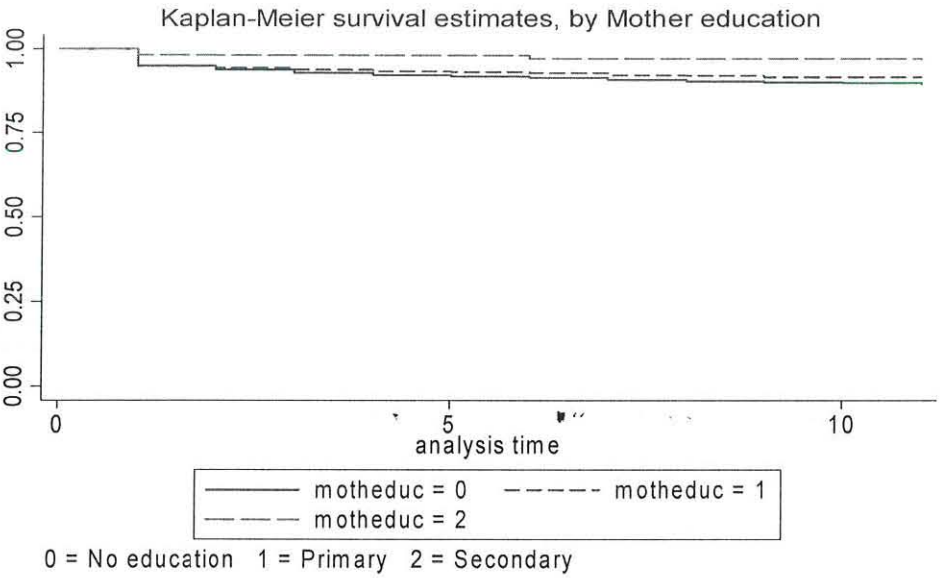
Table 12A: Categorical Variable Coding

		Frequency	(1)	(2)
Mother education	0=No education	5457	1	0
	1=Primary	1179	0	1
	2=Secondary & Higher	482	0	0
Source of Water	0=pipe	1560	0	0
	1=protected source	2516	1	0
	2=unprotected source	3042	0	1
Breastfeeding status	0=No	1991	1	
	1=Yes	5127	0	
sex	0=Female	3466	1	
	1=Male	3652	0	
Mother age	0=15-20	926	1	0
	1=20-34	4987	0	1
	2=>=35	1205	0	0
Birth order	0=1	1198	0	0
	1=2-4	3160	1	0
	2=>=5	2760	0	1

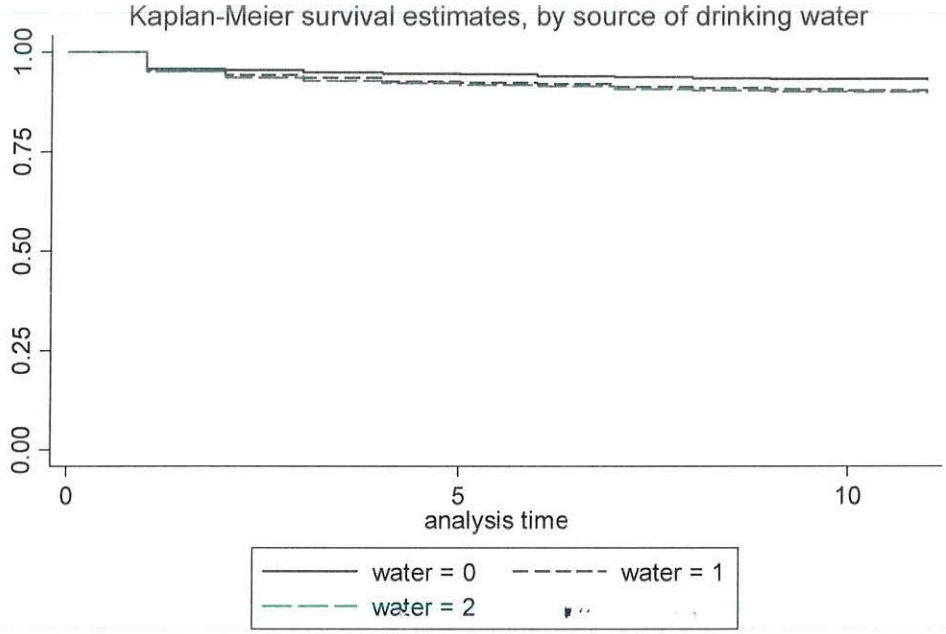
Figures 1A: Plots of Kaplan-Meier survivor functions, based on different factors



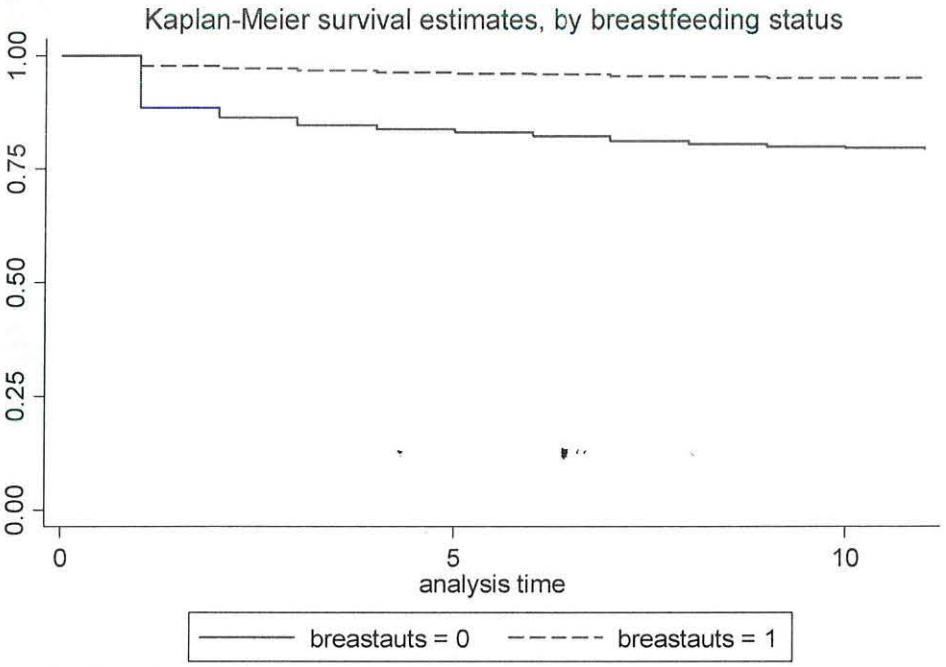
0 = Urban 1 = Rural



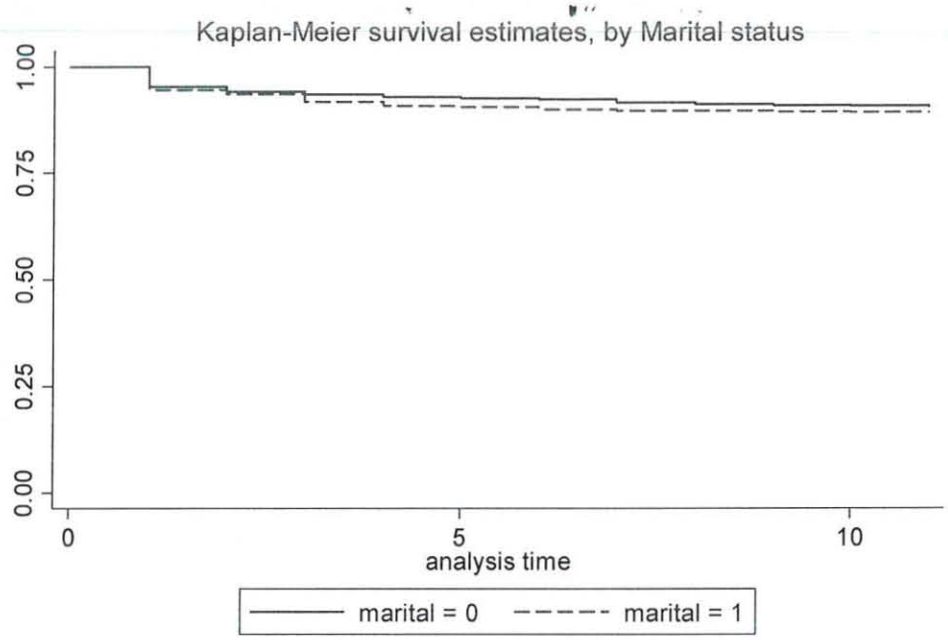
0 = No education 1 = Primary 2 = Secondary



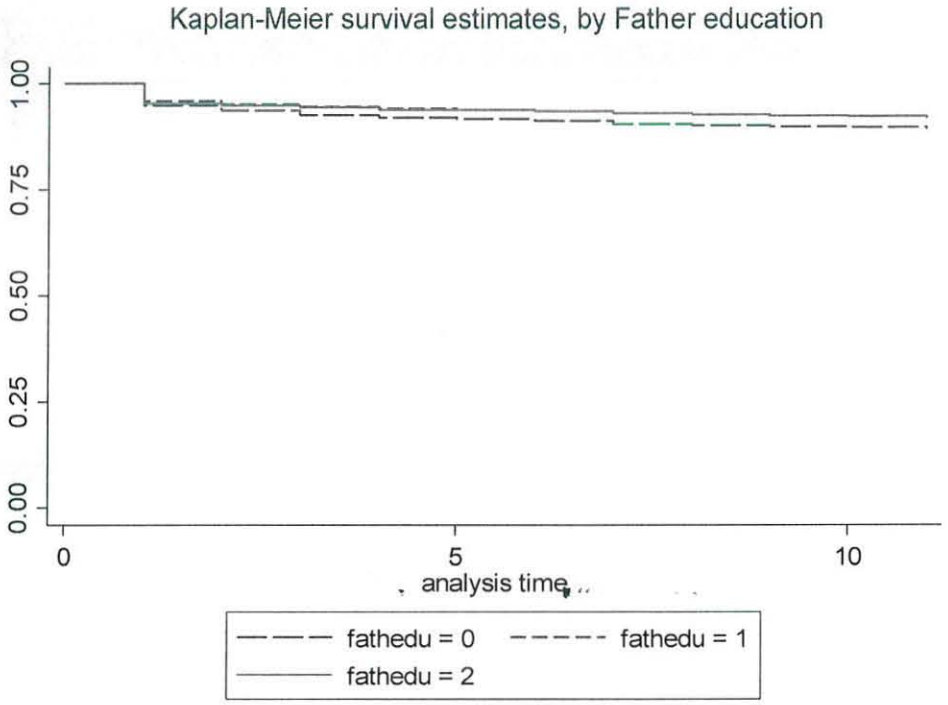
0 = Piped 1 = Protected source 2 = Unprotected source



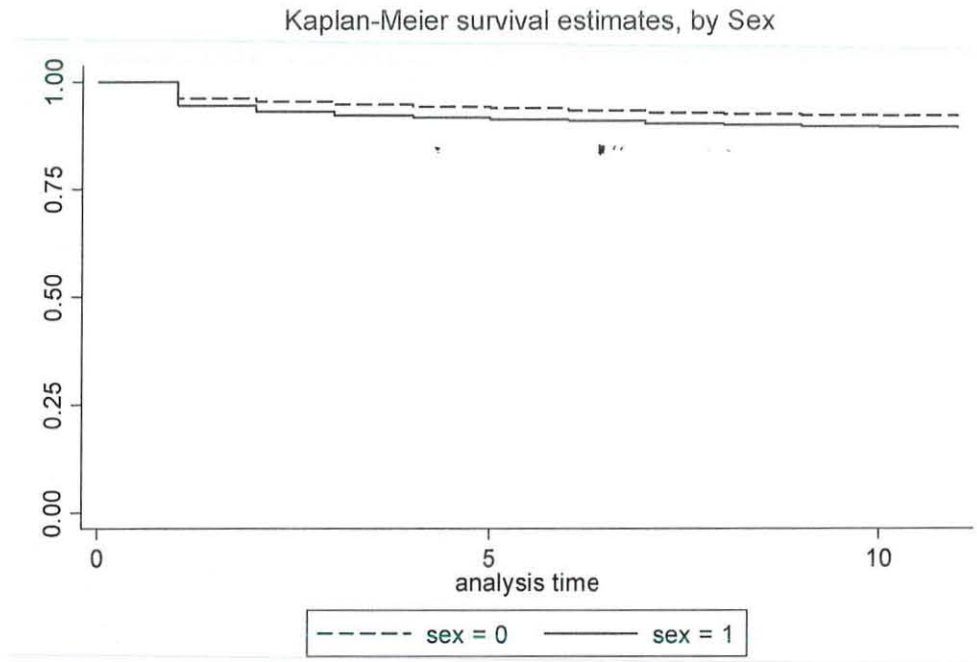
0 = No 1 = Yes



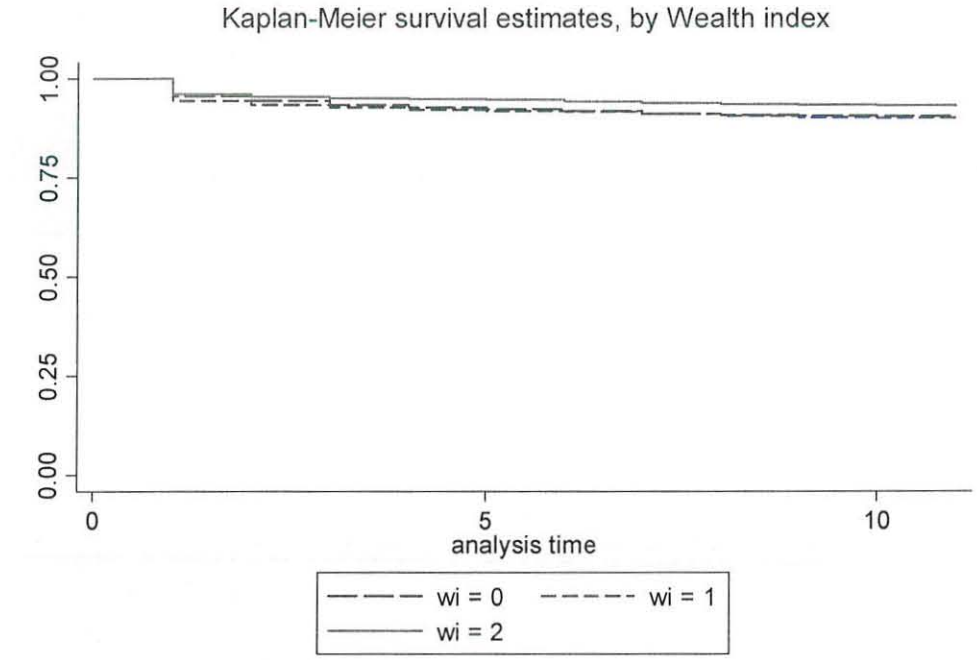
0 = Currently married 1 = Currently not married



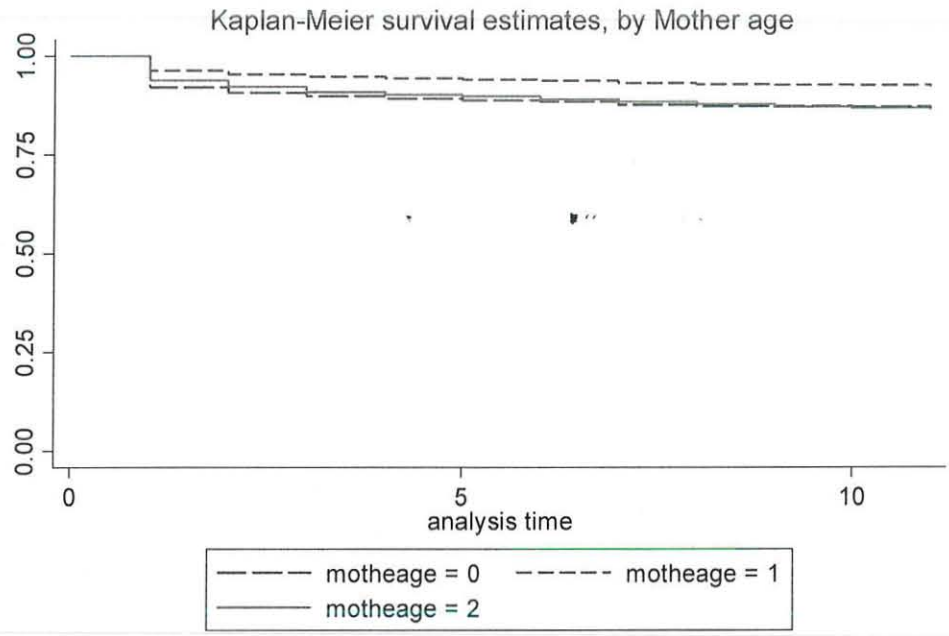
0 = No education 1 = Primary 2 = Secondary



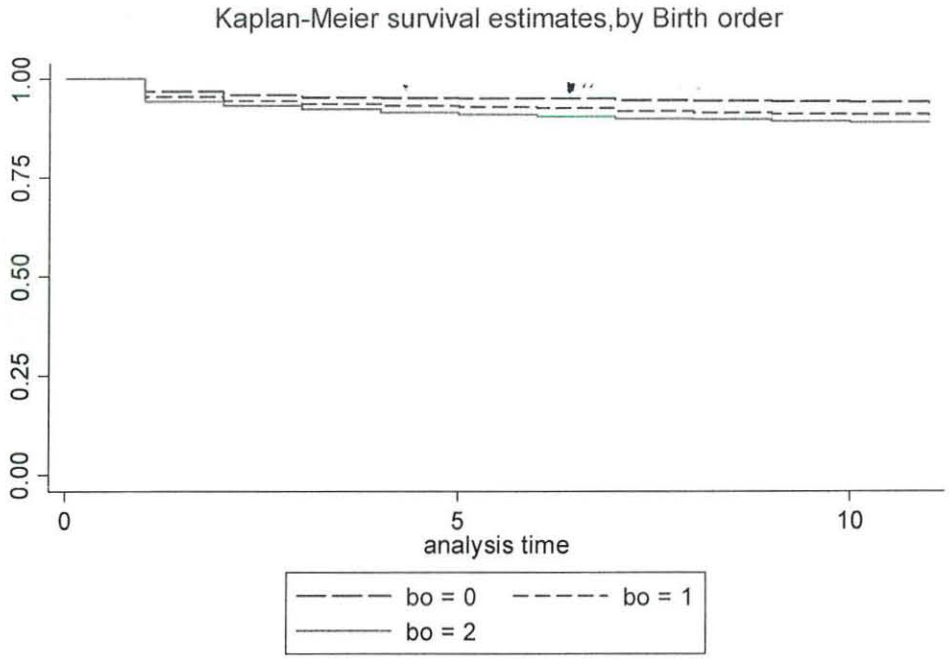
0 = Female 1 = Male



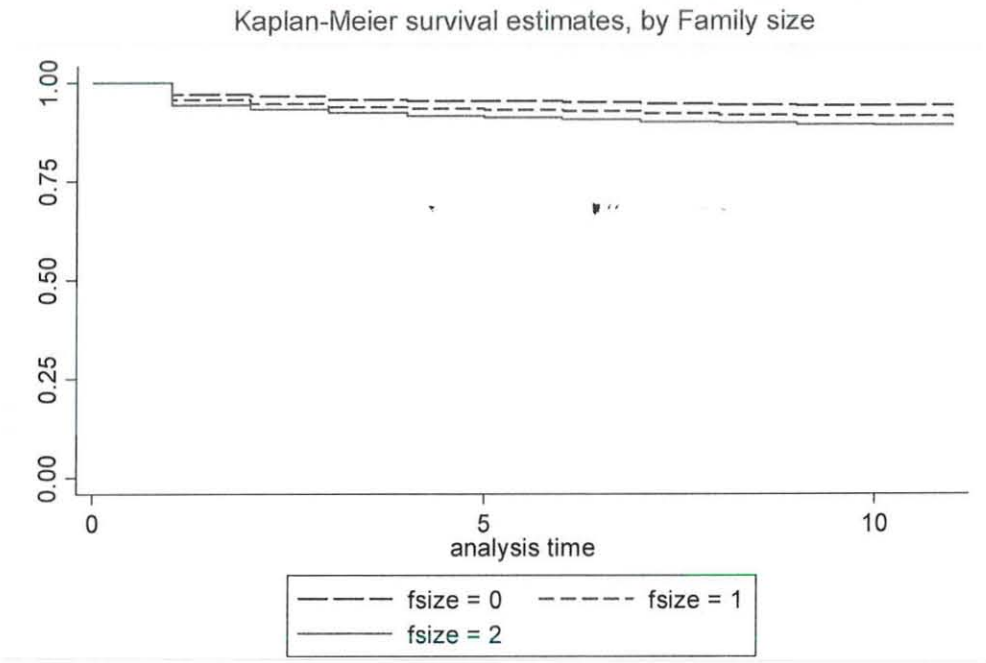
0 =Poor 1 = Medium 2 = Rich



0 = <20 1 = 20-34 2 = >=35



0 = 1 1 = 2-4 2 = >4



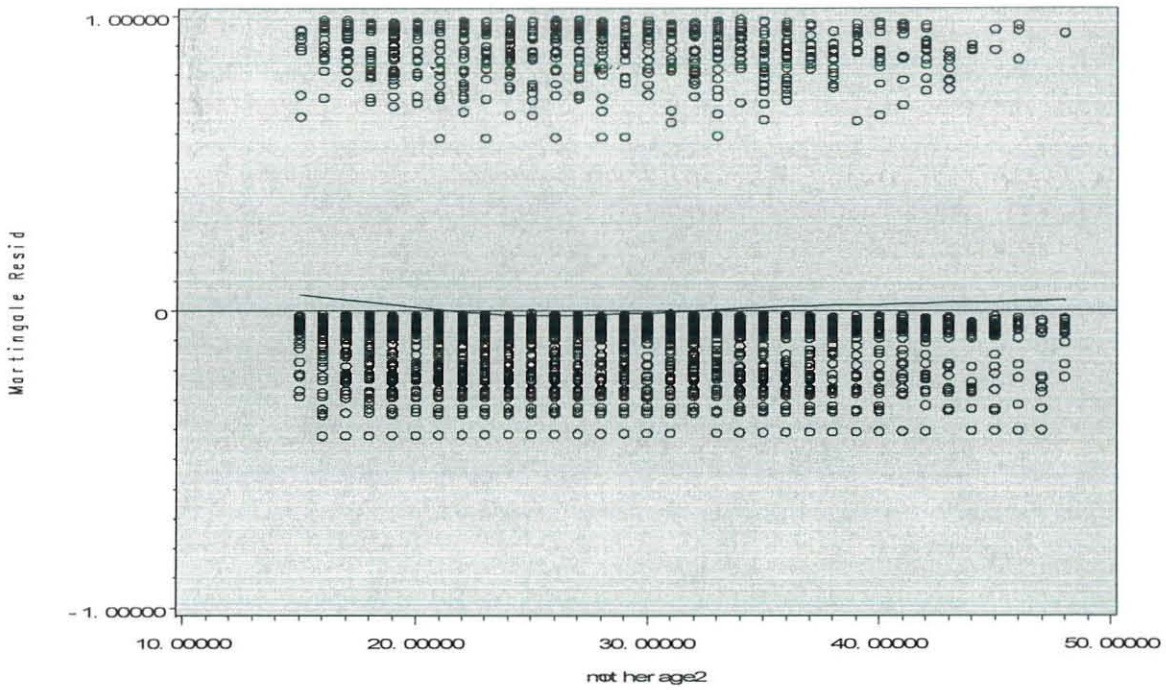
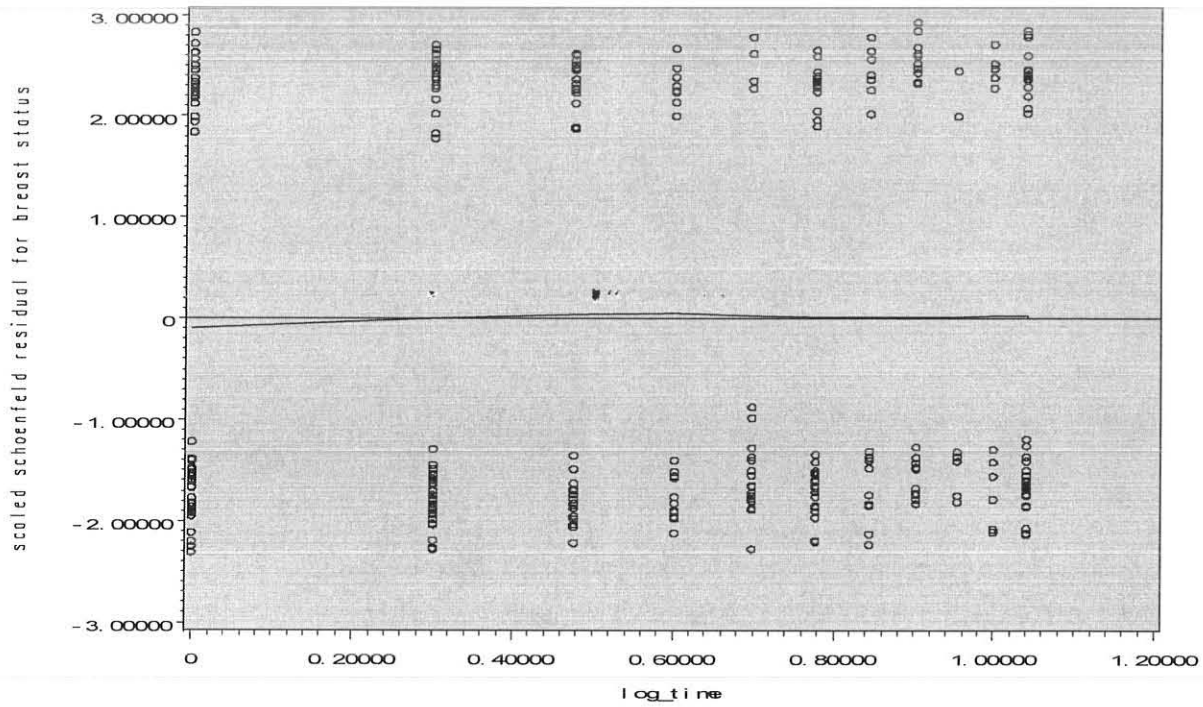
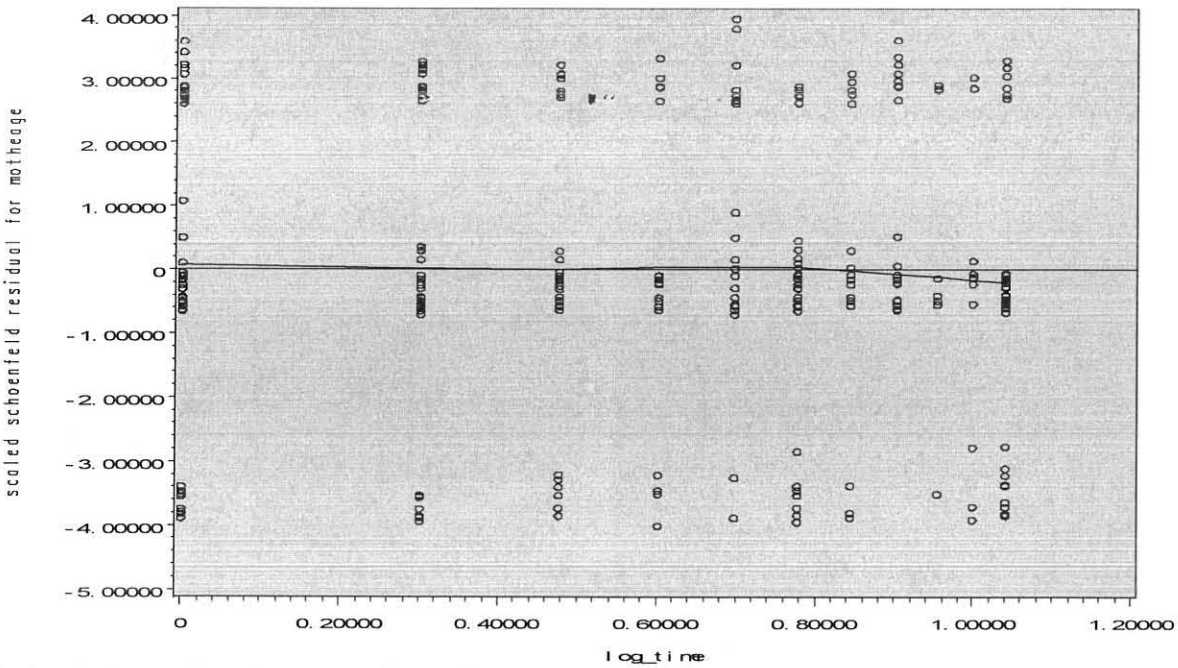


Figure 2A: Plots of Martingale Residuals and Lowess Smoothed Residuals for ungrouped mother age.



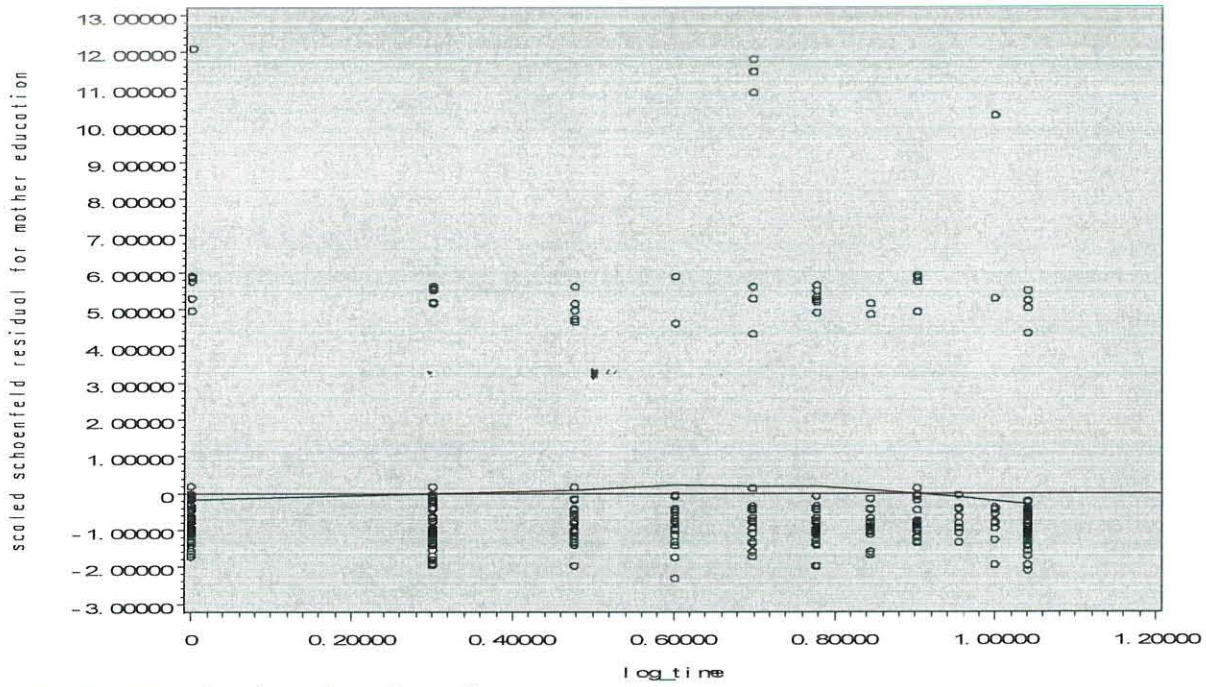
The line that passes through zero is a reference line.

Figure 3A: Plots of the Scaled Schoenfeld residuals and their lowess smooth obtained from the model in Table 8A for the covariate breast status.



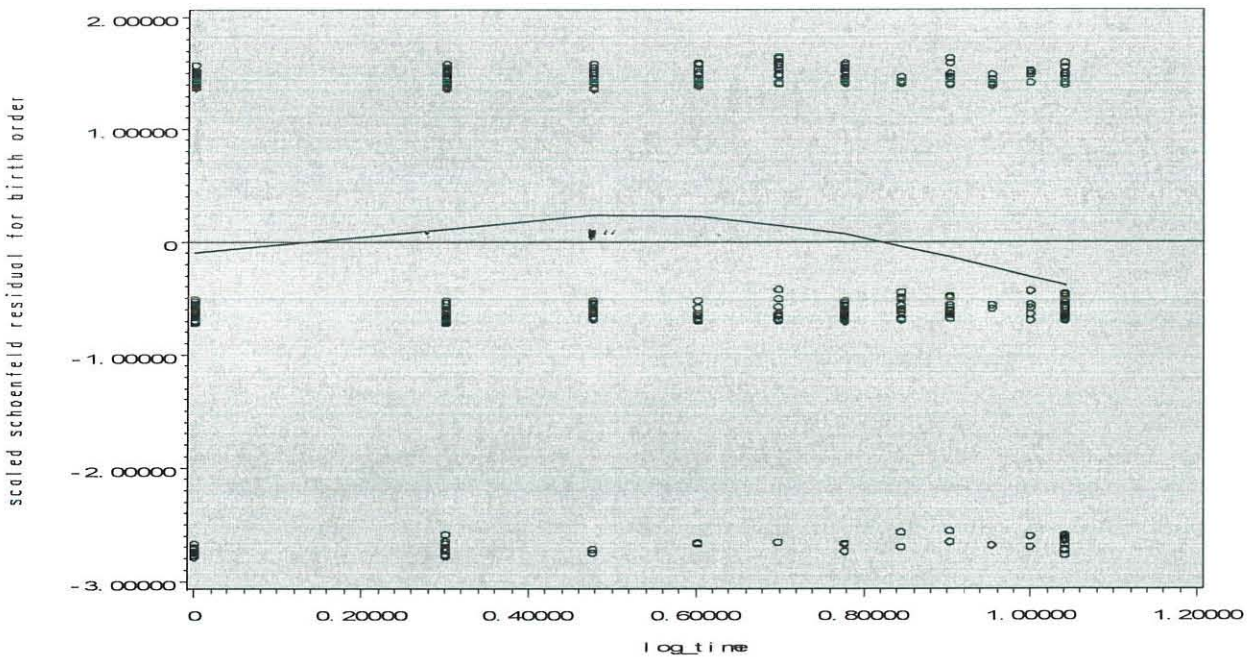
The line that passes through zero is a reference line.

Figure 4A: Plots of the Scaled Schoenfeld residuals and their lowess smooth obtained from the model in Table 8A for the covariate mother age.



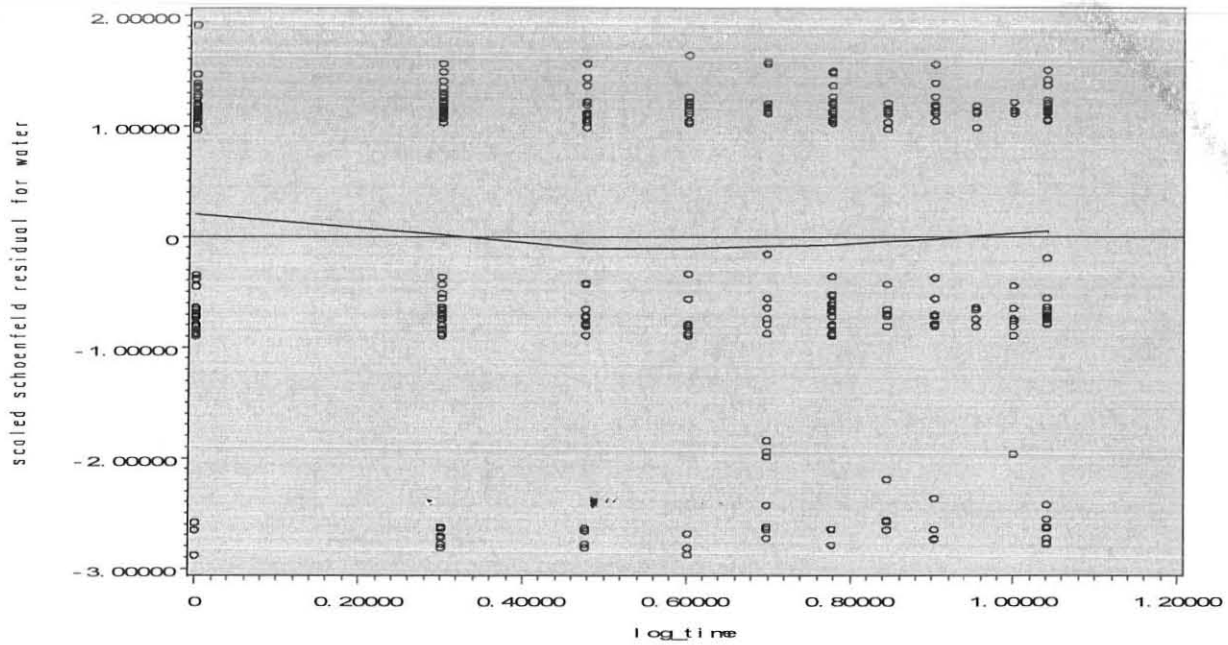
The line that passes through zero is a reference line.

Figure 5A: Plots of the Scaled Schoenfeld residuals and their lowess smooth obtained from the model in Table 8A for the covariate mother education.



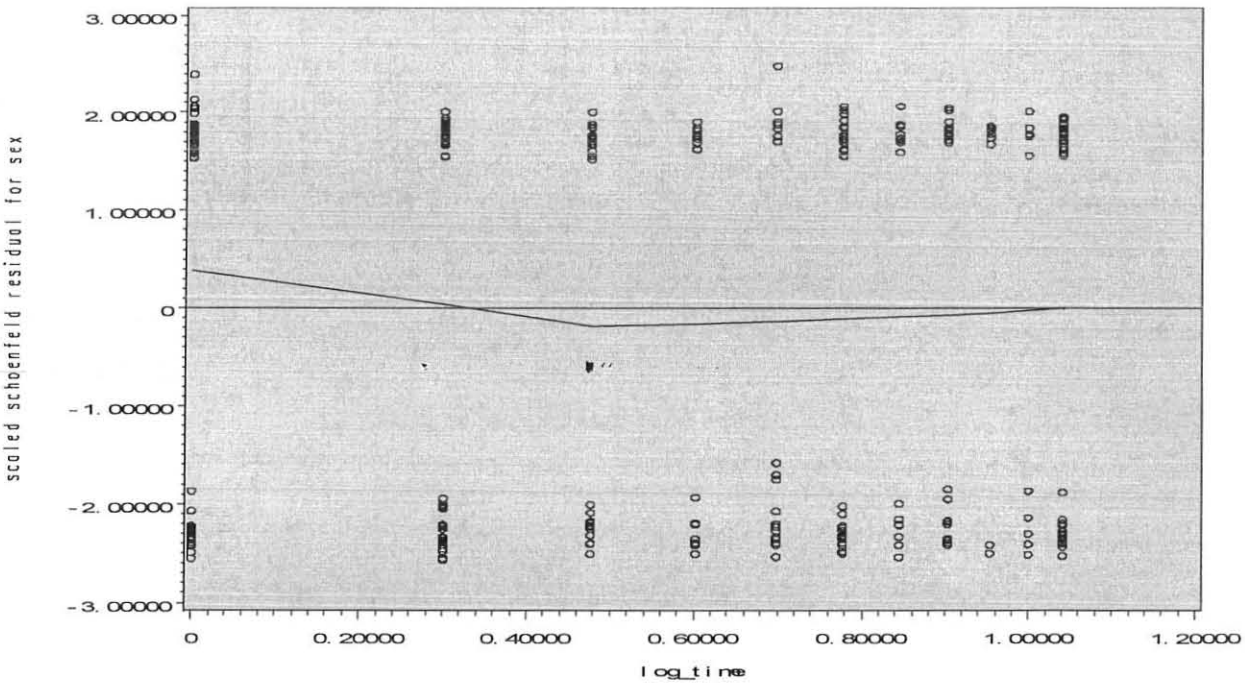
The line that passes through zero is a reference line.

Figure 6A: plot of the Scaled Schoenfeld residuals and their lowess smooth obtained from the model in Table 8A for the covariate birth order.



The line that passes through zero is a reference line.

Figure 7A: Plots of the Scaled Schoenfeld residuals and their lowess smooth obtained from the model in Table 8A for the covariate water.



The line that passes through zero is a reference line.

Figure 8A: Plots of the Scaled Schoenfeld residuals and their lowess smooth obtained from the model in Table 8A for the covariate sex.

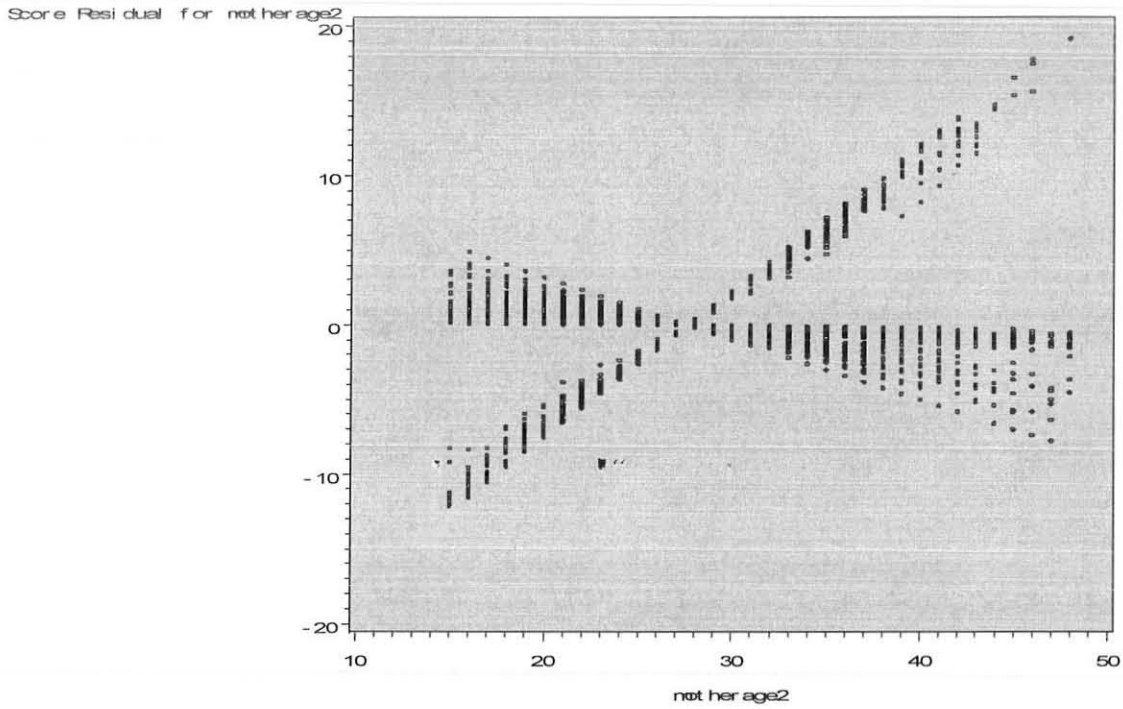


Figure 9A: Plots of the score residuals computed from the model in Table 8A for ungrouped mother age.

I, the undersigned, declare that this thesis is my original work and has not been presented for a degree in any other university, and that all sources of materials used for the thesis have been duly acknowledged.

Declared by:

Name: Samuel Mulinge

Signature: _____

Date: 10.06.2011

Confirmed by Advisor:

Name: Prof. Elizabeth Henda

Signature: _____

Date: 14/06/2011