27

ADDIS ABABA UNIVERSITY

GRADUATE STUDIES PROGRAM

FACULTY OF SCIENCE

DEPARTMENT OF STATISTICS

Survival Analysis of Twin and Singleton infants in South-West Ethiopia

By

Henok Asefa

A Thesis submitted to the Office of Graduate Programs of Addis Ababa University in partial fulfillment of the requirement for the Degree of Masters of Science in Statistics

August 2009

ADDIS ABABA UNIVERSITY

GRADUATE STUDIES PROGRAM

FACULTY OF SCIENCE

DEPARTMENT OF STATISTICS

Survival Analysis of Twin and Singleton infants in South-West Ethiopia

By

Henok Asefa

Approved by the Board of Examiners:

Sileshi Fanta _____

Department Head Signature

_____ _____

Examiner Signature

Sileshi Fanta _____

Examiner Signature

Acknowledgements

My most sincere and heartfelt thanks go to my advisor Dr. Fentaw Abegaz for his interest, devotion, guidance, consultation and encouragement till the end of this study.

I would also like to thank all the members of my family. My sincere thanks go to my father, Ato Asefa Neda, who was the source of especial strength towards the successful completion of the study.

I am extremely grateful to Ato Fasil Tessema for providing me the raw data for use in this thesis.

Finally, my heartfelt gratitude goes to all who were there with me.

ABSTRACT

In this study the effect of socioeconomic, demographic and biological variables on infant survival is investigated using data from "The Jimma infant survival differential longitudinal study" conducted in 46 urban and 64 rural 'kebeles' in South-West Ethiopia during September 1992 and October 1994. The study uses 7,851 singleton and 89 twin infants. Matched pair analysis is used to assess the survival experience between twins, and parametric regression model is used to examine the association between infant mortality and socioeconomic, demographic and biological variables. Data analyses were made using STATA 10 and SPSS 16.

The result of the matched pair analysis showed that, the hazard rates for the first and the second born twins are different at 5% level of significance. The Kaplan-Meier survival estimate for the first born twins is 0.7143 (95% CI: 0.6069 to 0.7971) and for the second born twins is 0.5634 (95% CI: 0.4523 to 0.6604).

In regard to the survival of singleton infants, the result of Kaplan-Meier estimator shows that the survival probability is 0.912 with standard error of 0.0033. That is, infant mortality rate of about 88 deaths per 1000 live births.

Different parametric regression models were fitted to select the significant factors affecting infant mortality in South-West Ethiopia. Almost all the models considered provided good fit for the data but log-normal regression model was found to be the best fit for the data.

Based on the result of the final log-normal regression model, infant mortality was significantly ($p < 0.05$) associated with birth weight, sex of the infant, mother's education, family size and marital status and the interaction between ethnicity and place of residence.

Contents

CHAPTER ONE

INTRODUCTION

1.1Background of the study and statement of the problem

"Infant" is derived from the Latin word "infans" meaning unable to speak. Thus, many define infancy as the period from birth to approximately one year of age, when language begins to flourish. It is an exciting period of "firsts"- first smile, first successful grasp, first evidence of separation anxiety, first word, first step, first sentences and so on. The infant is a dynamic, ever-changing being who undergoes an orderly and predictable sequence of neurodevelopment and physical growth. This sequence is influenced continuously by intrinsic and extrinsic forces that produce individual variation and make each infant developmental path unique. Intrinsic influences include the child's physical characteristics, state of wellness or illness, temperament, and other genetically determined attributes. Extrinsic influences during infancy originate primarily from the family, the personalities and style of care given by parents and siblings, the family's economic status with its impact on resources of time and money, and the cultural milieu into which the infant is born.

The survival of infants in their first year of life is commonly viewed as an indicator of general health and wellbeing of a population. The infant mortality rate (IMR) defined as the risk for a live born child to die before its first birthday is known to be one of the most sensitive and commonly used indicators of the social and economic development of a population (Masuy-Stroobant, 2001). Infant mortality rates are divided into two periods: neonatal and post-neonatal. 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 (23 percent).

Low birth weight and maternal complications in labor contribute significantly to neonatal mortality. In addition, there is a synergistic interrelationship between poverty and increased neonatal death. Post-neonatal mortality is death occurring between 28 and 365 days of life. Post-neonatal deaths are associated with environmental, social, and economic conditions and events that occur after delivery.

The association between deprivation and poor survival in infancy was already documented with survey data as early as 1824 (Masuy-Stroobant, 2001). The association between socio-economic factors and infant mortality was further reinforced when improvements in overall infant mortality levels over time ran parallel with general social and economic development in most industrialized countries during the twentieth century.

Even though death is a biological event mainly caused by a specific disease, the demographic study of the determinants of infant and child mortality will concentrate on the cultural, environmental, social and behavioral factors, which may influence the likelihood of ill health, disease and death in early infancy. Researches on the historical decline of infant and child mortality in Europe have thus identified retrospectively a wide series of determinants which are also known to explain the present-day situation in high mortality populations. Climatic and seasonal variations in mortality by diarrhea have shown the importance of ecological conditions; significant spatial correlations between regional IMRs and infant feeding practices were also abundantly documented; social factors as indicated by the excess mortality of illegitimate infants, or the striking rural-urban differences observed during the industrialization process played also an important role in European history. Finally, the high fertility patterns we have known did also exert an effect on infant and child survival, through shortened birth intervals, family size, etc. IMR has a strong link to fertility and life expectancy at birth. Increased survival of infants and young children is generally accompanied by a drop in fertility.

In fact, so many proximate determinants may be directly influenced by a mother's education to radically alter chances for child survival. Hence the wealth of results showing the positive association between mother's education and survival chances of her children: Caldwell (1979) states that « ... the most important finding at the individual level was the extraordinary stability of the relationship between maternal education and child survival across the different continents and across enormous differences in societal levels of education and mortality », and those findings hold true even after controlling for other socio-economic variables.

Each year four million babies die worldwide in the first four weeks of life. Ninety-nine percent of those deaths occur in low- and middle-income countries. According to the Lancet Neonatal Survival Steering Team in 2005, worldwide neonatal mortality cannot be reduced substantially without focusing on African countries (Lawn et al, 2005). The socioeconomic status of developing countries adversely affects maternal-child health because it limits access to adequate nutrition, quality health care, medications, safe water, adequate sanitation, and other basic social services. The factors associated with high infant mortality rates transcend national boundaries, making infant mortality a critical global health problem.

Historically, perinatal death rates (perinatal deaths include pregnancy losses of at least seven months' gestation (stillbirths) and deaths to live births within the first seven days of life (early neonatal deaths)) and morbidity have been higher in twin pregnancies than in singleton births. Asefa and Tessema (1997) in their study reported that infant mortality of 446.8 per 1000 among multiple-births, in South-West Ethiopia, is over four folds compared to singletons. This is a very high death toll and a matter of serious concern. Moreover, the second twin has been regarded as having a higher risk of a poor outcome. While the first-born twin is less likely to experience neonatal and perinatal complications and most likely to be heavier than the after coming twin (the second-born twin). Birth order effect is more extreme when the first-born twin weighs less than the second-born twin (Hacking et al, 2001). The effect of birth order was marked even when other factors were controlled, such as birth weight, infant sex, presentation and mode of delivery (Hacking et al, 2001).

Poverty is one of the most important factors affecting the infant mortality rate in Africa. Ethiopia is one of the poor African countries with, according to UNICEF (2009) report, a Gross National Income per capita of about $220 in 2007. Ethiopia has a correspondingly high infant mortality rate, estimated for 1991 at 130 death per1000 live birth (World Bank, 1993).

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. Data from the 2005 EDHS of Ethiopia show that infant mortality has declined by 19 percent over the 15-year period preceding the survey from 95 deaths per 1,000 live births to 77. Under-five mortality has gone down by 25 percent from 165 deaths per 1,000 live births to 123. The corresponding declines in neonatal and postneonatal mortality over the 15-year period are 15 percent and 22 percent, respectively. It is not surprising that there are various estimates of infant mortality in a country where there is no vital events registration. The variation mainly emanates from methodological difference in the study. According to the 1994 national census of Ethiopia, the country level estimated infant mortality rate was 116 deaths per1000 live births. In south-west Ethiopia, the estimates are 153 deaths per1000 live births for Kefecho Shekacho, 120 deaths per1000 live births for Illubabbor and 147 deaths per1000 live births for Jimma zones and 93 deaths per1000 live births for Jimma town (CSA, 1996). Shamebo et al (1993) in a cases-referent study at nine rural and one urban kebeles, reported an infant mortality rate of 136 deaths per1000 live births. Asefa and Tessema (1997) in an epidemiological survey based on one-year birth cohort presented an overall infant mortality rate of 105.8 deaths per1000 live births; the mortality rate was calculated using life table estimate. All the above findings show high infant mortality rates in Ethiopia.

A study based on sixty less developed countries from UN statistical yearbook showed that good nutrition and the presence of informally trained health care personnel are more significantly related to low rates of infant mortality (Jayachandran and Jarvis, 1988). Child health researchers agree that the cause of infant mortality in developing countries is multi-factorial. The baby's survival depends on the interaction of socio-economic, biological, behavioral and environmental factors. Hence to understand infant survival one has to see the

relations and interactions among the different socio-economic, biological, behavioral and environmental webs.

Understanding the epidemiological profile of infant death and the capabilities of health systems in dealing with childhood health problems is essential before appropriate public health interventions can be developed and implemented. Development of these interventions also requires an understanding and identification of the determinants of infant mortality. In order to understand and prioritize these determinant factors, there is a need to specify an appropriate model and method of analysis. Most often Cox proportional hazards regression model is employed in survival data analysis as it requires less stringent assumptions as compared to fully parametric regression survival analysis models. However, the most important assumption to be checked with regard to Cox-regression model is the assumption of proportional hazards. If one of the independent variable violates this assumption we need to look for another approach such as stratification or using parametric regression model to study survivorship. Thus, the purpose of this study is to identify determinants of infant mortality that would help in planning and implementing interventions to its reduction using parametric regression model.

1.2 Objectives of the study

General objective

The general objective of this study is to assess the socio-economic, demographic and biological factors that determine the survival of infants in South-West Ethiopia.

Specific objectives

The specific objectives of this study are

- To assess the extent of difference in survival experience between twins.
- To study the distribution of survival times of singleton infants.
- To investigate factors that determines the survival of infants in South-West Ethiopia.

1.3 Application of results

- Since the study has involved a large sample size and varied ethnic groups in South-West Ethiopia, and samples from both urban and rural settings, and wider geographical areas, the findings will contribute to improve child care and health in the part of the country were the study was undertaken.
- The result of this study provides information to government and other concerned bodies in setting policies, strategies, and further investigation for reducing infant mortality.

1.4 Limitations of the study

- The principal limitation of the study is that the number of deaths on which inference is based is very small relative to the censored observations.
- Also, an additional problem is that due to missing values related to dead infants, the analysis is based on smaller set of data for which complete information on survival times are available.

CHAPTER TWO

LITERATURE REVIEW

Each year, about 7.5 million babies born in Sub-Saharan Africa, Asia, Latin America, the Middle East and North Africa die before their first birthday. On average, 61 babies out of 1,000 live births die in developing countries, compared with eight deaths per 1,000 in developed countries before their first birthday; in some developing countries, the rates are much higher than the average. For example, in Sub-Saharan Africa, the world's poorest region, more than one in 10 infants die before age one in Benin, Burkina Faso, Central African Republic, Chad, Ethiopia, Guinea, Malawi, Mali, Mozambique, Niger, Tanzania and Zambia. Per capita income is below $2,000 in all of these countries; in many, it is $500–900. In countries where per capita income is higher, infant mortality rates are substantially lower. High infant mortality is, therefore, clearly a function of poverty. Poverty creates conditions, for example, lack of clean water, poor sanitation, malnutrition, endemic infections, poor or nonexistent primary health care services and low levels of spending on health care, in which babies who are not robust at birth do not receive the health care they need to overcome their vulnerability (AGI, 2002).

Perinatal and neonatal deaths are the most significant contributors to infant's mortality and reduced life expectancy. Perinatal mortality in the developed world indicates a rate of 10 per 1000 still births and live births or less, while in developing countries perinatal mortality range from 35 per 1000 still births and live births to 100 per 1000 still births and live births (Kulmala, 2000). In Ethiopia perinatal mortality has declined from 52 per 1000 still births and live births in 2000 to 32 per 1000 still births and live births in 2005 (EDHS, 2005).

Many literatures reported that, the second twin has been regarded as having a higher risk of a poor outcomes compared with the first-born twin. Birth order effect is more extreme when the second-born twin is heavier than the first-born twin which is a rare case (Hacking et al, 2001).

A study in Canada revealed that the second twin was at significantly greater risk of composite adverse outcome than the first twin. Second twins were more likely to experience perinatal adverse outcomes than first born co-twins, whether delivered at 37 weeks or more, 34–36 weeks, or less than 34 weeks of gestation. The result of this study confirmed that the second twin was more likely to suffer adverse perinatal outcome independent of birth weight, infant sex, presentation, or mode of delivery (odds ratio [OR] 3.77, 95% confidence interval 2.31–6.16) (Anthony et al, 2006).

Neonatal deaths (death in the first month of life) are generally associated with elements linked to maternal care during pregnancy and delivery while socio-environmental factors become more important determinants of infant survival during the post-neonatal period. It is estimated that neonatal deaths can account for nearly 50-60% of all infant deaths in developing countries. Fifty percent of infant deaths in Ethiopia occur during the first month of life (EDHS, 2005). Low socio economic status reflected by lack of education, low utilization of health services and poor environmental sanitation leads to poor pregnancy outcomes (stillbirth, low birth weight, early and late neonatal mortalities) (Karsten, 2002).

A study based on the 2003 Kenyan DHS by Hisham and Clifford (2008) showed that for the post-neonatal mortality, the significant factors are maternal education, province, 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 breast feeding 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 breast feeding 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, this study showed that biological and demographic variables are more important determinants of infant and post neonatal mortality.

Generally group differences in mortality have been explained by social characteristics hypothesis. This hypothesis emphasizes differences in socio-economic and demographic characteristics as the main cause of observed differences in mortality. The basic assumption is that irrespective of ethnic background, people who possess the same socio-economic and demographic characteristics should have similar mortality experiences. A study in India indicates that ethnicity per se does not have an effect on infant survival. Ethnic differences in infant mortality merely reflect socio-economic and household environmental differences between the members of the ethnic groups (Jatran, 2002).

Childhood mortality in general and infant mortality in particular are 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 order to achieve this goal, efforts are concentrated at identifying cost effective strategies as many international agencies have advocated for more resources to be directed to the health sector. One way of doing this is to identify and rank-order the importance of the socioeconomic and other factors that affect infant mortality. This will help in prioritizing the

factors that need to be manipulated for effective health interventions in the face of competing scarce resources.

Universally, there is huge literature that focused on the determinants of infant and child mortality. Most of the studies have shown significant association between socioeconomic and other factors and infant-child mortality. Socioeconomic (for example, Caldwell, 1979, Hisham and Clifford, 2008), demographic (Jatran, 2002), biological (for example, Forste, 1994) or environmental factors (for example, Mutunga, 2004) through making use of survey or censuses data.

In Ethiopian context studies about infant mortality have been conducted since the 1970s. Data from the 2005 Ethiopian DHS show neonatal mortality rate of 39 per 1,000 live births. This rate is similar to post-neonatal deaths (38 per 1,000 live births) during the same period; that is, the risk of dying for any Ethiopian child who survived the first month of life is the same as in the remaining 11 months of the first year of life. Thus 50 percent of infant deaths in Ethiopia occur during the first month of life, a similar pattern was observed in the 2000 EDHS. The infant mortality rate in the five years preceding the survey is 77 deaths per 1,000 live births and under-five mortality is 123 deaths per 1,000 live births for the same period. This means that one in every thirteen Ethiopian children dies before reaching age one, while one in every eight does not survive to the fifth birthday.

A hospital-based follow up study estimated a neonatal mortality of 71.8 deaths per1000 live birth in Addis Ababa, which was more than 60% of the infant mortality rate (Sahle Mariam and Berhane, 1997). Previous studies in Butajira (1991-92) found that parental factors, including paternal education, urban residence, and child feeding practices, lack of sanitation and pipe water were related to under-five mortality where the effect was stronger in infants compared to children 1-4 years (Shamebo et al, 1994).

Infant and child survival is influenced by the socioeconomic characteristics of mothers. Based on EDHS (2005) in Ethiopia mortality in urban areas is consistently lower than in rural areas. For example, infant mortality in urban areas is 66 deaths per 1,000 live births compared with 81 deaths per 1,000 live births in rural areas. The urban-rural difference is even more pronounced in the case of child mortality. Wide regional differentials in infant and under-five mortality are observed. For example, under-five mortality ranges from a low of 72 per 1,000 live births in Addis Ababa to a high of 157 per 1,000 live births in Benishangul-Gumuz. Under-five mortality is also relatively higher in Amhara and Gambela. As expected, mother's education is inversely related to a child's risk of dying. Under-five mortality among children born to mothers with no education (139 per 1,000 live births) is more than twice that of children born to mothers with secondary and higher level of education (54 per 1,000 live births). The beneficial effect of educating mothers is obvious for all childhood mortality rates.

Regarding the association between socioeconomic status and infant and child mortality, Caldwell (1979) reported on the effect of mother's education on reducing child mortality. He put up a theory that mother's education works through changing feeding and care practices, leading to better health seeking behavior and by changing the traditional familial relationships. In supporting Caldwell's explanation, 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. However, some studies showed a contrary conclusion that the effect of maternal education on child survival is weaker in sub-Saharan Africa.

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). However, in Kenya Hill et al (2001) has 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. Adjusting for a number of socioeconomic factors 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.

Other studies on determinants of infant and child mortality had supported the significant effect of some biological and demographic predictors on the phenomenon (Forste, 1994). Regarding the mortality differentials by gender, males had 50% higher IMR compared to females (Hisham and Clifford, 2008). According to Department of Health, South Australia, infant mortality rates are higher in males than females for almost all leading causes in Australia in the 15 year period 1982 to 1996, infant mortality of males was about 27% higher than females (DHSA, 2006). 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 (EDHS, 2005).

Sahile Mariam and Berhane (1997) studied neonatal mortality rate of 1334 singleton neonates born at three Addis Ababa hospitals prospectively followed during the neonatal period. They found that the neonatal mortality rate to be 71.8 deaths per 1000 live births, with early and late neonatal mortality rates of 50.9 and 20.9 per 1000 live births respectively. The main risk factors for perinatal mortality were being low birth weight and prematurity.

A child's birth weight or size at birth is an important indicator of the child's vulnerability to the risk of childhood illnesses and the chances of survival. Children whose birth weight is less than 2.5 kg are considered to have a higher than average risk of early childhood death. UNICEF (1999) estimates that at the global level about 17% of infants are born with birth weight less than 2,500 grams. The prevalence of low birth weight is not uniform throughout the world: in low income countries from about 10-30% as compared to developed countries 4-10%. The prevalence of low-birth weight among infants in Ethiopia was 20 percent from 2000 to 2007 (UNICEF, 2009). According to the 2005 Ethiopian DHS the percentage of low birth weight babies has increased in the past five years from 8 percent in 2000 to 14 percent in 2005. Birth weight is lower among children born to older women (age at birth greater than 35 years) and children of women with no education. The birth weight of a child also varies by mother's place of residence. Twenty-three percent of births in rural areas compared with 10 percent in urban areas have a reported birth weight less than 2.5 kg in Ethiopia (EDHS, 2005). Birth weight is determined also by duration of gestational age and intrauterine growth rate. Low birth weight is considered as one of the determinants of perinatal and neonatal mortality. Neonatal illness in general closely related to low birth weight, low birth weight babies also tend to have higher mortalities (Kulmala, 2000).

Maternal factors are known to influence the weight of the new born baby. Among the factors are: parity, birth interval, nutritional status as reflected by pre pregnancy weight, weight for height and weight gain, health status of the mother indicated as the presence of anemia, antenatal infections or complication of pregnancy and behavioral conditions like antenatal attendance and physical conditions during pregnancy (Kulmala, 2000). Maternal nutritional status is probably the most important determinant of the birth weight and the probability of neonatal survival. Multi gravidity (multiple pregnancy) and parity also affects the nutritional status of the mother. If a woman cannot recover fully from the effect of her last pregnancy and period of breast feeding before becoming pregnant again, her nutritional status might be expected to deteriorate with each successive pregnancy, which is called "Maternal Depression Syndrome". This condition increases the risk of premature birth and low birth weight babies with lower chance of survival (Kulmala, 2000, Karsten, 2002).

CHAPTER THREE

DATA AND METHODOLOGY

3.1 Source of Data

The data for this study are taken from "The Jimma infant survival differential longitudinal study" conducted in 46 urban and 64 rural 'kebeles', with an estimated population of 300,200, at the time of the study, in the administrative zones of Jimma, Illubabbor, and Keffecho, South-West Ethiopia during September 1992 and October 1994. The altitude of the study areas ranges from 1500m to just below 2000m. The sample involved a one-year birth cohort of 8,273 infants and their families. There were a total of 8, 161 births of which 8,050 singletons and 111 multiple births (110 sets of twins and one set of triplet). However, due to missing values in this study we considered 7,851 singleton and 89 twin infants. Data were collected by direct interview, observation and anthropometric measurement and dealt with socioeconomic, environmental, biomedical, behavioral and nutritional variables at birth and on a bimonthly follow-up visit.

Trained enumerators and Traditional Birth Attendants (TBAs) identify expectant mothers in their second trimester by house to house visits in their respective catchment kebeles. In order to identify all live births, all the traditional birth attendants (TBAs) in the kebeles were involved in the field work. TBAs are women residents of the kebele they serve and by tradition they visit and assist women during pregnancy and delivery. The TBAs had easy access to women in the fertile age group, and were able to assess their pregnancy status.

Each TBA was given responsibility for about 300 houses, and went house to house regularly to locate pregnant women in their second trimester. The TBA reported daily in person to the enumerator responsible for her kebele. For each three kebeles one high school completed girl enumerator was assigned. The enumerator registered the address of the expectant mother. After registration both the TBA and the enumerator monitored the expectant mother so as to reach her on time soon after delivery. For each three or more enumerators one supervisor (mainly

nurses) was assigned who checked and support data collection. In each kebele all one year live-birth cohorts were recruited for the study and followed for one year from 1992 to 1994.

3.2 Study Variables

3.2.1 The response variable

The response variable used in this study is "infant survival time", which is measured as the duration in days starting from the infant birth to death (if the event occurred) or from the infant birth to the end of the study period or to the date of lost for follow-up (censored observations). Due to the nature of the dependent variable in the regression model for survival data analysis, statistical packages for estimating this model require that the data on the dependent variable be composed of two parts: a dummy variable indicating whether or not the event (infant death) occurred during the study period; and a variable giving either the time of the event occurring or the time of censoring. In this study, for those infants who died during the study period, the second component of the dependent variable is the number of days from birth to death. For those infants who did not die by the end of the study or those who were lost to follow-up, the second component of the dependent variable is the number of days of follow-up. Since there are censored observations the coding of the first component of the dependent variable, event, is such that 1 indicates uncensored observation (those who died before celebrating their first birth day) and 0 for censored observations (those who are alive at the end of the study period or lost to follow-up).

3.2.2 Explanatory variables

The explanatory variables considered in this study include: socioeconomic, demographic and biological variables. The socioeconomic variables include maternal education level, type of place of residence (urban / rural) and source of drinking water. The demographic variables used are age of the mother at birth, ethnicity, religion, family size and marital status. The biological variables include total live birth, sex of the child and birth weight of the infant. These are specified and defined below.

Table1. The explanatory variables considered in this study

Variable	Variable description	Category
Meduc	Maternal educational level	0 = " illiterate" 1 = " 1 – 6 grade" 2 = " secondary & above"
Site	Place of residence	0 = " urban" 1 = " rural "
Momage	Age of the mother at birth	0= " <20 years" 1= " 20-34 years" 2= ">=35 years"
Sex	Sex of the child	0 = " female" 1 = " male"
Mometh	Ethnicity of the mother	0 = " Oromo" 1= " Other"
Momrel	Mother's religion	0 = " Christian" 1 = " Muslim"
Mstatus	Marital status of the mother	0 = " Currently Married" 1 = "Currently not Married"
Famsize	Family size	
Bwt	Birth weight of the infant	0= "Weight>=2,500gm" 1= "Weight<2,500gm"
Parity	Total number of live birth	1 = " Primi" 2 = " 2 – 4" 3 = " > 4"
Water	Source of drinking water	0 = " pipe" 1 = " protected well/ spring" 2 = " unprotected source"

3.3 Methodology

3.3.1 Survival analysis

Survival analysis is concerned with the technique for studying data related to time-to-an-event. Survival data are not amenable to standard statistical procedures used in data 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.

The main goal for a statistical analysis of survival data is to fit a model that will yield biologically plausible and interpretable estimates of the effect of covariates on survival time. The fact that some of our observations of the outcome variable, survival time are incomplete is a problem for conventional univariate summary statistics such as the mean, standard deviation, etc. In this situation, we must obtain an estimate of the cumulative distribution function in order to obtain values of statistics which summarize the survival data. Summary values are at the heart of statistical analysis. Survivor and hazard functions are the two functions of central interest in summarizing survival data (Hosmer and Lemeshow, 1998).

Let T be a random variable representing survival times and t be the realization of the random variable T. The cumulative distribution function is the probability that a subject selected at random will have a survival time less than some stated value, t. which is given as:

$$F(t) = P(T < t)$$

And its complement, the survivorship function is given as:

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

The survivor function can therefore be used to represent the probability of observing a survival time greater than or equal to some stated value t.

As already mentioned, the hazard function is another important analytic tool for the analysis of survival time data. The term "hazard" is used to describe the concept of the risk of "failure" in an interval after time t, conditional on the subject having survived to time t. That is, the function represents the instantaneous death rate for an individual surviving to time t. Thus, the hazard function, $h(t)$, is defined as:

$$h(t) = \lim_{\Delta t \to 0} \left\{ P\left(\frac{t \leq T \leq t + \Delta t | T > t}{\Delta t} \right) \right\}$$

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

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

where $f(t)$ is the probability density function of T.

The survival and hazard functions can be estimated using the Kaplan-Meier method, or Life Table method, or Nelson-Aalen method as a preliminary analysis. These methods are non-parametric or distribution-free, since these do not require specific assumption to be made about the underlying distribution of the survival times. When we have a large study whose mortality experience is presented in calendar time units (such as quarterly, semi-annually, etc) the Life Table estimator of the survivorship function may be used.

Unlike the Life Table estimator the Kaplan-Meier estimator of the survival function is based on individual observations and it is more precise than the Life Table estimator. In order to construct the Kaplan-Meier estimator, suppose we have n independent observations. Denote their survival times by $t_1, t_2, \cdots, t_n$ and indicator of censoring by $\delta_1, \delta_2 \cdots, \delta_n$.

where $\delta_i = \begin{cases} 1 \text{ , event or death} \\ 0 \text{ , censored} \end{cases}$

The survival data are denoted by (t_i, δ_i) for i=1, 2, $\cdots$,n

Suppose we have $m \le n$ failures occurred at distinct m times. The distinct ordered failure times observed among the n individuals are $t_{(1)}, t_{(2)}, \cdots t_{(m)}$, where $m \le n$.

The conditional probability of survival at time $t_{(i)}$, $P(t_{(i)})$, is then estimated by:

$$P(t_{(i)}) = \frac{n_i - d_i}{n_i} \quad ,$$

where n_i is the number of individuals at risk of dying or failing at time $t_{(i)}$ and d_i is the number of deaths at time $t_{(i)}$.

Then the Kaplan-Meier estimator of the survival function at time t, is the product of conditional survival probabilities and denoted as:

$$\hat{S}_{KM}(t) = \prod_{t_{(i)} \le t} \left(\frac{n_i - d_i}{n_i} \right).$$

For convenience we take $\hat{S}_{KM}(t) = 1$ if $t < t_{(1)}$

Based on $\hat{S}_{KM}(t)$ the cumulative hazard function is estimated by:

$$\hat{H}_{KM}(t) = -\ln\left(\hat{S}_{KM}(t)\right).$$

The variance of the Kaplan-Meier survival function estimator is estimated by the following formula which is estimated based on the delta method using first order Taylor series expansion.

$$\hat{Var}\left(\hat{S}_{KM}(t)\right) = \left[\hat{S}_{KM}(t)\right]^2 \sum_{t_{(i)} \le t} \frac{d_i}{n_i(n_i - d_i)} .$$

This is referred to as Greenwood's formula for estimating the variance of Kaplan-Meier survival function estimator.

The construction of point wise confidence interval for the Kaplan-Meier survival function estimator is given in Appendix B.

The variance of the cumulative hazard function estimator is estimated by:

$$\hat{Var}\left(\hat{H}_{KM}(t)\right) = \sum_{t_{(i)} \le t} \frac{d_i}{n_i\left(n_i - d_i\right)} .$$

Another estimator of the cumulative hazard function of interest is, the Nelson-Aalen estimator which is given as:

$$\hat{H}_{NA}(t) = \sum_{t_{(i)} \le t} \frac{d_i}{n_i} . \qquad \ldots\ldots\ldots (3.1)$$

The corresponding survival function estimator is given by

$$\hat{S}_{NA}(t) = e^{-\hat{H}_{NA}(t)} .$$

After obtaining statistics which provide a description of the overall survival experience, i.e.; the survival and hazard functions, we usually turn our attention to a comparison of the survivorship experience in key subgroups in the data. These groups might be defined by the values of a covariate which are thought to be related to survival times.

3.3.2 Tests for two or more groups

When comparing groups of subjects, it is always a good idea to begin with a graphical display of the data in each group. In studies of survival time, we plot the Kaplan-Meier estimator of the survivorship function for each of the groups. It can be considered as a visual aid for comparison. In general, the pattern of one survivorship function lying above another means the group defined by the upper curve lived longer, or had a more favorable survival experience,

than the group defined by the lower curve. In other words, at any point in time the proportion of subjects estimated to be alive is greater for one group (represented by the upper curve) than the other (represented by the lower curve).

The goals of comparing survivorship functions in different groups of subjects are identical to those of the two sample t-test, the non-parametric rank sum test and the analysis of variance. Namely, we wish to quantify differences between groups through point and interval estimates of key measures. Standard statistical procedures, such as those named above, may be used without modification when there are no censored observations. Since survival data are typically right skewed, we would likely use rank-based non-parametric tests followed by estimates and confidence intervals of medians (and possibly other quantiles) within groups. Modifications of these procedures are required when censored observations are present in the data (The detail discussion is found in Hosmer and Lemeshow, 1998).

Here we discuss on hypothesis tests that are based on comparing the Nelson-Aalen cumulative hazard rate estimator (the Kaplan-Meier survival function estimator could equally be used), obtained directly from the data, to an expected estimator of the cumulative hazard rate, based on the assumed model under the null hypothesis. Rather than a direct comparison of these two rates, we shall examine tests that look at weighted differences between the observed and expected hazard rates (The detail discussion is found in Klein and Moeschberger, 1997).

A method used in comparing hazard rates of $K(K \geq 2)$ populations, that is, a procedure to test the following set of hypotheses:

H_0 : $h_1(t)=h_2(t)=\ldots=h_K(t)$, for all $t \leq \tau$, versus

H_A : at least one of the $h_j(t)$'s is different for some $t \leq \tau$

Where $h_j(t)$, $j = 1, 2, \ldots$, K is the hazard rate of the j^{th} population at a time t.

Our inference is about the hazard rates for all time points less than τ, which is, typically, the longest time in the study. The alternative hypothesis is a global one in that we wish to reject the null hypothesis if, at least, one of the populations differs from the others at some time.

Suppose $t_1, t_2, \ldots, t_D$ are the distinct death times in the pooled sample. At time t_i we observe d_{ij} events in the j^{th} sample out of Y_{ij} individuals at risk, $j = 1, 2, \ldots, k$ and $i = 1, 2, \ldots, D$.

Let $d_i = \sum_{j=1}^{K} d_{ij}$ and $Y_i = \sum_{j=1}^{K} Y_{ij}$ be the number of deaths and the number at risk in the combined sample at time t_i, $i = 1, 2, \ldots, D$. The test of H_0 is based on weighted comparison of the estimated hazard rate of the j^{th} population under the null and alternative hypotheses, based on the Nelson-Aalen estimator in (3.1). If the null hypothesis is true, then, an estimator of the expected hazard rate in the j^{th} population under H_0 is the pooled sample estimator of the hazard rate d_i / Y_i whereas using only data from the j^{th} sample, the estimator of the hazard rate is d_{ij} / Y_{ij}. To make comparison let $W_j(t)$ be a positive weight function with the property that $W_j(t_i)$ is zero whenever Y_{ij} is zero. The test of H_0 is based on the statistics:

$$Z_j(\tau) = \sum_{i=1}^{D} W_j(t_i) \left\{ \frac{d_{ij}}{Y_{ij}} - \frac{d_i}{Y_i} \right\}, \quad j = 1, 2, \ldots, K \qquad \ldots\ldots\ldots (3.2)$$

If all the $Z_j(\tau)$'s are close to zero, then, there is little evidence to believe that the null hypothesis is false, whereas, if one of the $Z_j(\tau)$'s is far from zero, then, there is evidence that this population has a hazard rate differing from that expected under the null hypothesis.

Although the general theory allows for different weight functions for each of the comparisons in (3.2) in practice, all of the commonly used tests have a weight function $W_j(t_i) = Y_{ij} W(t_i)$. Here, $W(t_i)$ is a common weight shared by each group, and Y_{ij} is the number at risk in the j^{th} group at time t_i. With this choice of weight functions

$$Z_j(\tau)=\sum_{i=1}^{D} W(t_i)\left[d_{ij}-Y_{ij}\left(\frac{d_i}{Y_i}\right)\right], j=1,2,\ \ldots\ ,\mathrm{K}. \qquad \ldots\ldots\ldots (3.3)$$

Note that with this class of weights the test statistic is the sum of the weighted difference between the observed number of deaths and the expected number of deaths under H_0 in the j^{th} sample. The expected number of deaths in sample j at t_i is the proportion of individuals at risk Y_{ij}/Y_i that are in sample j at time t_i, multiplied by the number of deaths at time t_i.

The variance of $Z_j(\tau)$ in (3.3) is given by:

$$\hat{\sigma}_{jj}=\sum_{i=1}^{D} W(t_i)^2\frac{Y_{ij}}{Y_i}\left(1-\frac{Y_{ij}}{Y_i}\right)\left(\frac{Y_i-d_i}{Y_i-1}\right)d_i\ , j=1,2,\ \ldots\ ,K \qquad \ldots\ldots\ldots (3.4)$$

and the covariance of $Z_j(\tau)$, $Z_g(\tau)$ is expressed by:

$$\hat{\sigma}_{jg}=-\sum_{i=1}^{D} W(t_i)^2\frac{Y_{ij}}{Y_i}\frac{Y_{ig}}{Y_i}\left(\frac{Y_i-d_i}{Y_i-1}\right)d_i\ , j\neq g$$

The term $(Y_i-d_i)/(Y_i-1)$, which equals one if no two individuals have a common event time, is a correction for ties. The terms $\frac{Y_{ij}}{Y_i}\left(1-\frac{Y_{ij}}{Y_i}\right)d_i$ and $-\frac{Y_{ij}}{Y_i}.\frac{Y_{ig}}{Y_i}d_i$ arise from the variance and covariance of a multinomial random variable with parameters d_i, $p_j=\frac{Y_{ij}}{Y_i}$, $j=1,2,\ \ldots\ ,K$. The components vector $\left(Z_1(\tau),Z_2(\tau),\ \ldots\ ,Z_K(\tau)\right)$ are linearly dependent because $\sum_{j=1}^{K} Z_j(\tau)$ is zero. The test statistic is constructed by selecting any $K-1$ of the Z_j's. The estimated variance-covariance matrix of these statistics is given by the $(K-1)\times(K-1)$ matrix Σ, formed by the appropriate $\hat{\sigma}_{jg}$'s. The test statistics is given by the quadratic form

$$\chi^2 = \left(Z_1(\tau), Z_2(\tau), \ \ldots \ , Z_{K-1}(\tau)\right)\Sigma^{-1}\left(Z_1(\tau), Z_2(\tau), \ \ldots \ , Z_{K-1}(\tau)\right)'$$

When the null hypothesis is true, this statistic has a chi-squared distribution for large samples with K-1 degrees of freedom. An α level test of H_0 rejects when χ^2 is larger than the α^{th} upper percentage point of a chi-squared, random variable with $K-1$ degrees of freedom.

When $K = 2$ the test statistic can be written as:

$$Z = \frac{\sum_{i=1}^{D} W(t_i)\left[d_{i1} - Y_{i1}\left(\frac{d_i}{Y_i}\right)\right]}{\sqrt{\sum_{i=1}^{D} W(t_i)^2 \frac{Y_{i1}}{Y_i}\left(1-\frac{Y_{i1}}{Y_i}\right)\left(\frac{Y_i - d_i}{Y_i - 1}\right)d_i}}$$

which has a standard normal distribution for large samples when H_0 is true, using this statistic, the α level test of the alternative hypothesis:

$H_A : h_1(t) > h_2(t)$, for $t \le \tau$, is rejected when the calculated value of $Z \ge z_\alpha$, the α^{th} upper percentage point of a standard normal distribution.

A variety of weight functions have been proposed in the literature. A common weight function, leading to a test available in most statistical packages, is $W(t) = 1$ for all t . This choice of weight function leads to the so-called log rank test.

Now we shall see how these tests can be used to analyze the survival of matched pairs. Here we have paired event times (T_{1i}, T_{2i}) and their corresponding event indicators $(\delta_{1i}, \delta_{2i})$, for $i = 1, 2, \ \ldots \ , n$. We wish to test the hypothesis $H_0 : h_1(t) = h_2(t)$. Computing the statistics (3.3) and (3.4) we get (3.5) and (3.6).

$$Z_{1i}(\tau)=\begin{cases} W(T_{1i})(1-1/2)=W(T_{1i})/2 & \text{if } T_{1i}<T_{2i},\delta_{1i}=1 \\ & \text{or } T_{1i}=T_{2i},\delta_{1i}=1,\delta_{2i}=0 \\ W(T_{2i})(0-1/2)=-W(T_{2i})/2 & \text{if } T_{2i}<T_{1i},\delta_{2i}=1 \\ & \text{or } T_{2i}=T_{1i},\delta_{2i}=1,\delta_{1i}=0 \\ 0 & \text{Otherwise} \end{cases} \quad \ldots\ldots\ldots (3.5)$$

And

$$\hat{\sigma}_{11i}=\begin{cases} W(T_{1i})^2(1/2)(1-1/2)=W(T_{1i})^2/4 & \text{if } T_{1i}<T_{2i},\delta_{1i}=1 \\ & \text{or } T_{1i}=T_{2i},\delta_{1i}=1,\delta_{2i}=0 \\ W(T_{2i})^2(1/2)(1-1/2)=W(T_{2i})^2/4 & \text{if } T_{2i}<T_{1i},\delta_{2i}=1 \\ & \text{or } T_{2i}=T_{1i},\delta_{2i}=1,\delta_{1i}=0 \\ 0 & \text{Otherwise} \end{cases} \quad \ldots\ldots (3.6)$$

For any of the weight functions W we have:

$$Z_{1.}(\tau)=w\left(\frac{D_1-D_2}{2}\right)$$

and

$$\hat{\sigma}_{11.}=w^2\left(\frac{D_1+D_2}{4}\right)$$

where D_1 is the number of matched pairs in which the individual from sample 1 experiences the event first and D_2 is the number in which the individual from sample 2 experiences the event first. Here w is the value of the weight function W at the time when the smaller of the pair fails. Because these weights do not depend on from which group this failure came from, the test statistic is:

$$Z=\frac{D_1-D_2}{\sqrt{D_1+D_2}} \quad \ldots\ldots\ldots (3.7)$$

This has a standard normal distribution when the number of pairs is large under the null hypothesis. Note that matched pairs, where the smaller of the two times corresponds to a censored observation, make no contribution to Z_L or $\hat{\sigma}_{LL}$. Detail discussion is found in Klein and Moeschberger (1997).

3.3.3 Regression Model

Regression modeling of the relationship between an outcome variable and independent predictor variable(s) is commonly employed in virtually all fields. The popularity of this approach is due to the fact that plausible models may be easily fit, evaluated and interpreted. Statistically, the specification of a model requires choosing both the systematic and error components. The choice of the systematic component involves an assessment of the relationship between an "average" of the outcome variable and the independent variable(s). This may be guided by an exploratory analysis of the current data and/or past experience. The choice of an error component involves specifying the statistical distribution of what remains to be explained after the model is fit i.e.; the residuals.

In an applied setting, the task of model selection is, to a large extent, based on the goals of the analysis and on the measurement scale of the outcome variable. For example, we may wish to model the relationship between a continuous dependent variable and a set of independent variables. A good place to start would be to use a model with a linear systematic component and normally distributed errors, the usual linear regression model. Suppose instead we wish to model the relationship between a dichotomous dependent variable and a set of explanatory variables. If we assume the goal of this analysis is to estimate the "effect" of the various factors via an odds-ratio, then the logistic regression model would be a good choice. The logistic regression model has a systematic component that is linear in the log-odds and has binomial/Bernoulli distributed errors.

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 main goal for a statistical analysis of survival data is to fit a model that will yield meaningful and interpretable estimates of the effect of covariates on survival time. Mainly there are three types of regression models for survival data analysis. These models differ in the assumptions made about the relationship between covariates and survival experience. For example, in the accelerated failure time model log-of survival time is related additively to the linear combination of covariates; in proportional hazards model log-of the hazard rate is related additively to the linear combination of covariates or hazard rate is related multiplicatively to the linear combination of covariates; and in proportional odds model log-of odds of survival is related additively to the linear combination of covariates. Detail discussion is found in Hosmer and Lemeshow (1998).

3.3.4 Accelerated Failure-Time Model

Let T be a random variable representing survival time, a convenient and meaningful way to describe survival time with a set of covariates, $(X_1, X_2, \cdots, X_p)$, is with the equation

$$T = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p} \times \varepsilon \qquad \ldots\ldots\ldots (3.8)$$

the error component, ε, follows the exponential distribution with parameter equal to one and its density function is given by:

$$f(\varepsilon) = e^{-\varepsilon}$$

This model, (3.8), can be "linearized" by taking the natural logarithm of each side of the equation yielding

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \theta \qquad \ldots\ldots\ldots (3.9)$$

where $Y=\ln(T)$ and $\theta = \ln(\varepsilon)$. The error component θ follows an "extreme minimum value" distribution. This distribution is not encountered often outside of applications in survival analysis but plays central role in models of life-length and is often referred to as the Gumbel

distribution. The mean of this distribution is 0 and its shape parameter is 1 (denoted by $G(0,1)$). The density function of the $G(0,1)$ is given by:

$$f(\theta) = \exp\left[\theta - e^{(\theta)}\right]$$

The use of the distribution $G(0,1)$ in (3.9) is somewhat like using the standard normal distribution in linear regression. The standard normal distribution is denoted $N(0,1)$. From practical experience we know that, in linear regression, the errors rarely if ever have variance equal to one. The usual assumption is that the variance is neither a function of the outcome variable nor of the independent variables. It is assumed to be constant and equal to the parameter σ^2. This distribution is denoted by $N(0,\sigma^2)$. An additional parameter may be introduced into (3.9) by multiplying θ by σ to yield the model

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \sigma \times \theta \qquad \ldots\ldots\ldots (3.10)$$

The distribution of $\sigma \times \theta$ is denoted as $G(0,\sigma)$.

Under this model (3.10) survival time follows the exponential distribution when $\sigma = 1$ and the Weibull distribution when $\sigma \neq 1$. On the other hand when θ is distributed as standard normal, $N(0,1)$ or standard logistic, $L(0,1)$, the survival time follows log-normal distribution or log-logistic distribution respectively (Hosmer and Lemeshow, 1998).

Survival time models that can be linearized by taking log-transformation are called accelerated failure time models. The reason for this terminology is that the effect of the covariate is multiplicative on the time scale. That is, the effect of the covariate is said to "accelerate" survival time. In terms of the survival function, the accelerated failure-time model states that the survival function of an individual with covariate X at time t is the same as the survival function of an individual with a baseline survival function at a time $t\exp(\underset{\sim}{\beta}'X)$, where $\underset{\sim}{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p)$ is a vector of regression coefficients and $X = (x_0, x_1, \ldots, x_p)$. In other words, the accelerated failure time model is defined by the relationship:

$$S(t|X) = S_0\left[t\exp\left(\underset{\sim}{\beta}'X\right)\right], \text{ for all t}$$

where S_0 is a baseline survival function.

The factor $\exp\left(\underset{\sim}{\beta}'X\right)$ is called an acceleration factor which tells us how a change in covariate values changes the time scale from the baseline time scale. One implication of this model is that the hazard rate for an individual with covariate X is related to a baseline hazard rate by

$$h(t|X) = \exp\left(\underset{\sim}{\beta}'X\right)h_0\left[\exp\left(\underset{\sim}{\beta}'X\right)t\right], \text{ for all t}$$

where h_0 is the baseline hazard rate.

3.3.5 The Weibull Regression Model

The Weibull distribution is a very flexible model for lifetime data. It has a hazard rate which is monotone increasing, decreasing, or constant. It is one of the parametric regression models that have both a proportional hazards representation and an accelerated failure-time representation. The hazard function for the single covariate Weibull regression model is:

$$h\left(t,x,\underset{\sim}{\beta},\lambda\right) = \frac{\lambda t^{\lambda-1}}{\left(e^{\beta_0+\beta_1 x}\right)^{\lambda}} \qquad \ldots\ldots\ldots (3.11)$$

For convenience, we use $\lambda = 1/\sigma$. This hazard function may be re-expressed in proportional hazards or accelerated failure time form. The proportional hazards form of the function is obtained as follows:

$$h\left(t,x,\underset{\sim}{\beta},\lambda\right) = \lambda t^{\lambda-1}e^{-\lambda(\beta_0+\beta_1 x)}$$

$$= \lambda t^{\lambda-1}e^{-\lambda\beta_0}e^{-\lambda\beta_1 x}$$

$$h(t,x,\underset{\sim}{\beta},\lambda) = \lambda\gamma t^{\lambda-1}e^{-\lambda\beta_1 x}$$

$$= h_0(t)e^{\theta_1 x} \quad , \qquad \ldots\ldots\ldots (3.12)$$

where $\gamma = \exp(-\beta_0/\sigma) = \exp(\theta_0)$, $\theta_1 = -\beta_1/\sigma$ and the baseline hazard function is $h_0(t) = \lambda\gamma t^{\lambda-1}$ although the parameter σ is a variance-like parameter on the log-time scale, $\lambda = 1/\sigma$ is commonly called the shape parameter. So, we will refer to σ as the shape parameter. The parameter γ is called a scale parameter. The expression for the hazard function in (3.12) leads to a hazard ratio interpretation of the parameter θ_1 i.e.; the proportional hazards model representation.

The accelerated failure time form of the hazard function is obtained by re-expressing (3.11) as follows:

$$h(t,x,\underset{\sim}{\beta},\lambda) = \lambda t^{\lambda-1}e^{-\lambda(\beta_0+\beta_1 x)}$$

$$= \lambda\gamma\left(te^{-\beta_1 x}\right)^{\lambda-1}e^{-\beta_1 x} \quad . \qquad \ldots\ldots\ldots (3.13)$$

The rationale for presenting both forms of the hazard function is that different software packages use different parameterizations when fitting a model. For example, SAS use the accelerated failure time model parameterization in (3.13) and report estimates of the $\underset{\sim}{\beta}$ form of the coefficients and σ where $\sigma = 1/\lambda$. STATA, on the other hand, offers the option of providing estimates of either parameterization. If the accelerated failure time or log-time form is chosen, estimates of $\underset{\sim}{\beta}$ and σ are provided. If the log-hazard form is chosen (3.12) representation, estimates of $\underset{\sim}{\theta}$ and σ are provided such that the relationship between the two sets of coefficients is $\underset{\sim}{\theta} = -\underset{\sim}{\beta}/\sigma$. In the remainder of this section, we use the accelerated failure time formulation of the model.

The survivorship function corresponding to the accelerated failure time form of the hazard function in (3.13) is

$$S\left(t,x,\underset{\sim}{\beta},\sigma\right)=\exp\left\{-t^{\lambda}\exp\left[\left(-1/\sigma\right)\left(\beta_{o}+\beta_{1}x\right)\right]\right\}.$$

We obtain the equation for the median survival time by setting the survivorship function equal to 0.5 and solving for time yielding

$$t_{50}\left(x,\underset{\sim}{\beta},\sigma\right)=\left[-\ln\left(0.5\right)\right]^{\sigma}e^{\beta_{o}+\beta_{1}x}. \qquad \ldots\ldots\ldots (3.14)$$

For example, if the covariate is dichotomous and coded 0/1, the time ratio at the median survival time is:

$$TR\left(x=1,x=0\right)=\frac{t_{50}\left(x=1,\underset{\sim}{\beta},\sigma\right)}{t_{50}\left(x=0,\underset{\sim}{\beta},\sigma\right)}=\frac{\left[-\ln\left(0.5\right)\right]^{\sigma}e^{\beta_{0}+\beta_{1}}}{\left[-\ln\left(0.5\right)\right]^{\sigma}e^{\beta_{0}}}=e^{\beta_{1}}. \qquad \ldots\ldots\ldots (3.15)$$

Alternatively, the relationship between the two median times is:

$$t_{50}\left(x=1,\underset{\sim}{\beta},\sigma\right)=e^{\beta_{1}}t_{50}\left(x=0,\underset{\sim}{\beta},\sigma\right). \qquad \ldots\ldots\ldots (3.16)$$

From (3.16), we can see that the median survival time in the group with $x=1$ is e^{β_1} times the median survival time in the group with $x=0$. From this the multiplicative covariate effect on the time scale is clear and easy to understand. The result in (3.15) holds not only for the median but for all percentiles.

The $(1-\alpha)100$ percent confidence interval for the time ratio in (3.15) is constructed based on the $(1-\alpha)100$ percent confidence interval of $\hat{\beta}_1$, estimator of β.

The Weibull regression model with P covariates can be expressed as

$$Y=X'\underset{\sim}{\beta}+\sigma W$$

where $X=\left(x_0,x_1,\ \ldots\ ,x_p\right)$, $x_0=1$ and $Y=\ln\left(T\right)$ and the error term, σW, follows an "extreme minimum value" distribution, i.e.; $W\sim G\left(0,\sigma\right)$. $\underset{\sim}{\beta}'=\left(\beta_0,\beta_1,\ldots,\beta_p\right)$ represents the vector of regression coefficients. The probability density function of W is:

$$f_W(w) = \exp(w - e^w) .$$

Thus, the underlying probability density function and survival function, respectively, for Y, are

$$f_Y(y) = \frac{1}{\sigma}\exp\left\{\frac{\left(y - X'\underset{\sim}{\beta}\right)}{\sigma} - e^{\frac{\left(y - X'\underset{\sim}{\beta}\right)}{\sigma}}\right\}$$

and

$$S(y) = \exp\left(-e^{\left[\frac{\left(y - X'\underset{\sim}{\beta}\right)}{\sigma}\right]}\right).$$

Parameter Estimation

The full maximum likelihood estimation method is used to estimate parameters of the model. The first step in maximum likelihood estimation is to create the specific likelihood function to be maximized. In simplest terms, the likelihood function is an expression that yields a quantity similar to the probability of the observed data under the model. In case of regression models for survival data, we construct the actual likelihood function by considering the contribution of the individual observations separately. Those subjects who died or experience the event contribute through their probability density function and censored observations contribute through their survival function to the likelihood function. Thus, under the assumption of independent observations, the full likelihood function is obtained by multiplying the respective contributions of the individual observations. Detail discussion is found in Hosmer and Lemeshow (1998).

For right censored observations the likelihood is:

$$L\left(\underset{\sim}{\beta},\sigma\right) = \prod_{i=1}^{n}\left[f\left(y_i;\underset{\sim}{x}_i,\underset{\sim}{\beta},\sigma\right)\right]^{\delta_i}\left[S\left(y_i;\underset{\sim}{x}_i,\underset{\sim}{\beta},\sigma\right)\right]^{1-\delta_i}$$

$$= \prod_{i=1}^{n} \left[\frac{1}{\sigma} e^{\left(\frac{y_i - x_i \beta}{\sigma} \right)} \right]^{\delta_i} \exp\left[-e^{\left(\frac{y_i - x_i \beta}{\sigma} \right)} \right] .$$

In order to avoid complication in the calculation of the likelihood let

$$z_i = \frac{y_i - x_i \beta}{\sigma}$$

and $f(z)$ is replaced with $\frac{1}{\sigma} f(z)$. This yields the log-likelihood function

$$l(\beta, \sigma) = \sum_{i=1}^{n} \delta_i (-\ln(\sigma) + z_i) - e^{z_i} \qquad \ldots\ldots\ldots (3.17)$$

The score equation for the j^{th} regression coefficient is obtained by taking the derivative of (3.17) with respect to β_j and setting it equal to zero, yielding

$$\frac{\partial l(\beta, \sigma)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{-x_{ij}}{\sigma} (\delta_i - e^{z_i}) = 0 \ , j = 0, 1, 2, \ldots, p . \qquad \ldots\ldots\ldots (3.18)$$

The score equation for the shape parameter, σ, is:

$$\frac{\partial l(\beta, \sigma)}{\partial \sigma} = \frac{-m}{\sigma} + \sum_{i=1}^{n} \frac{z_i}{\sigma} (\delta_i - e^{z_i}) = 0 \qquad \ldots\ldots\ldots (3.19)$$

where $m = \sum_{i=1}^{n} \delta_i$, is the number of non-censored observations.

Some software packages, for example, STATA, parameterize the log-likelihood function in terms of $-\ln(\sigma)$ since this makes the score equation easier to solve from a computational point of view. Regardless of what form is used, all packages report an estimate of σ. The solution to (3.18) and (3.19) are denoted as $\hat{\beta}$ and $\hat{\sigma}$, respectively.

As is the case in any application of maximum likelihood, the estimator of the covariance matrix of the parameter estimator is obtained from the observed information matrix. The individual elements of this matrix to be evaluated are:

$$-\frac{\partial^2 l\left(\underset{\sim}{\beta},\sigma\right)}{\partial\beta_j\partial\beta_k}=\frac{1}{\sigma^2}\sum_{i=1}^{n}x_{ij}x_{ik}e^{z_i}$$

$$-\frac{\partial^2 l\left(\underset{\sim}{\beta},\sigma\right)}{\partial\beta_j\partial\sigma}=\frac{1}{\sigma^2}\sum_{i=1}^{n}x_{ij}z_i e^{z_i}$$

and

$$-\frac{\partial^2 l\left(\underset{\sim}{\beta},\sigma\right)}{\partial\sigma\partial\sigma}=\frac{m}{\sigma^2}+\frac{1}{\sigma}\sum_{i=1}^{n}z_i^2 e^{z_i}$$

When evaluated at the solution to the likelihood equations, the information matrix may be expressed as

$$I\left(\hat{\underset{\sim}{\beta}},\hat{\sigma}\right)=\frac{1}{\hat{\sigma}^2}\begin{bmatrix}X'\hat{V}X & X'\hat{V}\hat{z}\\ \hat{z}'\hat{V}X & \hat{z}'\hat{V}\hat{z}+m\end{bmatrix}$$

where X is an n by $p+1$ matrix containing the values of the covariates, $\hat{V}=\mathrm{diag}\left(e^{\hat{z}_i}\right)$, an n by n diagonal matrix, and $\hat{z}'=\left(\hat{z}_1,\hat{z}_2,\ldots,\hat{z}_n\right)$, with

$$\hat{z}_i=\frac{y_i-\underset{\sim}{x}_i'\hat{\underset{\sim}{\beta}}}{\hat{\sigma}}.$$

The estimator of the covariance matrix for the estimators of the parameters is

$$\widehat{\mathrm{Var}}\left(\hat{\underset{\sim}{\beta}},\hat{\sigma}\right)=\left[I\left(\hat{\underset{\sim}{\beta}},\hat{\sigma}\right)\right]^{-1}.$$

The end-points of a $(1-\alpha)100$ percent Wald statistic based confidence interval for the j^{th} coefficient is given by:

$$\hat{\beta}_j \pm Z_{1-\alpha/2} \hat{SE}\left(\hat{\beta}_j\right)$$

where $\hat{SE}\left(\hat{\beta}_j\right)$ denotes the estimator of the standard error of the estimator of the coefficient.

3.3.6 Log-normal Regression Model

Given a set of covariates $X = \left(x_0, x_1, \ldots, x_p\right)'$, the logarithm of the time-to-event follows the usual normal regression model, that is,

$$Y = \ln(T) = \beta' X + \sigma W$$

where W follows a standard normal distribution. For the log-normal distribution the survival function of the time to event T is give by:

$$S(t) = 1 - \Phi\left\{\left[\log(t) - \beta' X\right]/\sigma\right\}$$

where $\Phi\{\ \}$ is the standard normal cumulative distribution function.

The hazard function does not have a closed form (Klein and Moeschberger, 1997). The likelihood construction, parameter estimation and so on follow the same step like the Weibull distribution discussed in section (3.3.5).

3.3.7 Log-logistic Regression Model

An alternative model to the Weibull regression model is the log-logistic regression model. This distribution has a hazard rate which is hump-shaped, that is, it increases initially and, then, decreases. It has a survival function and hazard rate that has a closed form expression, as contrasted with the log-normal distribution which also has a hump-shaped hazard rate.

The single covariate log-logistic accelerated failure time model may be expressed as

$$\ln(T) = \beta_0 + \beta_1 x + \sigma\varepsilon \qquad \ldots\ldots\ldots (3.20)$$

where the error term ε follows the standard logistic distribution. The appealing feature of the log-logistic model is that the slope coefficient can be expressed in such a way that it can be interpreted as an odds-ratio. In order to develop the odds-ratio interpretation, we begin by expressing the survivorship function for the model in (3.20) as

$$S\left(t, x, \underset{\sim}{\beta}, \sigma\right) = \left[1 + \exp(z)\right]^{-1} \qquad \ldots\ldots\ldots (3.21)$$

where z, as before, is the standardized log-time outcome variable, i.e., $z = (y - \beta_0 - \beta_1 x)/\sigma$ and $y = \ln(t)$. The odds of a survival time of at least t is

$$\frac{S\left(t, x, \underset{\sim}{\beta}, \sigma\right)}{1 - S\left(t, x, \underset{\sim}{\beta}, \sigma\right)} = \exp(-z) \qquad \ldots\ldots\ldots (3.22)$$

As an example, assume that the covariate is dichotomous and coded 0/1. The odds-ratio at time t formed from the ratio of the odds in (3.22) evaluated at binary covariate $x = 1$ and $x = 0$ is

$$\mathrm{OR}(t, x = 1, x = 0) = \frac{\exp\left[\dfrac{-(y - \beta_0 - \beta_1 \times 1)}{\sigma}\right]}{\exp\left[\dfrac{-(y - \beta_0 - \beta_1 \times 0)}{\sigma}\right]} = \exp(\beta_1/\sigma) \qquad \ldots\ldots\ldots (3.23)$$

Note that the ratio in (3.23) is independent of time. If the odds-ratio in (3.23) was 2, the interpretation would be that the odds of survival beyond time t among subjects with $x = 1$ is twice that of subjects with $x = 0$, and this holds for all t.

An alternative interpretation is obtained when we express the median survival time as a function of the regression coefficients. Setting the survivorship function in (3.21) equal to 0.5 and solving, we obtain an equation for the median survival time of

$$t_{50}\left(x, \underset{\sim}{\beta}, \sigma\right) = \exp(\beta_0 + \beta_1 x)$$

and the time ratio at the median time is

$$\mathrm{TR}(t_{50}, x = 1, x = 0) = \exp(\beta_1) . \qquad \ldots\ldots\ldots (3.24)$$

As expected with an accelerated failure time model, the exponentiated coefficient provides the acceleration factor on the time scale. In particular, if the time ratio in (3.24) is 2, the median survival time in the group with $x = 1$ is twice that of the group with $x = 0$. Since the percentiles of the survival time distribution in (3.21) are of the form

$$t_p\left(x, \underset{\sim}{\beta}, \sigma\right) = [(1-p)/p]^{\sigma} \exp(\beta_0 + \beta_1 x) .$$

The result in (3.24) holds at all values of time.

As was the case with the Weibull model, maximum likelihood is the method usually employed to fit a log-logistic model to a set of data subject to right censoring. It follows from results for the standard logistic distribution that the contribution of noncensored observations to the likelihood is

$$(1/\sigma)\exp(z)/[1+\exp(z)]^2 .$$

And that of censored observations is

$$[1+\exp(z)]^{-1}$$

where, for a multivariable model, $z=\left(y-X'\underset{\sim}{\beta}\right)/\sigma$. It follows that the log-likelihood function for a sample of n independent observations of time, covariates and censoring indicator, denoted $(t_i, X_i, \delta_i), i=1,2, \ldots, n$ is

$$L\left(\underset{\sim}{\beta},\sigma\right)=\sum_{i=1}^{n}\delta_i\{-\ln(\sigma)+z_i-2\ln[1+\exp(z_i)]\}-(1-\delta_i)\ln[1+\exp(z_i)]$$

The parameter estimation and the derivation of important statistics follow same step like the Weibull regression model discussed in section (3.3.5) (Hosmer and Lemeshow, 1998).

3.3.8 Model selection

We have presented different parametric models that can be used to study the effects of covariates on survival time. In this section, we focus on methods which help us for selecting the appropriate parametric model for the observed data using graphical displays and information criteria without covariates and after adjusting for the effect of covariates. First we check the appropriateness of the models graphically and then we compare the Akaikie information criterion (AIC) and the Bayesian information criterion (BIC) as final criterion to choose the best model which fits the data.

The graphical methods of comparison, without covariates, discussed here serve as empirical checks for the appropriateness of a given parametric model, not to "prove" that a particular parametric model is correct. In fact, in many applications, several parametric models may provide reasonable fits to the data and provide quite similar estimates of key quantities (Klein and Moeschberger, 1997).

For simplicity we shall examine the problem of checking for the appropriateness of a given model in the univariate setting i.e.; without covariates. The key tool is to find a function of the cumulative hazard rate which is linear in some function of time. The basic plot is made by estimating the cumulative hazard rate by the Nelson–Aalen estimator in (3.1). A function of the Kaplan-Meier survival function estimator could also be used for this purpose, as an alternative

to the Nelson-Aalen cumulative hazard function estimator. To illustrate this technique, consider a check of the appropriateness of the log-logistic regression model. Here, the cumulative hazard rate is $H(t) = \ln(1+\lambda t^{\alpha})$ a simple mathematical manipulation will give:

$$\ln\{\exp[H(t)]-1\} = \ln(\lambda)+\alpha\ln(t)\ . \qquad \ldots\ldots\ldots (3.25)$$

So, a plot of $\ln\left\{\exp\left[\hat{H}(t)\right]-1\right\}$ versus $\ln(t)$ should be approximately linear with intercept $\ln(\lambda)$ for the log-logistic model to be appropriate for the observed survival data.

For the other parametric models discussed, the following plots are made to check the fit of the models to the observed data.

For Log-normal regression model: $\Phi^{-1}\left[1-\exp\left(-\hat{H}(t)\right)\right]$ versus $\ln(t)$ ..... (3.26)

For Exponential regression model: $\hat{H}(t)$ versus t ..... (3.27)

For Weibull regression model: $\ln\left(\hat{H}(t)\right)$ versus $\ln(t)$ ..... (3.28)

where $\hat{H}(t)$ is the Nelson-Aalen cumulative hazard rate estimator.

Next, we discuss methods to compare parametric regression models for survival data after adjusting for covariates, using the Cox-Snell diagnostic plot and information criteria.

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. Central to the evaluation of model adequacy in any setting is an appropriate definition of a residual.

For all regression models, a specifically designed statistic to evaluate the accuracy of a postulated model (goodness-of-fit) is called a residual value. Residual values reflect the difference between the model-estimated values and the observed data that generated the model estimates. The first such residual is the Cox–Snell residual that provides a check of the overall fit of the model. The Cox-Snell diagnostic plot is used to assess the overall goodness of fit of

the model. The Cox-Snell diagnostic plot is the plot of the estimated Nelson-Aalen cumulative hazard function against the estimated Cox-Snell residuals, which is the same as the estimated cumulative hazard function from the model.

For example, the estimator of the cumulative hazard function for the Weibull regression model is:

$$\hat{H}\left(t_i, x_i, \underset{\sim}{\hat{\beta}}, \hat{\sigma}\right) = \exp\left[\left(y_i - X_i' \underset{\sim}{\hat{\beta}}\right) \Big/ \hat{\sigma}\right]$$

$$= \exp\left(\hat{z}_i\right)$$

$$\hat{H}\left(t_i, x_i, \underset{\sim}{\hat{\beta}}, \hat{\sigma}\right) = \left(t_i e^{-X_i' \underset{\sim}{\hat{\beta}}}\right)^{\hat{\lambda}}$$

for $i = 1, 2, \ldots, n$.

If the model is good, the plot of Cox-Snell residuals versus the Nelson-Aalen cumulative hazard estimates should lie along the 45 degree diagonal line that passes through the origin. Alternatively, one may calculate the Cox-Snell residuals from the martingale residuals as

$$\hat{H}\left(t_i, x_i, \underset{\sim}{\hat{\beta}}, \hat{\sigma}\right) = \delta_i - \hat{M}_i$$

where the martingale residuals for each observation is given by:

$$\hat{M}_i = \delta_i - \exp\left(\hat{z}_i\right)$$
$$= \delta_i - t_i^{\hat{\lambda}} \exp\left(-\hat{\lambda} X_i' \underset{\sim}{\hat{\beta}}\right)$$

which are used for checking the scale or linearity of continuous covariates in the model.

With the exception of the Weibull and log-normal regression models, which are nested under the generalized gamma model, it is difficult to use a formal statistical test to discriminate between parametric models for survival data because the models are not nested in a larger model which includes all the regression models (Klein and Moeschberger, 1997). In order to

compare alternative models fitted to an observed set of survival data, a statistic, which measures the extent to which the data are fitted by a particular model, is required. One way of selecting an appropriate parametric model is to base the decision on minimum Akaikie information criterion (AIC) and the Bayesian information criteria (BIC). The AIC is given by:

$$\text{AIC} = -2\log(\textit{likelihood}) + 2(p+k)$$

where $k = 1$ for the exponential model, $k = 2$ for the Weibull, log-logistic, and log-normal models and $k = 3$ for the generalized gamma model.

So, the model with the smallest AIC and BIC will be considered as the best fit for the observed survival data.

3.3.9 Variable Selection Procedures

Covariates may be selected for inclusion in a regression model for survival data using stepwise selection methods that operate in an identical manner to those used in other regression models, such as linear or logistic regression. The statistical test used as a criterion is most often the likelihood ratio test. However, the score test and Wald test are also often used by software packages. From the conceptual point of view it does not matter which test is actually used. However, the likelihood ratio test is the best of the three tests and should be used when there is a choice.

Stepwise selection process consists of forward selection followed by backward elimination. The forward selection adds to the model the covariate that is most statistically significant among those not in the model. The backward elimination process checks each covariate in the model for continued significance. Most software packages that have implemented stepwise selection of covariates treat all the covariates available for selection as if they were continuous. This implies that to consider nominal scale covariates with more than two levels correctly, one must create and include in the list of covariates the individual design variables. An additional problem is that the individual design variables are not considered as a unit, and the program

may select a subset of them. When this occurs, there has been an implicit recoding of the covariate and the user must make sure that the recoding makes clinical sense or must add the unselected design variables into the model when it is examined in more detail (The detail discussion is found in Hosmer and Lemeshow, 1998).

The stepwise procedure is described below, as it is currently implemented by default, using single degree-of-freedom tests for entry and removal of covariates. Here the test to be used is the likelihood ratio test; actually the default in STATA is likelihood ratio test.

Step 0: Assume that there are p possible covariates, denote $x_j, j = 1,2,\ldots,p$. This list is assumed to include continuous covariates as well as all design variables for nominal scaled covariates. At step 0 the likelihood ratio test and its p-value for the significance of each covariate is computed by comparing the log likelihood of the model containing x_j to the log likelihood of model zero (i.e., the model containing no covariates). The test statistic is

$$G^{(0)}(j) = -2\left[L^{(0)}(j) - L(0)\right], j = 1,2,\ldots,p \qquad \ldots\ldots\ldots (3.29)$$

where $L(0)$ is equal to the log likelihood of model zero, the no covariate model, and $L^{(0)}(j)$ is equal to the log likelihood of the model containing covariate x_j. The test's significance level is

$$p^{(0)}(j) = p\left[\chi^2(1) \geq G^{(0)}(j)\right]. \qquad \ldots\ldots\ldots (3.30)$$

Evaluation of (3.29) and (3.30) requires fitting p separate regression models for survival data. The parenthesized superscript in (3.29) and (3.30) denotes the step, and j indexes the particular covariate.

The candidate for entry into the model at step 1 is the most significant covariate and is denoted by x_{e_1}, where

$$p^{(0)}(e_1) = \min_j\left[p^{(0)}(j)\right].$$

For the variable x_{e_1} to be entered into the model, its p-value must be smaller than some pre-chosen criterion for significance, denote p_E. If the variable selected for entry is significant (i.e., $p^{(0)}(e_1) < p_E$), then the selection process continues to step 1; otherwise it stops.

Step 1: This step begins with variable x_{e_1} in the model. Then p-1 new regression models for the survival data (each including one remaining variable along with x_{e_1}) are fit, and the results are used to compute the likelihood ratio test of the fitted two-variable model to the one-variable model containing only x_{e_1},

$$G^{(1)}(j) = -2\left[L^{(1)}(j) - L(x_{e_1})\right],\ j = 1,2,\ldots,p \text{ and } j \neq e_1 . \qquad \ldots\ldots\ldots (3.31)$$

The p-value for the test of the significance of adding x_j to the model containing x_{e_1} is

$$p^{(1)}(j) = p\left[\chi^2(1) \geq G^{(1)}(j)\right] . \qquad \ldots\ldots\ldots (3.32)$$

The variable selected as the candidate for entry at step 2 is x_{e_2} where

$$p^{(1)}(e_2) = \min_{j \neq e_1}\left[p^{(1)}(j)\right] .$$

If the selected covariate x_{e_2} is significant, $p^{(1)}(e_2) < p_E$, then the selection process goes to step 2; otherwise it stops.

Step 2: This step begins with both x_{e_1} and x_{e_2} in the model. During this step, two different evaluations occur. The step begins with a backward elimination check for the continued contribution of x_{e_1}. That is, does x_{e_1} still contribute to the model after x_{e_2} has been added? This is essentially an evaluation of (3.31) and (3.32) with the roles of the two variables reversed. From an operational point of view, we choose a different significance criterion for this check, denoted p_R. We choose this value such that $p_R > p_E$ to eliminate the possibility of entering and removing the same variable in an endless number of successive steps. Assume the variable entered at step 1 is still significant.

The program fits p-2 regression models for survival data (each including one remaining variable along with x_{e_1} and $x_{e_2}^{j}$) and computes the likelihood ratio test and its p-value for the addition of the new covariate to the model, namely

$$G^{(2)}(j) = -2\left[L^{(2)}(j) - L\left(x_{e_1}, x_{e_2}\right)\right], \; j = 1,2,\ldots,p \text{ and } j \neq e_1, e_2$$

and

$$p^{(2)}(j) = p\left[\chi^2(1) \geq G^{(2)}(j)\right] \; .$$

The covariate x_{e_3} selected for entry at step 3 is the one with the smallest p-value, that is,

$$p^{(2)}(e_3) = \min_{j \neq e_1, e_2}\left[p^{(2)}(j)\right] .$$

The program proceeds to step 3 $p^{(2)}(e_3) < p_E$; otherwise it stops.

Step 3: Step 3, if reached, is similar to step 2 in that the elimination process determines whether all variables entered into the model at earlier steps are still significant. The selection process then followed is identical to the selection part of earlier steps. This procedure is followed until the last step, step S.

Step S: At this step one of two things may happen: (1) all the covariates are in the model and none may be removed or (2) each covariate not in the model has $p^{(s)}(j) > p_E$. At this point, no covariates are selected for entry and none of the covariates in the model may be removed.

The number of variables selected in any application will depend on the strength of the associations between covariates and survival time and the choice of p_E and p_R. Due to the multiple testing that occurs, it is nearly impossible to calculate the actual statistical significance of the full stepwise process. Research in linear regression indicates that use of $p_E = 0.05$ excludes too many important covariates and that one should choose a level of significance of 15 percent. In many applications it may make sense to use 25-50 percent to allow more variables to enter than will ultimately be used and then narrow the field of selected variables using $p < 0.15$ to obtain a multivariable model for further analysis (Hosmer and

Lemeshow, 1998). An unavoidable problem with any stepwise selection procedure is the potential for the inclusion of "noise" covariates and the exclusion of important covariates. This might be considered as the limitation of the selection procedure.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Twins Survival

Birth order of twins was not clearly indicated in the data set under consideration, but we used birth weight as a proxy variable to identify which one of the two is born first, that is because almost always the first born twin is heavier than the second born twin (Pawel et al, 2008).

There were 110 twins but after cleaning the data for missing birth weights and survival times, the number of twins reduced to 89 of which there were 63 deaths, that is infant mortality of 354 deaths per 1,000 live births among the twins.

The survival experience of the first and second born calculated based on the Kaplan-Meier survival function estimator are displayed in Figure 1. Similarly, the cumulative hazard estimates are displayed in Figure 2. As can be seen from Figure 1 the first born twin have better survival during the study period than the second born twin and from Figure 2, the hazard rate of the second twins is higher than the first born twins. In addition to the graphical comparison of hazard rates for the first and second born, we employed formal test of significance using matched-pair analysis for survival data. That is, using the statistic in (3.7), we test the hypothesis that there is no difference in the hazard rates between the first born twin and the second born twin; where D_1 (number of first-born twins died first) =12 and D_2 (number of second-born twins died first) =31. The value of the test statistic equals Z= 2.897 with p-value 0.004, so there is sufficient evidence that the hazard rates for the first and second born are different at 5% level of significance. The estimated survival probabilities and the corresponding 95% confidence interval at the end of their first year is 0.7143 (95% CI: 0.6069 to 0.7971) and 0.5634 (95% CI: 0.4523 to 0.6604) for the first-born and second-born respectively. Since twins cannot be treated as independent observations it is not logical to combine them with singletons in the study of infant survival. So, in this study we investigate factors affecting the mortality of singleton infants separately. However, it is suggested that one

can investigate factors affecting the survival of twin infants by using multivariate survival analysis, which will be a topic of further investigation.

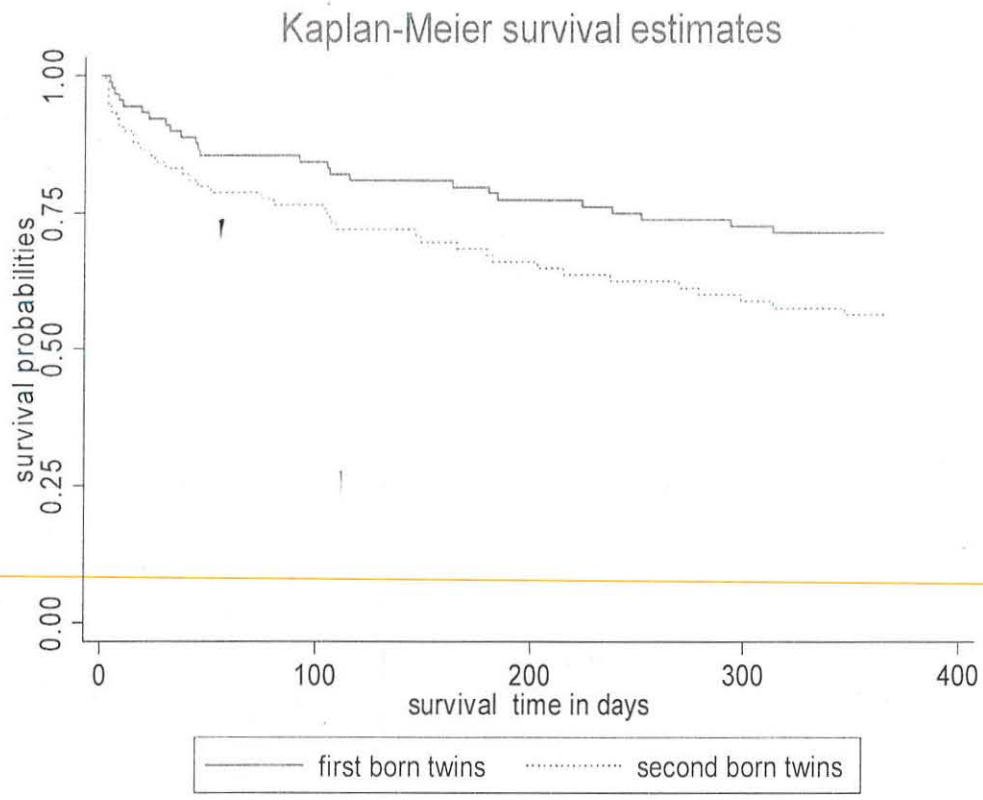

Figure 1. The Kaplan-Meier survival functions estimates of twins by birth order

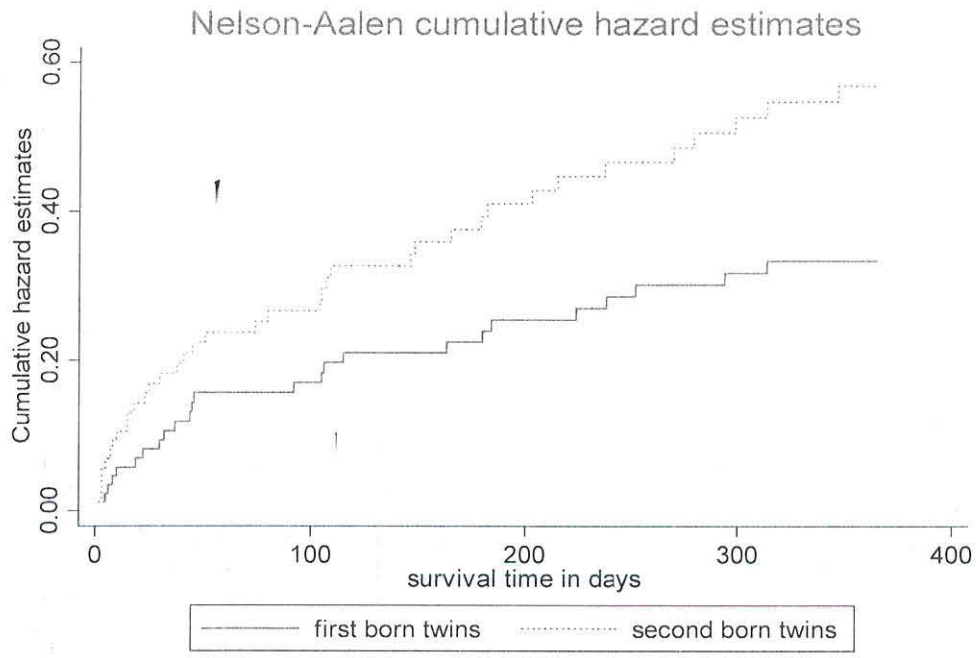

Figure 2. The Nelson-Aalen cumulative hazard functions of twins by birth order

4.2 Data Analysis for Singletons

Data analysis for the survival of singletons begins by providing descriptive summary measures followed by bivariate analysis of the categorical variables and grouped continuous covariates using log-rank test and then fitting both un-adjusted bivariate and multivariate parametric models. The results in the un-adjusted bivariate model describe the gross effect of the covariates on infant mortality, while the results in the multivariate model describe the net effect, that is, the effects of each covariate after controlling for the effects of other variables in the model. In this study a variable is considered significantly associated with mortality when its p-value is less than 0.05. The likelihood ratio test was used to examine the strength of the association between independent variable and infant mortality.

Some continuous variables were categorized for comparison of infant mortality among the various categories. For example, age of mother at birth was categorized into three age groups such as: less than 20, 20-34 and 35 and more. These breakpoints were selected because of the

need to cover the different reproductive trajectories: adolescence, young adult and adult women. On the other hand categorized variables were further edited by combining some of their groups in one or two groups either because of the small number of observations in those categories or to make the analysis and the interpretation more meaningful. For instance, in the variable "Mother's ethnic group", ethnic groups like Amhara, Tigre, Gurage and so on were all combined in one group because of the small number of observations in those ethnic categories relative to the major Oromo ethnic group in the study area. Marital status is recoded as currently married and not currently married.The rest of the variables are used in the same way as they have been used by Assefa and Tesema (1997) who originally collected the data and used Cox proportional hazards model except that in this study the covariate family size is treated as continuous variable.

4.3 Descriptive Analysis

The principal weakness here is the small size of observed deaths relative to censored observations on which inferences are based. The total number of observations used in this analysis, after cleaning the data for missing values, contains 7,851 singleton births, of which 661 died during the one year study period. This means that we have only 8.42 percent complete information about the survival times of the infants or equivalently, more than 91 percent of the information on survival times of the infants is incomplete or censored.

The majority of the deaths took place early. Approximately 22 percent of the total number of deaths occurred in the first month of life, which is the neonatal period. The highest risk for infant death is in this period. Low birth weight maternal complications in labor contribute significantly to neonatal mortality (Karsten, 2002).

Summary statistics based on quartiles are misleading; this is due to extremely large censored observations. The value of any estimated time percentile, such as the median survival time, depends on the estimated value of the constant term $\left(\hat{\beta}_0\right)$ in (3.14), which, in turn, depends on the observed number of actual survival times, which is the observed number of deaths.

Thus, the estimated time percentile values are only useful for descriptive purposes if the observed survival times are about 80 percent of the data. But the ratio of any two time percentiles depends only on the coefficient of the covariate of interest. See equation (3.15). Hence, the ratio is a more useful summary measure of survival experience than estimates of percentiles in this study, where only 8.42 percent of the data with the observed survival times.

The overall survival and cumulative hazard functions are given in Figure 3 and 4. As can be seen from Figure 3, the estimated survival probability of singleton infants at the end of their first year of life is about 0.9123 with standard error of 0.0033 and 95% confidence interval between 0.9056 and 0.9185. That is, infant mortality rate is estimated to be about 88 deaths in 1000 live births.

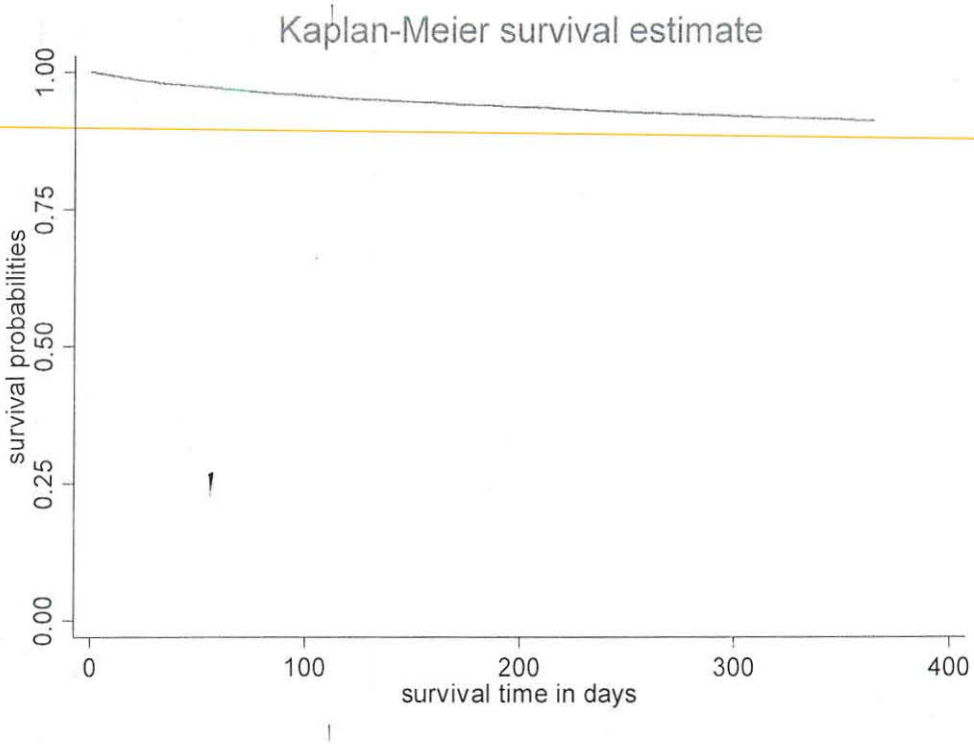

Figure 3. The overall estimated Kaplan-Meier survival function for singleton

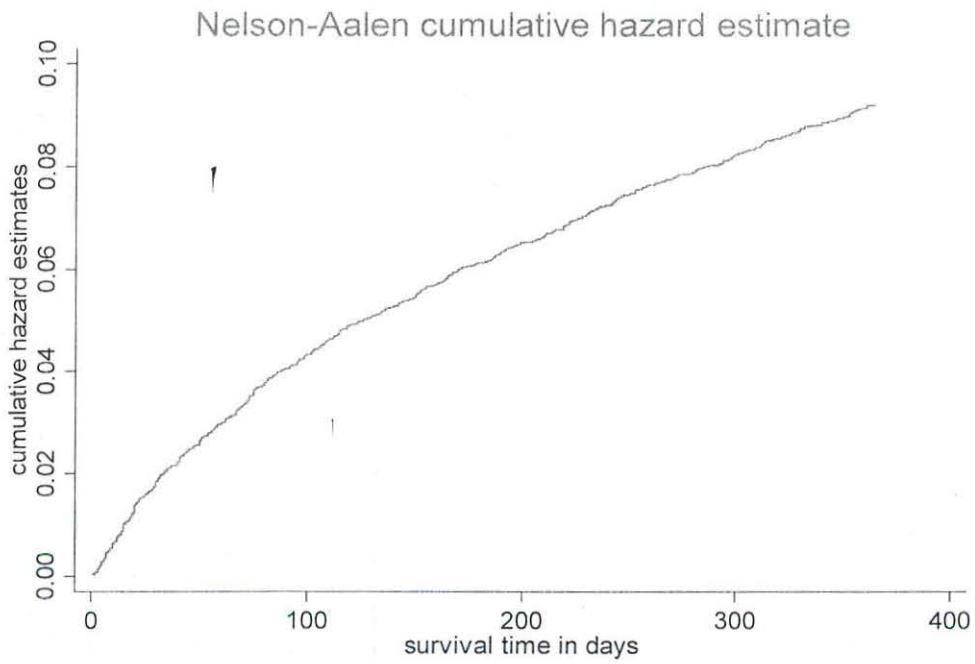

Figure 4. The Nelson-Aalen estimated cumulative hazard function for singletons

Next, we compare the survival experience of infants graphically based on Kaplan-Meier survival estimates and using the log-rank test. In order to compare the survival of groups of subjects graphically, we plot the Kaplan-Meier estimator of the survival function for each of the groups. In general, the pattern of one survivorship function lying above another means the group defined by the upper curve had a better survival than the group defined by the lower curve. The detailed discussion is found in Hosmer and Lemeshow (1998).

As can be seen from Figure 5 (Appendix A), the survival curve of females lies above the survival curve of males showing better survival of females than males. The reduction in survival times of infants whose mothers are illiterate is reflected in Figure 6. Figure 7 shows the importance of marriage for the better survival of infants, particularly during the post-neonatal period. Figure 8 could be an important evidence of the influence of birth weight for the survival of infants. Infants who weighs at least 2.5kg at birth better survive their first year of life than those who weigh less than 2.5kg at birth (See appendix A for Figures 5 -8).

4.4 Log-rank test for comparing infant survival

The difference in survival outcomes between the different levels of each covariate is tested using the non-parametric log-rank test. The results are presented in Table 2.

Table2. Log-rank test results for the equality of infant survival in the different categories of covariates.

Variables	Log-Rank Statistics	Df	Significance level
Momeduc	21.73	2	0.0000
Momage	3.38	2	0.1845
Site	7.55	1	0.0060
Weight	117.85	1	0.0000
Sex	5.81	1	0.0161
Mometh	5.19	1	0.0227
Momrel	5.06	1	0.0244
Mstatus	16.23	1	0.0001
Parity	5.77	2	0.0560
Water	5.27	2	0.0715

The results of the log-rank test in Table 2 indicate that for all the variables except age of mothers at birth, parity and source of drinking water; the hypothesis that the survival functions or hazard rates are equal among the groups was rejected at 5% significance level. That is, infant survival from group to group varies significantly in most of the covariates under consideration. This does not, however, mean that those variables which are not significant such as age of the mother, parity and source of drinking water, are not important and that they should not be considered for further analysis. But since these covariates may have a joint effect when they are taken together with other explanatory variables it is necessary to consider them in the multivariable analysis. Similarly, as suggested in the

literature, see for example Hosmer and Lemeshow (1998), using a modest level of significance of 20% to include potential covariates in the multivariable analysis all the covariates listed in Table 2 are eligible to be included in the multiple covariate analysis in the next sections.

4.5 Parametric models comparison

In this section, we present parametric model comparisons using graphical displays and information criteria without covariates and after adjusting for the effects of covariates. The choice of fitting parametric models than the most widely used Cox-proportional hazards model is due to the fact that the assumption of proportionality is violated for most important covariates. Even after using stratified proportional hazards model as a possible solution, the assumption still failed. See detailed results in Appendix B.

We start model comparison graphically. The following four graphs are based on equations (3.25) to (3.28) for the most commonly used parametric models such as exponential, Weibull, log-logistic and log-normal. Figures 9-12 are obtained without taking into account covariates, using the relationship between the Nelson-Aalen cumulative hazard rate estimates and survival time for each model. Figure 9 is a plot of cumulative hazard estimates versus time for the exponential model. If the fit is good, we expect a straight line through the origin. Figure 10 is a plot of log of cumulative hazard estimates versus log of analysis time for the Weibull model. If the fit is good we expect approximate linear relationship with an intercept. Figure 11 is a plot of probit of one minus the survival estimates versus log-of analysis time for the log-normal model. If the fit is good we expect approximate linear relationship with an intercept. Figure 12 is a plot of the log-of exponentiated cumulative hazard minus one versus log-of analysis time for the log-logistic model. If the fit is good we expect approximate linear relationship with an intercept.

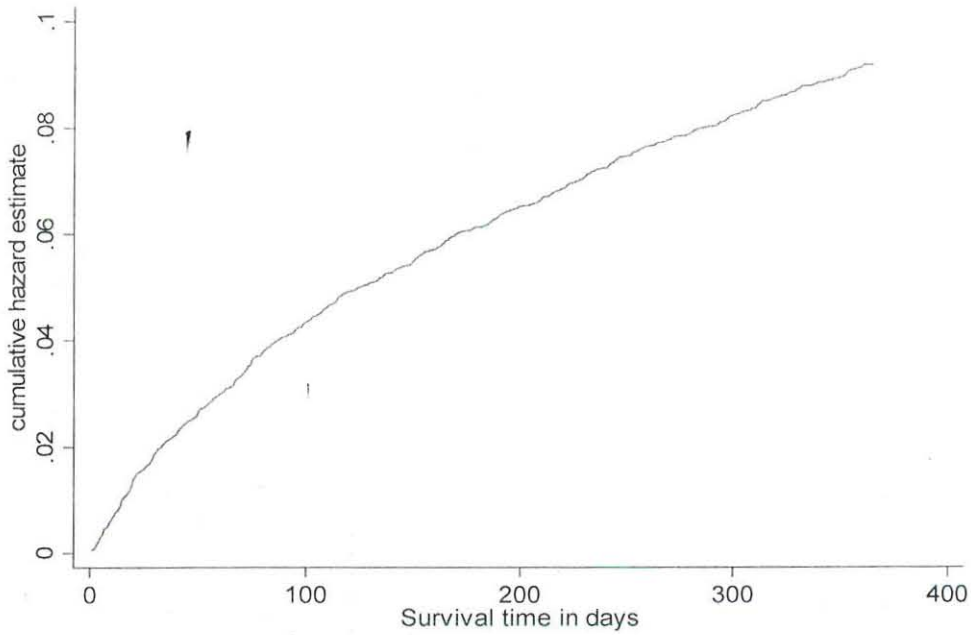

Figure 9. The plot to check the appropriateness of exponential model

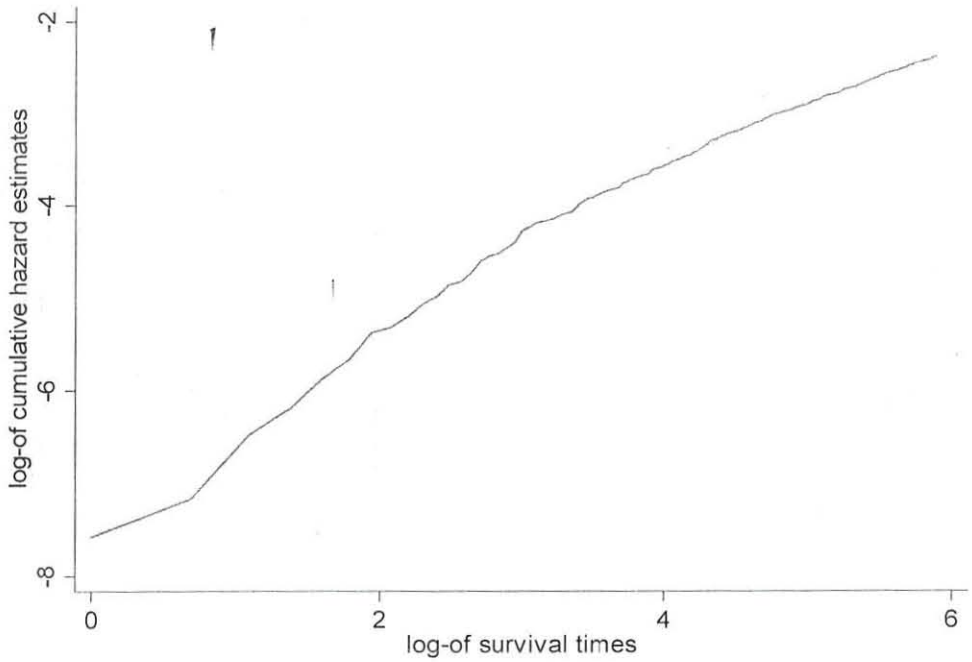

Figure 10. The plot to check the appropriateness of Weibull model

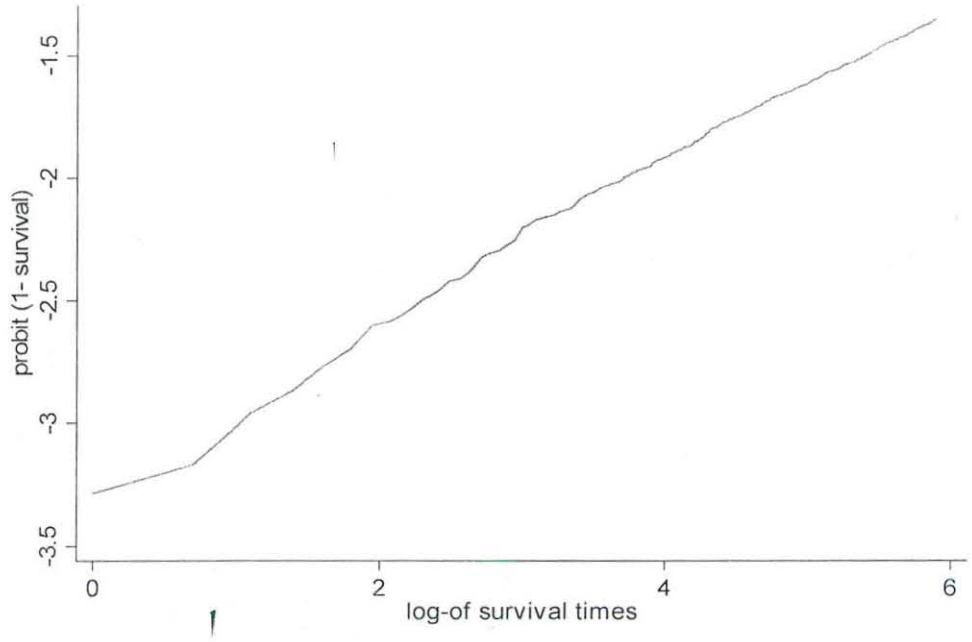

Figure 11. The plot to check the appropriateness of log-normal model

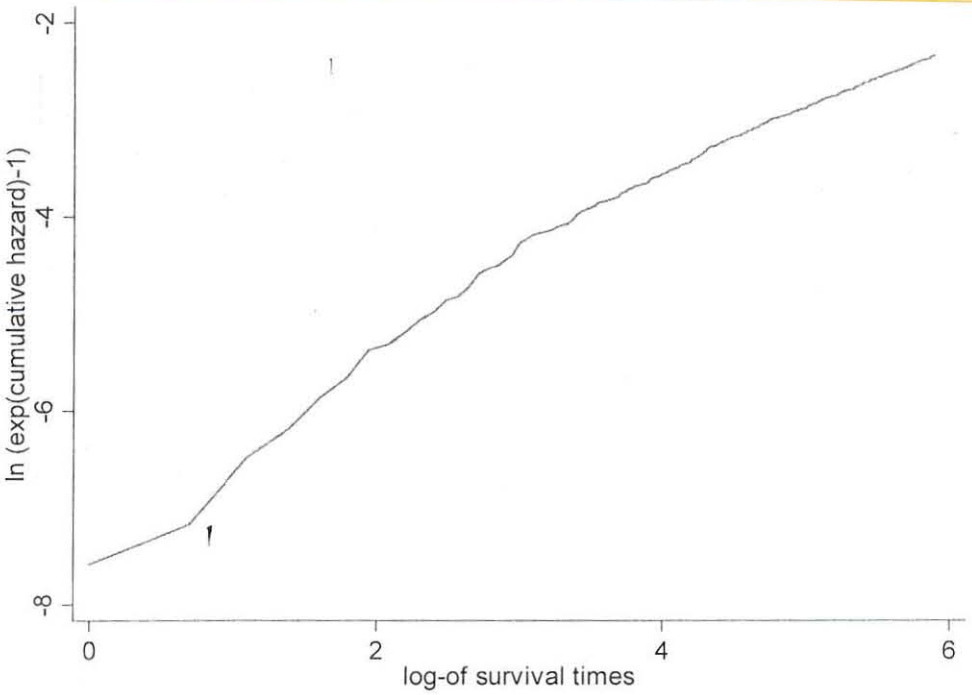

Figure 12. The plot to check the appropriateness of log-logistic model

As can be seen from Figure 9, exponential regression model seems not appropriate for the data compared with the remaining three regression models. Even though, the remaining three seems equally appropriate for the observed data, based on visual inspection slightly the appropriateness of the log-normal model (Figure 11) outshines.

Next, we consider model comparison after adjusting for the effect of covariates. In this case the graphical displays are based on the Cox-Snell plots. That is, if the model is good, the plot of Cox-Snell residuals versus Nelson-Aalen cumulative hazard estimates should lie along the 45 degree diagonal line that passes through the origin. Using all the covariates presented in Table 2, we fitted three parametric regression models the Weibull, log-logistic and log-normal models and the detail results are presented in Appendix B. Here we present the Cox-Snell plots for model comparison in Figures 13 to 15.

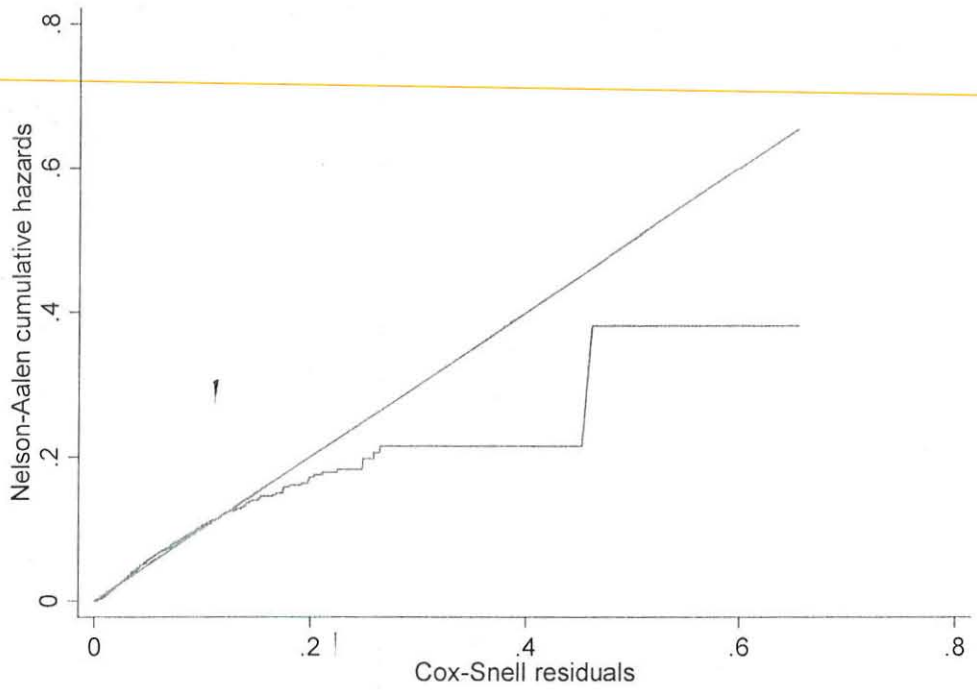

Figure 13. The Cox-Snell plot after fitting Weibull regression model

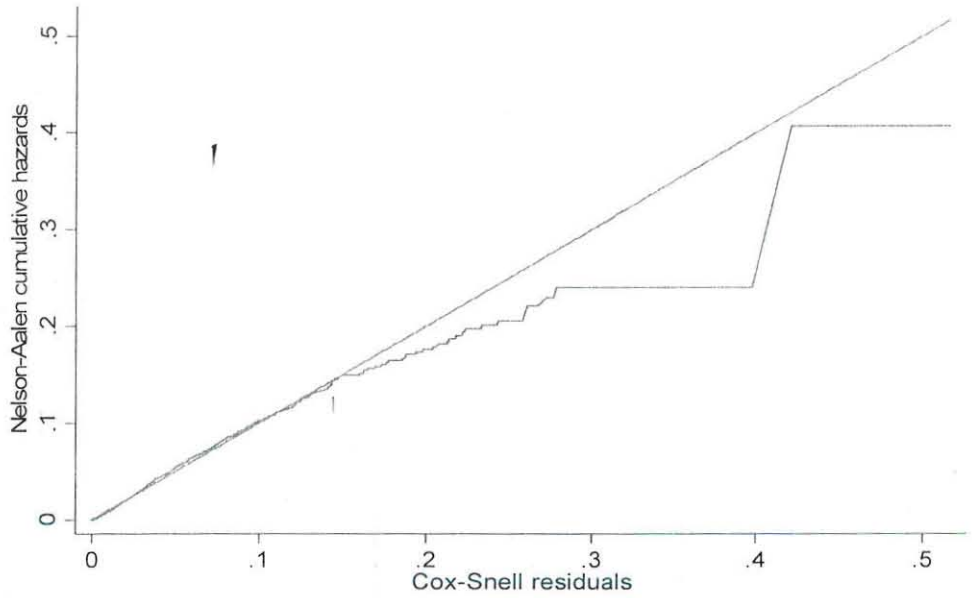

Figure 14. The Cox-Snell plot after fitting log-normal regression model

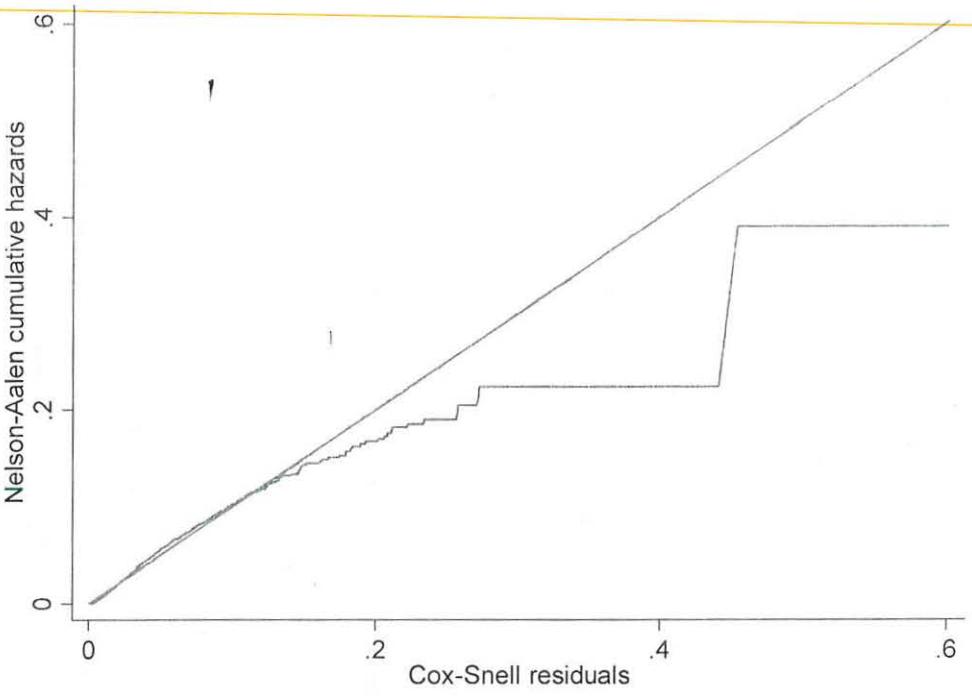

Figure 15. The Cox-Snell plot after fitting log-logistic regression model

From the Cox-Snell plots in Figures 13 to15 it is observed that log-normal regression model seems the best fit among the three models.

In addition to the graphical comparison of the three parametric regression models, we used Akaikie information criterion (AIC) and Bayesian information criterion (BIC) to choose the best model out of the four possible models. The results are shown in Table 3.

Table 3. Statistics for model comparison

Model	Obsn.	ll(null)	ll(model)	Df	AIC	BIC	LR	LR p-value
Exponential	7851	-3238.30	-3160.29	10	6340.58	6410.27	156	0.0000
Weibull	7851	-3181.24	-3105.25	11	6232.49	6309.15	151	0.0000
Log-normal	7851	-3167.16	-3079.59	11	6181.18	6257.83	175	0.0000
Log-logistic	7851	-3179.65	-3100.37	11	6222.75	6299.40	158	0.0000

Where in Table 3, LR is the value of the likelihood ratio test statistic and LR p-value is the corresponding p-value of the test.

According to the results in Table 3, the log-normal model with the smallest value of AIC and BIC seems to be the best fit of the four models. Thus, the log-normal regression model is considered further to discuss the effect of covariates on infant survival.

To check for model adequacy, we used the likelihood ratio test. The results are shown in Table 3. All the p-values of the likelihood ratio tests in Table 3 showed, all the four models are good fit for the data. However, based on the Akaikie information criterion (AIC) and Bayesian information criterion (BIC) we choose log-normal as the best fit for the observed survival data.

4.6 Bivariate un-adjusted log-normal regression model

To have an idea about the individual effects of the different explanatory variables on infant survival, we fitted log-normal regression model separately for each explanatory covariate. The results are shown in Table 4.

Table 4. The result of un-adjusted bivariate analysis using log-normal regression model

Covariates	coef.	std.Err	p-value	TR	95% CI of TR	
Family size	0.1617	0.0315	0.000	1.1755	1.1050	1.2505
Maternal Education						
Illitrate						
1-6 grade	0.6538	0.1643	0.000	0.1907	1.3935	2.6533
Secondary & above	0.7586	0.1907	0.000	2.1354	1.4693	3.1034
Place of residence						
Urban						
Rural	0.4109	0.1296	0.002	0.6630	.5143	.8548
Maternal age at child birth						
< 20 years						
20 - 34 years	0.3801	0.1918	0.047	1.4627	1.0043	2.1297
> = 35 years	0.3508	0.2339	0.134	1.4202	.8980	2.2461
Sex of the child						
Female						
Male	0.3056	0.1277	0.017	0.7367	.5736	.9461
Ethnicity						
Oromo						
Others	0.375	0.1397	0.007	1.4550	1.1066	1.9131
Mother's religion						
Christian						
Muslim	0.4028	0.143	0.005	0.6684	.5050	.8847
Marital status of the mother						
Married						
Not married	0.8166	0.2293	0.000	0.4419	.2820	.6927
Birth weight of the infant						
>=2,500gm						
<2,500gm	1.9171	0.1851	0.000	0.1470	.1023	.2114
Parity						
Primi						
2 - 4 children	0.3962	0.1587	0.013	1.4862	1.0890	2.0283
> = 4 children	0.4245	0.1704	0.013	1.5289	1.0947	2.1353
Source of drinking water						
Pipe						
Protected well/spring	0.1752	0.1919	0.361	0.8393	.5761	1.2227
Unprotected source	-4098	0.1769	0.021	0.6638	.4693	.9388

Note: all the first groups of categorical covariates are used as reference groups.

TR stands for time ratio.

From the un-adjusted lognormal regression it can be seen that all the covariates including mother's age and source of drinking water have significant gross effect on infant survival. In particular, the un-adjusted model reveals that, the survival of infants of young adults (age between 20 and 34 years) is much better than infants of adolescent mothers (age less than 20 years). And there is no significant difference in survival of infants of adult women (age > 35 years) and adolescent mothers (age less than 20 years). This difference was not captured by the non-parametric log-rank test. Similarly, the survival times of infants who drink water from unprotected sources is pretty much reduced relative to those who drink pipe water. And there is no statistically significant difference in survivorship between infants who drink pipe water and protected well/spring water. This difference was also not revealed by the non-parametric log-rank test.

4.7 Multiple covariate log-normal regression model for infant survival

When there are a number of explanatory variables of possible relevance, the effect of each term cannot be studied independently of the others. The effect of any given term therefore depends on the other terms currently included in the model. However, in the bivariate analysis technique the relations which are obtained for one factor do not take into account the other factors. So the multivariable analysis is used to know the most important factors associated with infant mortality in relation to the covariates included in the model.

We fitted log-normal regression model using all covariates and the results are presented in Table 5.

Table 5. Summary of parameter estimates for the preliminary model

Covariates	coef.	std.Err	p-value	TR	95% CI of TR	
Family size	0.1633	0.0394	0.000	1.1773	1.0898	1.2720
Maternal Education						
Illitrate						
1-6 grade	0.5905	0.1722	0.001	1.8049	1.2878	2.5296
Secondary & above	0.6749	0.2245	0.003	1.9638	1.2647	3.0495
Place of residence						

Urban						
Rural	0.1438	0.1831	0.432	1.1546	.8064	1.6532
Maternal age at child birth						
< 20 years						
20 - 34 years	0.1962	0.2171	0.366	1.2167	.7950	1.8622
> = 35 years	0.1134	0.2897	0.695	1.1201	.6349	1.9762
Sex of the child						
Female						
Male	-0.3973	0.1245	0.001	0.6721	.5265	.8579
Ethnicity						
Oromo						
Others	0.0471	0.1946	0.809	1.0483	.7159	1.5349
Mother's religion						
Christian						
Muslim	-0.1234	0.2056	0.548	0.8839	.5907	1.3225
Marital status of the mother						
Married						
Not married	-0.8807	0.2302	0.000	0.4145	.2640	.6508
Birth weight of the infant						
>=2,500gm						
<2,500gm	-1.7629	0.1873	0.000	0.1715	.1188	.2476
Parity						
Primi						
2 - 4 children	0.0325	0.1813	0.858	1.0330	.7241	1.4738
> = 4 children	-0.1739	0.2423	0.473	0.8404	.5226	1.3513
Source of drinking water						
Pipe						
Protected well/spring	0.0464	0.1928	0.810	1.0474	.7179	1.5283
Unprotected source	0.0914	0.2067	0.659	1.0957	.7307	1.6431
Constant term	9.1935	0.3825	0.000		8.4438	9.9432
-ln(sigma)	1.1119	0.0336	0.000		1.0461	1.1778
Sigma	3.0402	0.1022			2.8464	3.2473

Note: all the first groups of categorical covariates are used as reference groups.

For the covariate, family size, included in the analysis we checked the assumption of linearity using martingale residuals. The plot in Figure 16 presented in Appendix A suggests that the assumption of linearity is satisfied for the covariate family size. In addition, we checked for possible meaningful interaction terms to be included in the model. Among many interaction terms considered, the interaction between place of residence and ethnicity was statistically

significant. Hence, this interaction term was included in the model along with all other covariates.

In order to select the most important covariates in the final model, we used stepwise variable selection as discussed in section 3.3.9. The results of the stepwise selection are displayed in Table 6.

Table 6. Summary of parameter estimates of the final model

Covariates	coef.	std.Err	p-value	TR	95% CI of TR	
Family size	0.138983	0.030655	0.000	1.1491	1.0821	1.2203
Maternal Education						
Illitrate						
1-6 grade	0.616231	0.169393	0.000	1.8519	1.3287	2.5812
Secondary & above	0.695736	0.212287	0.001	2.0052	1.3227	3.0399
Place of residence						
Urban						
Rural	0.334965	0.183019	0.067	1.3979	.9765	2.0011
Sex of the child						
Female						
Male	-0.39163	0.124435	0.002	0.6760	.5297	.8627
Ethnicity						
Oromo						
Others	0.353819	0.195129	0.070	1.4245	.9718	2.0881
Marital status						
Married						
Not married	-0.85926	0.226403	0.000	0.4235	.2717	.6600
Birth weight of the infant						
>=2,500gm						
<2,500gm	-1.77674	0.186633	0.000	0.1692	.1174	.2439
Interaction term						
Rural X non-Oromo	-0.68848	0.325636	0.034	0.5023	.2653	.9510
Constant term	9.239217	0.28453	0.000		8.6816	9.7969
-ln(sigma)	1.111933			1.1117	1.0458	1.1776
Sigma	3.040228			3.0395	2.8458	3.2465

Note: all the first groups of categorical covariates are used as reference groups.

Hence, Table 6 presents the final log-normal regression model results of the determinants of infant mortality. The results include the estimates of coefficients of the model, the standard error of the coefficients and the corresponding p-values, the estimated time ratio and the

corresponding 95% confidence interval of the critical predictors of infant mortality in this study. As can be seen from Table 6, family size, maternal education, marital status, sex of the child and weight of the child at birth show statistically significant association with infant mortality. In addition, model result shows significant interaction between ethnicity and place of residence.

4.8 Discussion

The cumulative proportion surviving to the end of the first year was 0.9123; this is an infant mortality rate of around 88 deaths per 1,000 live births, which is smaller when compared with the country level estimated infant mortality rate of 116 deaths per 1,000 live births in the 1994 national census of Ethiopia.

In this study larger family size was beneficial for infant survival. The effect of an increase in the size of the family by one person increases survival times of infants by an estimated amount of 15 percent (95% CI: 8% to 22%). This finding needs further investigation.

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 (Hill et al, 2001, DHSA, 2006). The survival time of males is 67 percent of the survival time of females (95% CI: 53% to 86%) during infancy. In addition this result is in line with the 2005 DHS report that, more boys die before their first birth day than girls in Ethiopia.

There is a higher mortality in children whose mothers were not educated. Elementary education had decreased the probability of dying before the first year of life significantly by an estimated amount of 85 percent compared to mothers with no education (95% CI: 33% to 158%). And the survival times of infants whose mother's had secondary and above education

is estimated to be twice that of infants of illiterate mother's (95% CI: 1.32 to 3.04). Alternatively, this means that as the level of maternal education increases to secondary and above, the survival of infants is also increases relative to illiterate mothers. This increase is estimated to be between 32 percent and 204 percent with a probability 0.95.

The risk of infant death is higher in mothers who are not currently married than currently married. The survival times of infants whose mother's are not currently married is estimated to be approximately 42 percent of the survival times of infants of currently married mothers (95% CI: 27% to 66%). This result suggests a critical role of fathers in providing financial assistance and emotional support for the mother.

Early weights of infants strongly predicted mortality, as one would expect (Kulmala, 2000). The mortality in the low birth-weight group (weight <2,500 g) is higher than in those weighing above 2,500g. In this study it was found that survival times of underweight infants is estimated to be approximately only 17 percent of infants who weigh at least 2,500g. The main risk factors for perinatal mortality were being low birth weight and prematurity (Sahle Mariam and Berhane, 1997). Neonatal illness in general closely related to low birth weight, low birth weight babies also tend to have higher mortalities (Kulmala, 2000)

The fitted model result also shows significant interaction between ethnicity and place of residence. It indicates that the relationship between ethnicity and infant mortality is affected by the level of a third factor which is place of residence (effect modifier). Thus, we cannot generalize the effect of ethnicity on infant mortality without specifying the place of residence where the child grew. In rural areas infants who belong to Oromo ethnic group survive better than others; the survival of infants whose mother doesn't belong to Oromo ethnic group is about 72 percent of the survival times of infants of Oromo ethnic group. But in urban areas infants of other ethnic groups survive better than infants of Oromo ethnic group; that is the survival times of infants of Oromo ethnic groups is only 42 percent of the survival times of

infants of other ethnic groups. These ethnic differences in infant mortality merely reflect socio-economic and households environmental differences between the members of the ethnic groups. In general, when adjusting for the effect of place of residence infants of other ethnic groups better survive compared with infants of Oromo ethnic groups (see Figure 21).

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

In this study, matched pair analysis is used to compare the hazard rate of first-born and second-born twins. It was found that the hazard rates for the first and second born twins are statistically different ($p < 0.05$). The Kaplan-Meier survival estimate for the first born is 0.7143 with standard error of 0.0484 and for the second born twin is 0.5634 with standard error of 0.0535, that is, the survival of the second-born twin is about three fourth of the survival of the first-born twin. It is reasonable not to treat twins as independent observations. So, it is not logical to combine them with singletons in the study of infant survival. We investigate factors affecting the mortality of singleton infants separately using parametric regression models.

This study has examined the socioeconomic, demographic and biological determinants of infant mortality in South-West Ethiopia. Results from the fitted log-normal regression model show biological and demographic variables are more important determinants of infant and post neonatal mortality.

The predictors of the survival probability are dominated by biological and demographic factors. Birth weight is the most important predictor of infant mortality; infants having birth weight less than 2.5 kg are at a greater risk of dying relative to infants having birth weight greater than 2.5kg. The survival times of infants whose weight is less than 2.5kg at birth is reduced by more than 80 percent relative to infants whose weight is greater than 2.5kg at birth.

Birth weight is lower among children of women with no education. In this study, 77.14 percent of infants whose weight at birth is less than 2.5kg belong to illiterate mothers, 16.24 percent

belong to mothers who had elementary education and only 6.62 percent belong to mothers who had secondary and above education. In addition, family size, sex and marital status have strong association with infant mortality.

Maternal education is also the most important predictor of infant mortality found in this study. Primary education of mothers improves the survival of their infants by an estimated amount of 85 percent compared with illiterate mothers. Infant mortality among children born to mothers with no education is twice that of children born to mothers with secondary and higher level of education.

The findings suggest policies and efforts have to be put in place to improve women education. If women, who are the primary caretakers of children, are empowered through education, the health and survival of their children will be enhanced.

References

1. Alan Guttmacher Institute (AGI), (2002). Issues in Brief, 2002 Series, No. 2.
2. Anthony B., Colleen O., Vidia P., K. S. Joseph, David C., and Thomas F. (2006). *Determinants of Perinatal Mortality and Serious Neonatal Morbidity in the Second Twin.* By the American college of Obstetricians and Gynecologists. Vol.108, No. 3, Part 1.
3. 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.
4. Caldwell J. C. (1979). *Education as a Factor in Mortality Decline: An Examination of Nigerian Data.* Population Studies 33(3):395-413.
5. CSA. (1996). *The 1994 Population and Housing census of Ethiopia.* Statistical report, 1996; I.
6. Department of Health South Australia (DHSA) (2006). Maternal, Perinatal and Infant Mortality Committee (MPIMC).
7. Ethiopian Demographic and Health Survey (EDHS) (2005). Central Statistical Agency, Addis Ababa, Ethiopia and ORC Macro, Calverton, Maryland, USA.
8. Forste R. (1994). *The Effects of Breastfeeding and Birth Spacing on Infant and Child Mortality in Bolivia.* Population Studies 48:497-511.
9. Hacking D., Watkins A., Fraser S., Wolfe R., Nolan T. on behalf of Australia and New Zealand Neonatal Network. (2001). *Respiratory distress syndrome and birth order in premature twins.* Arch Dis Child Fetal Neonatal Ed; 84:F117–F121
10. 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
11. 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 ISSN 1934-7227 Volume 2, Issue 2: 1-16.
12. Hobcraft J. (1993). *Women's education, child welfare and child survival: a review of the evidence.* Health Transition Review 3(2):159-173.
13. Hosmer D.W, Lemeshow S. (1998). *Applied Survival Analysis Regression Modeling of Time to Event Data.* John Wiley and Sons, Inc. New York.

14. Jatran S. (2002). Infant Mortality in a Backward Region of North India: Does Ethnicity Matter? Asian MetaCentre Research Paper Series. No. 14.
15. Jayachandran J., Jarvis G. (1988). Socioeconomic developed countries. Social Biology, 1988; 33(3-4): 301-315.
16. Karsten K. (2002). *The Missing Link: Neonatal Mortality in Rural Communities*. Bulletin of the WHO 80(9); 759-760.
17. Klein, J.P. and Moeschberger, M.L. (1997), Survival Analysis Techniques for Censored and Truncated Data. Springer-Verlag. New York.
18. Kulmala T. (2000). Maternal health and pregnancy outcome. University of Tampere, School of Public Health Printing Press.
19. Lawn JE, Cousens S, Zupan J. (2005). *4 million neonatal deaths: When? Where? Why?* Lancet; 365(9462):891-900.
20. Madise, N. (2003). Infant mortality in Zambia: Socioeconomic and demographic correlates. Social Biology. Accessed from www.findarticles.com/p/articles/mi_qa3998
21. Masuy-Stroobant G. (2001). The determinants of infant mortality: *how far are conceptual frame works really modeled?* Institut de demographie, Universite Catholique de Louvain (UCL).
22. Mturi A.J. and Curtis S.L.. (1995). *The determinants of infant and child mortality in Tanzania*. Health Policy Plan 10:384-94.
23. Mutunga C. (2004). *Environmental Determinants of Child Mortality in Kenya*. Kenya Institute for Public Policy Research and Analysis (KIPPRA), Nairobi, Kenya.
24. Pawel K., Maria K., Jaroslaw K., Anita C., Malgorzata P. (2008). *Adaptation period of the first and second twin-comparative analysis of somatometric, biochemical, and clinical parameters*. Archives of Perinatal Medicine 14(4): 28-32
25. Sahle Mariam Y. and Berhane Y. (1997). *Neonatal Mortality among Hospital Delivered Babies in Addis Ababa Ethiopia*. Eth. J. of Health Development 11(3): 275-281, 1997.
26. Shamebo D., Muhe L., Sandström A., Freij L., Krantz I., Wall S. (1994). The Butajira Rural Health Project in Ethiopia: a nested case-referent (control) study of under-5 mortality and its health and behavioral determinants. Ann Trop Pediatric 1994; 14(3):201-9.

27. Shamebo D., Sandstrom A., Muhe L., Freij L., Krantz I., Lomberg G., and Wall S. (1993) The Butajira rural health project in Ethiopia: *A nested case-referent study of under-five mortality and its public health determinants*. Bulletin of the WHO 1993; 71(3/4): 389-396.
28. UNICEF (1999). State of World's Children 1999.
29. UNICEF (2009). The state of world children 2009.
30. UNICEF. (2006): State of World's Children 2006.
31. World Bank. (1993). World development report, investing in health: world development indicators Oxford, Oxford University Press, 1993.

APPENDIX A

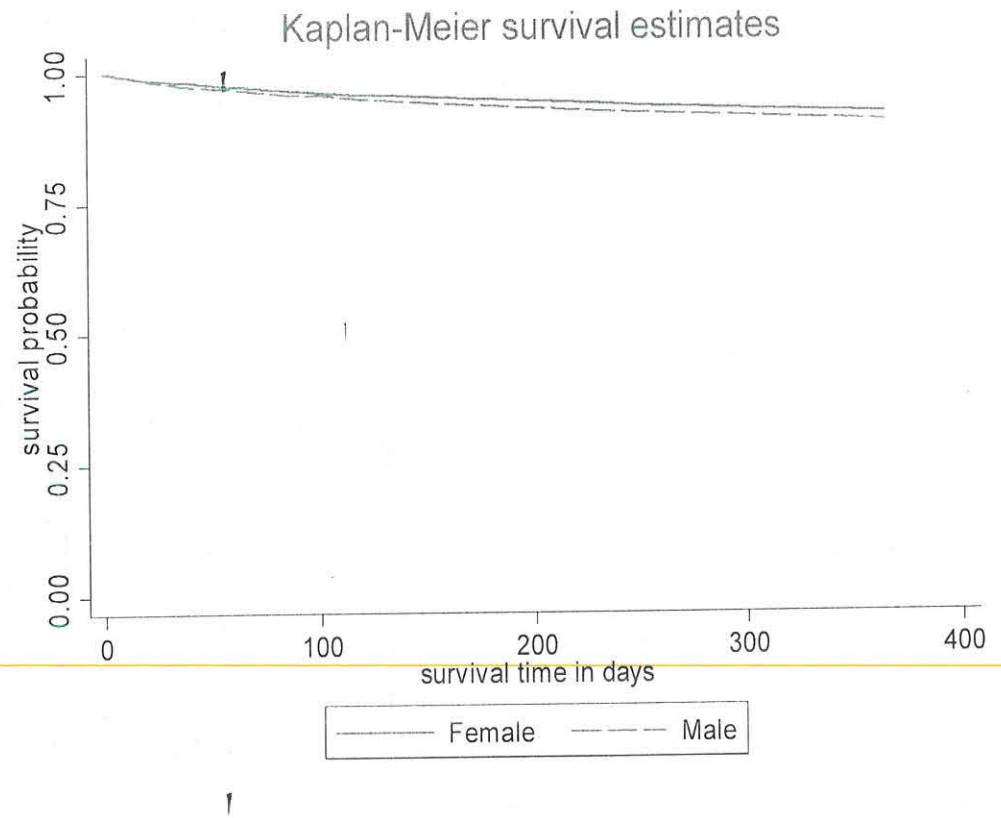

Figure 5. The Kaplan-Meier survival functions estimates by sex

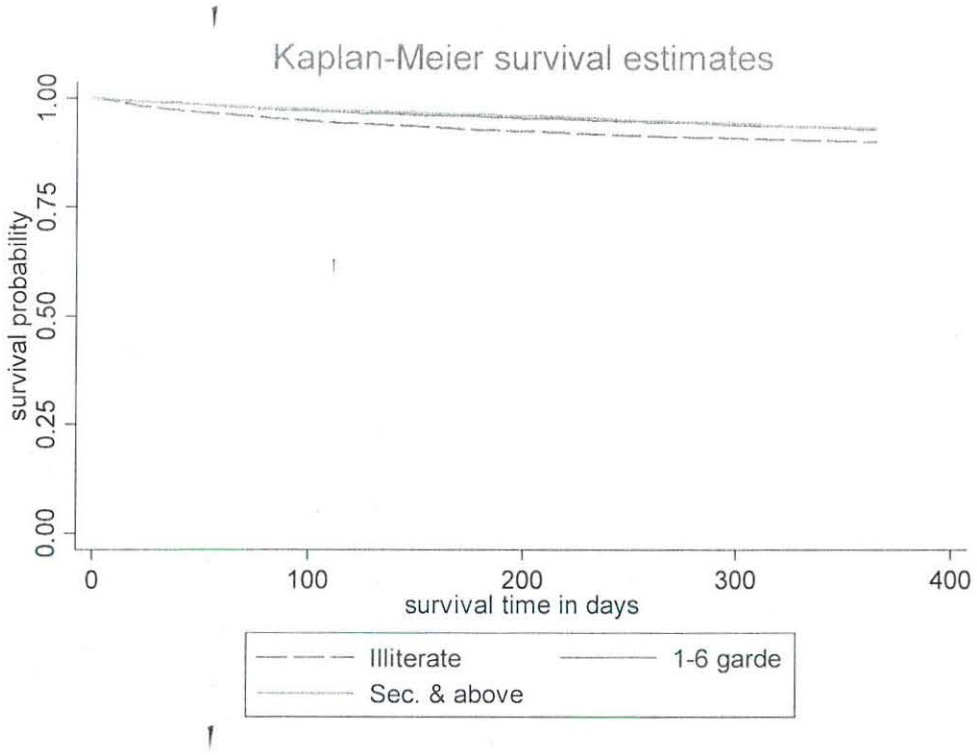

Figure 6. The Kaplan-Meier survival functions estimates by mother's education

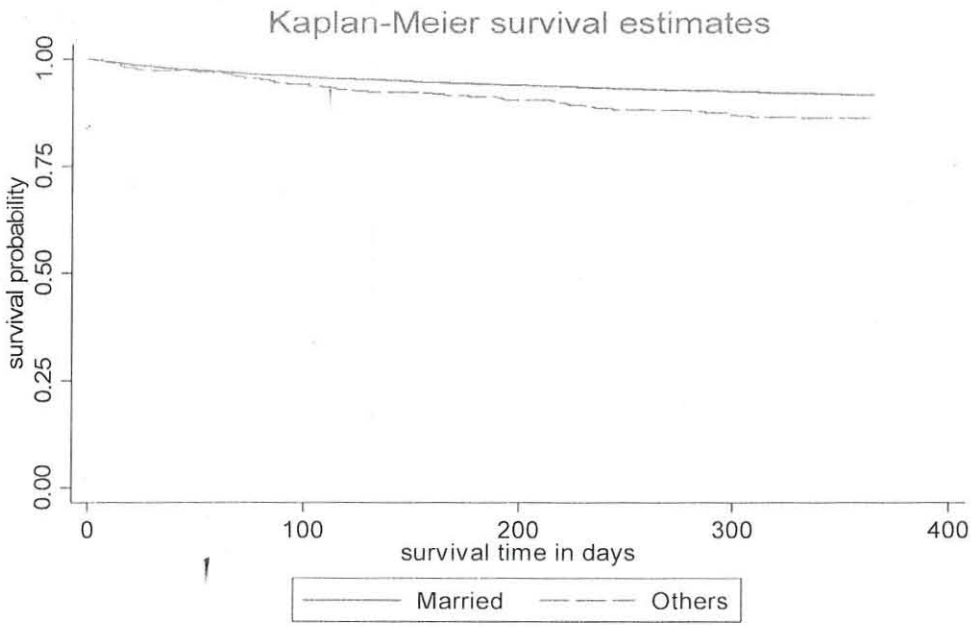

Figure 7. The Kaplan-Meier survival functions estimates by marital status

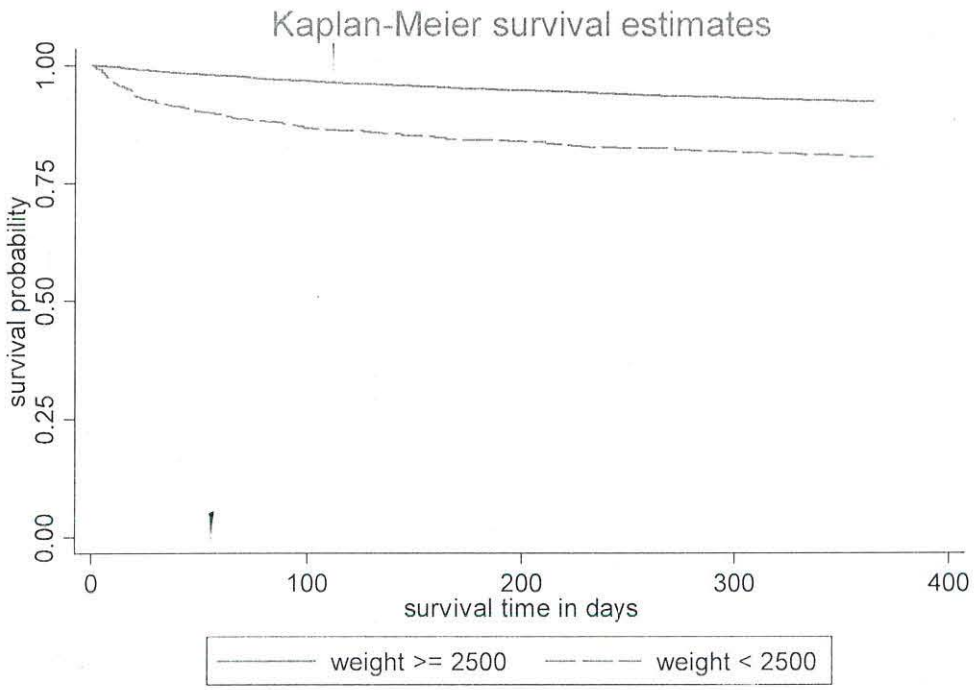

Figure 8. The Kaplan-Meier survival functions estimates by birth weight

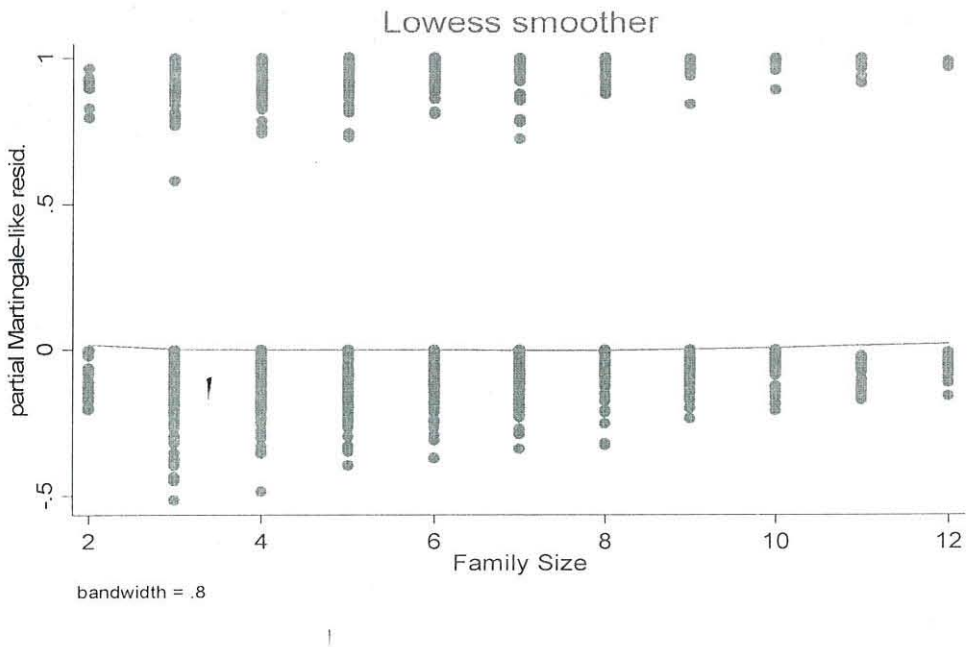

Figure 16. The plot of martingale residuals versus family size

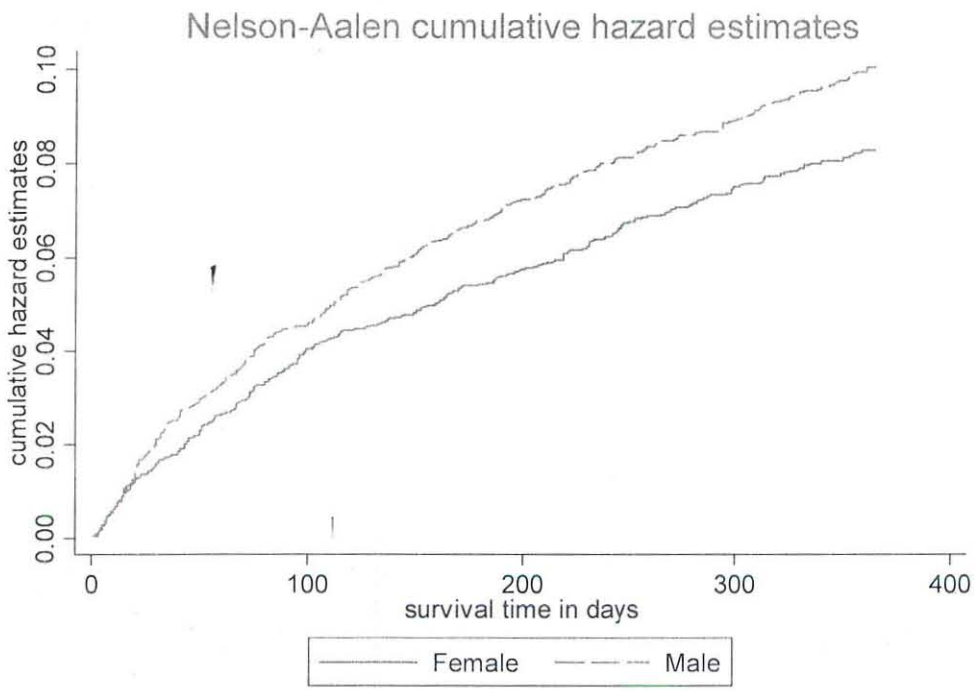

Figure 17. The Nelson-Aaelen cumulative hazard functions by sex

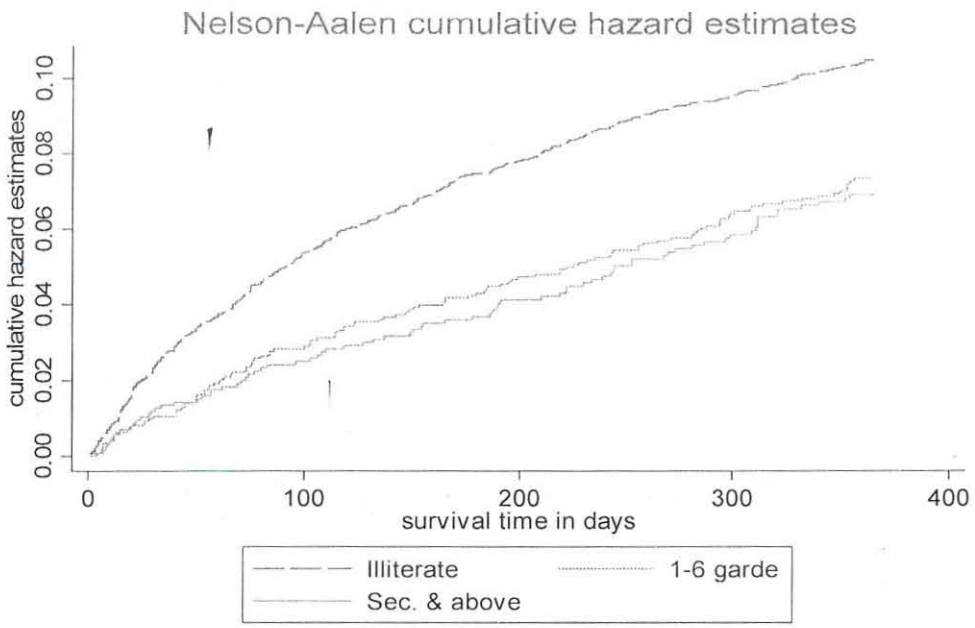

Figure 18. The Nelson-Aaelen cumulative hazard functions by mothers education

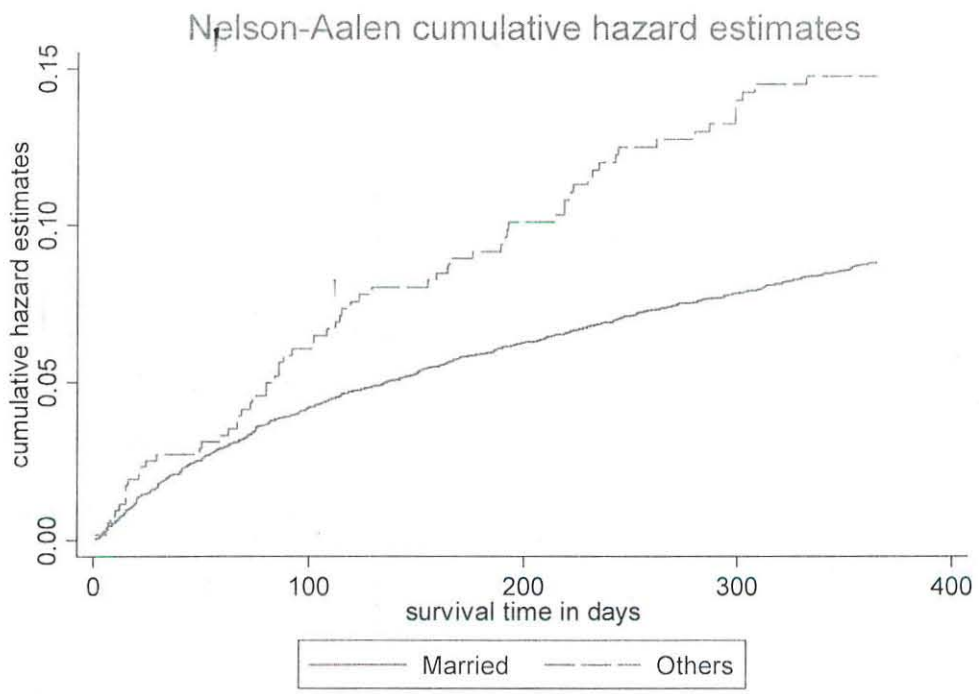

Figure 19. The Nelson-Aaelen cumulative hazard functions by marital status

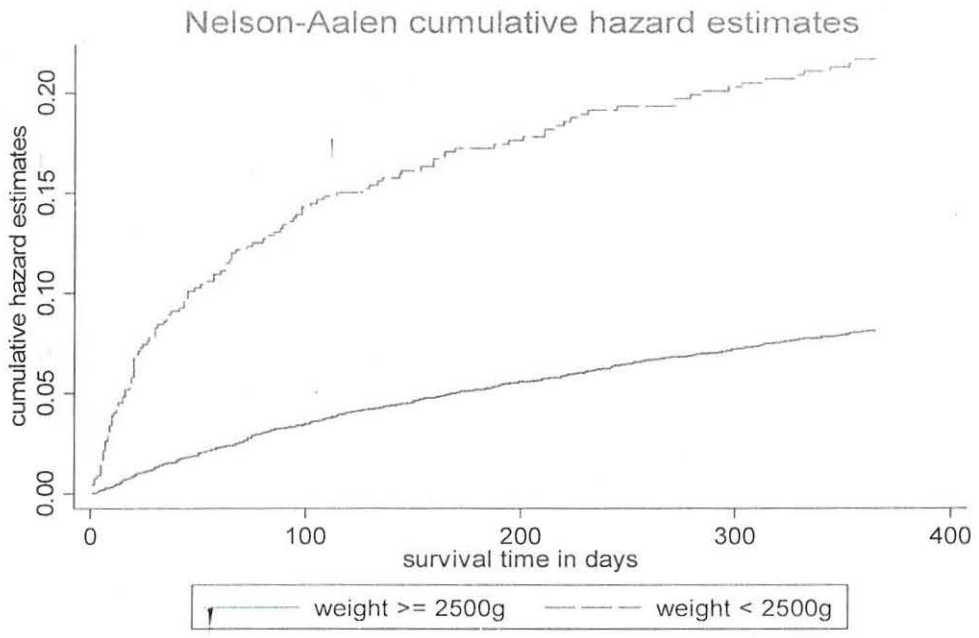

Figure 20. The Nelson-Aaelen cumulative hazard functions by birth weight

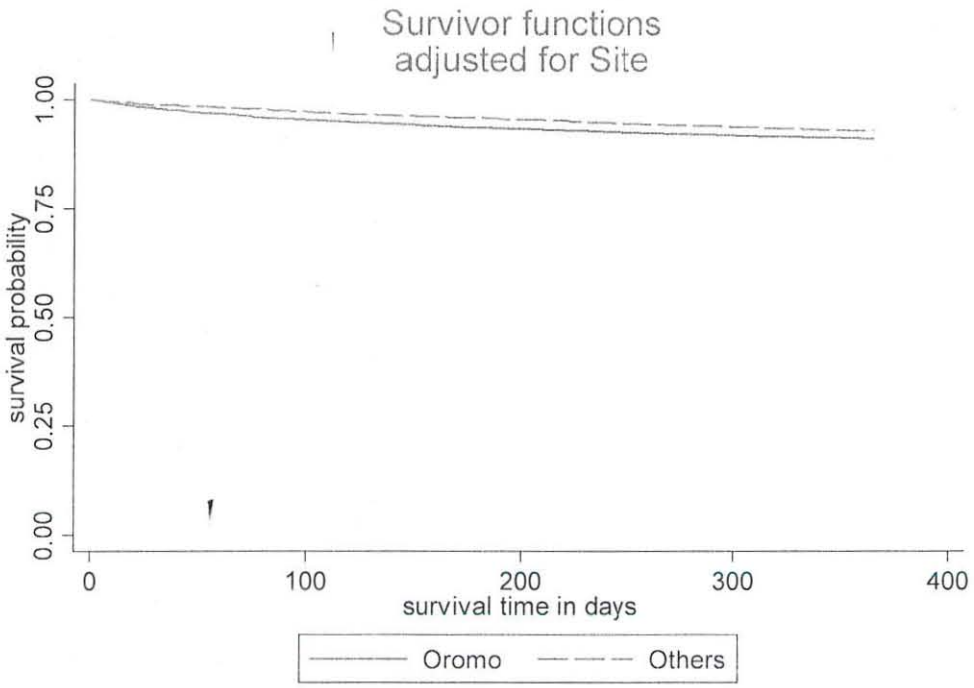

Figure 21. The Kaplan-Meier survival functions estimates by ethnic group adjusted for place of residence

APPENDIX B

Cox regression -- Breslow method for ties

No. of subjects =	7851	Number of obs =	7851
No. of failures =	661		
Time at risk =	2548291		
		LR chi2(22) =	189.83
Log likelihood =	-5778.2138	Prob > chi2 =	0.0000

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
rh						
famsize	.8648953	.0320523	-3.92	0.000	.8043011	.9300544
_Imomrel_2	1.044507	.1361273	0.33	0.738	.8090534	1.348484
_Iwater_2	.985871	.1227251	-0.11	0.909	.7724301	1.258291
_Iwater_3	.9713105	.1288618	-0.22	0.826	.7489122	1.259753
_Imeduc_1	.6000749	.0828763	-3.70	0.000	.4577687	.7866197
_Imeduc_2	.4702799	.1104002	-3.21	0.001	.2968475	.7450399
_Ibwt_1	4.254812	.6603433	9.33	0.000	3.138879	5.767481
_ISite_1	.9468231	.1683919	-0.31	0.759	.6681642	1.341697
_Isex_1	1.397366	.1762023	2.65	0.008	1.091382	1.789136
_Imometh2_2	.8076874	.1482045	-1.16	0.244	.5637081	1.157264
_Imstatus2_2	1.660575	.3528755	2.39	0.017	1.094906	2.518492
_Image3_1	.8885184	.1192856	-0.88	0.379	.6829528	1.155958
_Image3_2	.9540496	.1717018	-0.26	0.794	.6704693	1.357573
_Iparity_2	1.038791	.1183754	0.33	0.738	.8308632	1.298754
_Iparity_3	1.179544	.1817299	1.07	0.284	.8721103	1.595354
t						
famsize	1.000274	.0001882	1.46	0.145	.9999054	1.000643
meduc	1.001227	.0006194	1.98	0.047	1.000014	1.002442
Site	.999598	.0010087	-0.40	0.690	.9976229	1.001577
bwt	.9955104	.0011638	-3.85	0.000	.993232	.9977941
sex	.9993319	.000755	-0.88	0.376	.9978533	1.000813
mometh2	1.001295	.0009781	1.32	0.185	.9993795	1.003214
mstatus2	1.00039	.0012495	0.31	0.755	.9979439	1.002842

Note: second equation contains variables that continuously vary with respect to time; variables are interacted with current values of _t.

It can be seen that the p-value of the interaction between birth weight and log of time is 0.000 and the p-value of the interaction between mother's education and log of time is 0.047, theses shows that the two variables violates the proportional hazards assumption. Next we fitted stratified proportional hazards model using place of residence (site), birth weight or mothers education as a stratifying variable.

Stratified Cox regr. -- Breslow method for ties

No. of subjects =	7851	Number of obs =	7851
No. of failures =	661		
Time at risk =	2548291		
		LR chi2(20) =	169.76
Log likelihood =	-5339.7862	Prob > chi2 =	0.0000

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
rh						
famsize	.8650712	.0320489	-3.91	0.000	.8044829	.9302227
_Imomrel_2	1.044933	.1361815	0.34	0.736	.8093845	1.34903
_Iwater_2	.9857511	.1227087	-0.12	0.908	.7723383	1.258134
_Iwater_3	.9710085	.1288252	-0.22	0.825	.7486741	1.25937
_Imeduc_1	.6012479	.0828973	-3.69	0.000	.4588744	.7877951
_Imeduc_2	.4723745	.110613	-3.20	0.001	.2985148	.747493
_Ibwt_1	4.272216	.6653056	9.32	0.000	3.148451	5.797083
_Isex_1	1.396654	.1761205	2.65	0.008	1.090814	1.788244
_Imometh2_2	.8097014	.1482192	-1.15	0.249	.5655993	1.159153
_Imstatus2_2	1.657365	.3516416	2.38	0.017	1.093502	2.511984
_Image3_1	.8886409	.1192964	-0.88	0.379	.6830555	1.156103
_Image3_2	.9542415	.1717356	-0.26	0.795	.6706051	1.357843
_Iparity_2	1.038643	.118354	0.33	0.739	.8307514	1.298557
_Iparity_3	1.179491	.1817126	1.07	0.284	.8720842	1.595258
t						
famsize	1.000272	.0001881	1.45	0.148	.9999037	1.000641
meduc	1.001213	.0006169	1.97	0.049	1.000005	1.002423
bwt	.9954695	.0011718	-3.86	0.000	.9931755	.9977688
sex	.9993359	.000755	-0.88	0.379	.9978572	1.000817
mometh2	1.001279	.0009743	1.31	0.189	.9993712	1.00319
mstatus2	1.000406	.0012461	0.33	0.744	.9979671	1.002852

Stratified by Site

Note: second equation contains variables that continuously vary with respect to time; variables are interacted with current values of _t.

It can be seen that the p-value of the interaction between birth weight and log of time is 0.000 and the p-value of the interaction between mother's education and log of time is 0.049, theses shows that the two variables violates the proportional hazards assumption even after stratifying by the variable place of residence (site).

Stratified Cox regr. -- Breslow method for ties

No. of subjects =	7851	Number of obs =	7851
No. of failures =	661		
Time at risk =	2548291		
		LR chi2(20) =	79.56
Log likelihood =	-5470.2446	Prob > chi2 =	0.0000

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
rh						
famsize	.8646682	.0320789	-3.92	0.000	.8040263	.9298839
_Imeduc_1	.5982735	.0829043	-3.71	0.000	.4559808	.7849699
_Imeduc_2	.4669004	.110151	-3.23	0.001	.2940418	.7413775
_Iwater_2	.9858746	.1227285	-0.11	0.909	.7724284	1.258303
_Iwater_3	.9711024	.1288376	-0.22	0.825	.7487465	1.259491
_Iparity_2	1.03837	.1183485	0.33	0.741	.8304931	1.298279
_Iparity_3	1.17941	.1816931	1.07	0.284	.8720342	1.59513
_ISite_1	.9495526	.1696624	-0.29	0.772	.6690053	1.347747
_Isex_1	1.394527	.1757604	2.64	0.008	1.089294	1.78529
_Imomrel_2	1.044306	.1361268	0.33	0.739	.8088586	1.34829
_Imometh2_2	.8062134	.1484215	-1.17	0.242	.5620129	1.156522
_Imstatus2_2	1.661549	.3524189	2.39	0.017	1.096405	2.517997
_Image3_1	.8892524	.1194053	-0.87	0.382	.6834851	1.156967
_Image3_2	.9548215	.1718539	-0.26	0.797	.6709936	1.358708
t						
famsize	1.000277	.0001886	1.47	0.142	.9999076	1.000647
meduc	1.00125	.000624	2.00	0.045	1.000027	1.002473
sex	.9993408	.0007549	-0.87	0.383	.9978623	1.000822
mometh2	1.001307	.0009828	1.33	0.183	.9993824	1.003235
mstatus2	1.000392	.0012454	0.31	0.753	.9979535	1.002836
Site	.9995816	.0010157	-0.41	0.680	.9975928	1.001574

Stratified by bwt

Note: second equation contains variables that continuously vary with respect to time; variables are interacted with current values of _t.

It can be seen that the p-value of the interaction between mothers' education and log of time is 0.045, theses shows that the variable mothers' education (meduc) violates the proportional hazards assumption even after stratifying by the variable birth weight (bwt).

Stratified Cox regr. -- Breslow method for ties

No. of subjects =	7851	Number of obs =	7851
No. of failures =	661		
Time at risk =	2548291		
		LR chi2(19) =	153.09
Log likelihood =	-5246.528	Prob > chi2 =	0.0000

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
rh						
famsize	.8652075	.0320569	-3.91	0.000	.8046042	.9303755
_Ibwt_1	4.258858	.6619893	9.32	0.000	3.140392	5.77567
_Iwater_2	.9867813	.1228477	-0.11	0.915	.773129	1.259476
_Iwater_3	.9726391	.1290723	-0.21	0.834	.7498849	1.261563
_Iparity_2	1.038421	.1183303	0.33	0.741	.8305713	1.298284
_Iparity_3	1.178307	.1815377	1.06	0.287	.8711983	1.593677
_ISite_1	.9437981	.1682387	-0.32	0.746	.6654975	1.33848
_Isex_1	1.399465	.1764745	2.67	0.008	1.09301	1.791842
_Imomrel_2	1.045084	.1361866	0.34	0.735	.809524	1.349189
_Imometh2_2	.8079195	.1480242	-1.16	0.244	.5641751	1.15697
_Imstatus2_2	1.664499	.3532767	2.40	0.016	1.098051	2.523157
_Image3_1	.8879071	.119189	-0.89	0.376	.6825048	1.155126
_Image3_2	.9534891	.171582	-0.26	0.791	.6701014	1.356722
t						
famsize	1.000272	.0001882	1.45	0.148	.9999032	1.000641
bwt	.9954984	.0011676	-3.85	0.000	.9932125	.9977895
sex	.99932	.0007551	-0.90	0.368	.9978412	1.000801
mometh2	1.001288	.0009754	1.32	0.186	.9993784	1.003202
mstatus2	1.000368	.001247	0.29	0.768	.9979264	1.002815
Site	.9996294	.0010085	-0.37	0.713	.9976547	1.001608

Stratified by meduc

Note: second equation contains variables that continuously vary with respect to time; variables are interacted with current values of _t.

.

It can be seen that the p-value of the interaction between birth weight (bwt) and log of time is 0.000, theses shows that the variable birth weight (bwt) violates the proportional hazards assumption even after stratifying by the variable mothers' education.

Exponential regression -- accelerated failure-time form

No. of subjects	= 7851	Number of obs	= 7851
No. of failures	= 661		
Time at risk	= 2548291		
		LR chi2(9)	= 156.01
Log likelihood	= -3160.2915	Prob > chi2	= 0.0000

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
famsize	.0889526	.0198796	4.47	0.000	.0499893	.1279159
_Isex_1	-.2493336	.0786933	-3.17	0.002	-.4035697	-.0950974
_Imeduc_1	.3758504	.1089541	3.45	0.001	.1623044	.5893965
_Imeduc_2	.4343099	.1378435	3.15	0.002	.1641415	.7044783
_Imstatus2_2	-.5608076	.1316004	-4.26	0.000	-.8187396	-.3028756
_Ibwt_1	-.9777678	.1028387	-9.51	0.000	-1.179328	-.7762076
_Isited_1	.2335737	.1159585	2.01	0.044	.0062991	.4608482
_Imometh2_2	.20211	.1250388	1.62	0.106	-.0429615	.4471815
IsitXmom~2	-.408752	.2016752	-2.03	0.043	-.8040281	-.0134758
_cons	7.769582	.1605047	48.41	0.000	7.454999	8.084166

Weibull regression -- accelerated failure-time form

No. of subjects	= 7851	Number of obs	= 7851
No. of failures	= 661		
Time at risk	= 2548291		
		LR chi2(9)	= 151.99
Log likelihood	= -3105.2472	Prob > chi2	= 0.0000

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
famsize	.1257926	.0291444	4.32	0.000	.0686707	.1829146
_Isex_1	-.3553525	.1148737	-3.09	0.002	-.5805008	-.1302042
_Imeduc_1	.5376594	.1592824	3.38	0.001	.2254717	.8498472
_Imeduc_2	.6252674	.2013339	3.11	0.002	.2306603	1.019875
_Imstatus2_2	-.7884589	.1929173	-4.09	0.000	-1.16657	-.410348
_Ibwt_1	-1.399476	.1575679	-8.88	0.000	-1.708304	-1.090649
_Isited_1	.3221849	.1685171	1.91	0.056	-.0081026	.6524723
_Imometh2_2	.2943825	.1817653	1.62	0.105	-.061871	.6506359
IsitXmom~2	-.5867609	.2933948	-2.00	0.046	-1.161804	-.0117176
_cons	8.665589	.2573021	33.68	0.000	8.161286	9.169891
/ln_p	-.3722196	.0378699	-9.83	0.000	-.4464431	-.297996
p	.6892029	.0261			.6399001	.7423043
1/p	1.450952	.0549473			1.347156	1.562744

Log-normal regression -- accelerated failure-time form

No. of subjects	=	7851	Number of obs = 7851
No. of failures	=	661	
Time at risk	=	2548291	
			LR chi2(9) = 175.21
Log likelihood	=	-3079.4644	Prob > chi2 = 0.0000

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
famsize	.1389828	.0306552	4.53	0.000	.0788997	.1990658
_Isex_1	-.3916331	.1244346	-3.15	0.002	-.6355205	-.1477456
_Imeduc_1	.6162313	.1693934	3.64	0.000	.2842264	.9482362
_Imeduc_2	.6957362	.2122872	3.28	0.001	.279661	1.111811
_Imstatus2_2	-.8592568	.2264033	-3.80	0.000	-1.302999	-.4155144
_Ibwt_1	-1.776743	.1866331	-9.52	0.000	-2.142537	-1.410949
_Isited_1	.3349647	.1830194	1.83	0.067	-.0237468	.6936762
_Imometh2_2	.3538188	.1951287	1.81	0.070	-.0286265	.7362641
IsitXmom~2	-.688475	.3256356	-2.11	0.034	-1.326709	-.050241
_cons	9.239217	.2845301	32.47	0.000	8.681549	9.796886
/ln_sig	1.111703	.0336047	33.08	0.000	1.045839	1.177566
sigma	3.039529	.1021423			2.845784	3.246464

Log-logistic regression -- accelerated failure-time form

No. of subjects	=	7851	Number of obs = 7851
No. of failures	=	661	
Time at risk	=	2548291	
			LR chi2(9) = 158.55
Log likelihood	=	-3100.3733	Prob > chi2 = 0.0000

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
famsize	.1306197	.0296089	4.41	0.000	.0725873	.1886522
_Isex_1	-.3702133	.1174905	-3.15	0.002	-.6004904	-.1399362
_Imeduc_1	.5592916	.1618391	3.46	0.001	.2420928	.8764903
_Imeduc_2	.6418962	.2037789	3.15	0.002	.2424969	1.041296
_Imstatus2_2	-.8230752	.2021778	-4.07	0.000	-1.219336	-.4268139
_Ibwt_1	-1.508227	.1667853	-9.04	0.000	-1.835121	-1.181334
_Isited_1	.3269055	.172159	1.90	0.058	-.0105199	.6643309
_Imometh2_2	.3064026	.1846945	1.66	0.097	-.0555919	.6683971
IsitXmom~2	-.6043659	.3022583	-2.00	0.046	-1.196781	-.0119505
_cons	8.464221	.2591739	32.66	0.000	7.956249	8.972192
/ln_gam	.3404402	.0372865	9.13	0.000	.26736	.4135204
gamma	1.405566	.0524086			1.306511	1.512132

The Greenwood's formula for the variance of KM survival function estimator is given by:

$$\hat{Var}\left(\hat{S_{KM}}(t)\right)=\left[\hat{S_{KM}}(t)\right]^2\sum_{t_{(i)}\le t}\frac{d_i}{n_i(n_i-d_i)}$$

Point wise confidence interval (CI) for the survival function

$S(.)$ at $t=t_0$: $\hat{S}_{KM}(t_0)\pm z_{\alpha/2}s.e\left(\hat{S}_{KM}(t_0)\right)$

where $s.e\left(\hat{S}_{KM}(t_0)\right)=\sqrt{\hat{Var}\left(\hat{S}_{KM}(t_0)\right)}$; $0\le S(t)\le 1$

Sometimes, this CI gives values less than *0* and greater than *1* (which violates the rule of $0\le P(x)\le 1$). In order to achieve $0\le S(t)\le 1$, the log-log transformation will be applied.

$\ln\left(-\ln\left(\hat{S}_{KM}(t)\right)\right)$, $0\le S(t)\le 1$.

$\Rightarrow -\infty<\ln\left(\hat{S}_{KM}(t)\right)\le 0 \Rightarrow 0<-\ln\left(\hat{S}_{KM}(t)\right)<\infty$

$\Rightarrow -\infty<\ln\left(-\ln\left(\hat{S}_{KM}(t)\right)\right)<\infty$

The variance estimator of $\ln\left(-\ln\left(\hat{S}_{KM}(t)\right)\right)$ is given by:

$$\hat{Var}\left[\ln\left(-\ln\left(\hat{S}_{KM}(t)\right)\right)\right]=\left[\frac{1}{\ln\left(\hat{S}_{KM}(t)\right)}\right]^2\sum_{t_{(i)}\le t}\frac{d_i}{n_i(n_i-d_i)}$$

$\therefore$ point wise $(1-\alpha)100$ percent CI for $\ln\left(-\ln\left(\hat{S}_{KM}(t_0)\right)\right)$ is given by :

$$\ln\left(-\ln\left(\hat{S}_{KM}(t_0)\right)\right)\pm z_{\alpha/2}\ s.e\left(\ln\left(-\ln\left(\hat{S}_{KM}(t_0)\right)\right)\right)$$

Let $U=\ \ln\left(-\ln\left(\hat{S}_{KM}(t_0)\right)\right)+z_{\alpha/2}\ s.e\left(\ln\left(-\ln\left(\hat{S}_{KM}(t_0)\right)\right)\right)$

$$L = \ln\left(-\ln\left(\hat{S}_{KM}(t_0)\right)\right) - z_{\alpha/2}\ \text{s.e}\left(\ln\left(-\ln\left(\hat{S}_{KM}(t_0)\right)\right)\right)$$

$$\Rightarrow\ L \le \ln\left(-\ln\left(S_{KM}(t_0)\right)\right) \le U$$

$$\Rightarrow\ e^{L} \le -\ln\left(S_{KM}(t_0)\right) \le e^{U}$$

$$\Rightarrow\ e^{-e^{U}} \le S_{KM}(t_0) \le e^{e^{L}}$$

$\therefore$ point wise $(1-\alpha)100$ percent CI for KM survival estimate at $t = t_0$ is between $\exp(-\exp(U))$ and $\exp(-\exp(L))$

DECLARATION

I, the undersigned, declare that the thesis is my original work, has not been presented for degrees in any other University and all sources of material used for the thesis have been duly acknowledged.

Name: Henok Asefa

Signature: _____

Place: Faculty of Science, Addis Ababa University.

Date: 19/08/09

This thesis has been submitted for examination with my approval as a University advisor

Name: Fentaw Abegaz, PH.D

Signature: _____

Date: 19/08/09