SURVIVAL ANALYSIS AND FRAILTY MODELLING OF TIME-TO-EVENT DATA: AN APPLICATION TO INFANT MORTALITY IN MALAWI

MASTER OF SCIENCE(BIOSTATISTICS) THESIS

ESTHER KHUNDI

UNIVERSITY OF MALAWI

APRIL 2024

SURVIVAL ANALYSIS AND FRAILTY MODELLING OF TIME-TO-EVENT DATA: AN APPLICATION TO INFANT MORTALITY IN MALAWI

MSc. (Biostatistics) Thesis

BY

ESTHER KHUNDI

BSc. (Statistics)- University of Malawi

Thesis submitted to the Mathematical Sciences Department, Faculty of Science, in partial fulfilment for the Degree of Master of Science (Biostatistics)

University of Malawi

April 2024

DECLARATION

I, the undersigned, hereby declare that this thesis is my original work and it has never been submitted for similar purposes to this or any other university or institution of higher learning. Where other people's work has been used acknowledgements have been made. All errors contained herein are the author's sole responsibility.

ESTHER KHUNDI

____

Full Legal Name

____

Signature

30th April 2024

____

Date

CERTIFICATE OF APPROVAL

The undersigned certifies that this thesis represents the student’s work and effort, and it makes acknowledgments where other sources of information are used. The thesis is submitted with our approval.

Signature_____________________________Date:________________________________

Mavuto Mukaka, Ph.D (Professor)

Supervisor

DEDICATION

I dedicate this thesis to my daughter Ivana Candice Banda.

ACKNOWLEDGEMENTS

I thank God for giving me such a great opportunity and for giving me life to come this far.

A special thanks to my supervisor Professor Mavuto Mukaka for his constructive suggestions and comments on my research work. His selfless support during my period of study is highly appreciated.

A special thanks to my parents.

ABSTRACT

Infant mortality rate is one of the important health and development indicators in a country or community and that is why reduction of infant mortality has been the main target of public health policies for the past decades. Malawi, like many countries in the sub Saharan Africa is a country that suffers from the highest rates of infant mortality across the globe. Studies have been conducted to identify factors associated with infant mortality in Malawi but none of these studies used recent data. This study used the most recent survey data to identify the factors associated with infant mortality in Malawi by using survival analysis techniques and frailty modelling to control for unobserved heterogeneity using. The data used for this study was from the 2015-16 Malawi Demographic Healthy Survey (2015-16 MDHS) and was obtained from DHS program website: https://www.dhsprogram.com. Bivariate analysis was conducted to identify variables that had significant association with infant mortality in Malawi using both Kaplan-Meier and log rank test and were subsequently considered into the cox proportional hazard model analysis to estimate their strength of effect on infant mortality in Malawi. The variables were also modelled using both the semi parametric cox frailty model and parametric frailty models to find the best fit model using the maximum likelihood estimation. The results showed that sex of household head, mothers' age group, source of drinking water, religion, type of birth and place of delivery were significantly associated with infant mortality and that there are unobservable family effects which make infant deaths to cluster in some households/families.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CMC	Century Month Code
DHS	Demographic and Health Survey
IMR	Infant Mortality Rate
UNAIDS	United Nations Programme on HIV/ Acquired immune deficiency syndrome
UNICEF	United Nations Children’s Fund
UNIGME	United Nations Inter-agency Group on child Mortality Estimation
NSO	National Statistical Office
NMR	Neonatal Mortality Rate
MDG	Millennium Development Goals
SHH	Sex of Household Head
U5MR	Under Five Mortality Rate

CHAPTER 1

INTRODUCTION

This chapter introduces the study by first discussing the background and context, followed by the statement problem and objectives of the study.

1.1 Background

Infant mortality refers to the death of a child before reaching one year of age, and it is a global burden especially in developing countries like Malawi (Ndawala, 2015). Infant mortality includes prenatal mortality, neo natal mortality and postnatal mortality which are defined as death that occur in the first week after birth, death that occur within 28 days of birth and death that occur between 28 and 364 days after birth respectively (Ndawala, 2015) . Infant mortality rate (IMR) is sensitive to general structural factors like socio-economic development and basic living conditions, as such it is regarded as an important national health indicator (Sartorius, Kurt, Sartorius, 2014).When a country has high infant mortality rate, it is an indication of unmet human health needs such as sanitation, medical care, nutrition and education (Treibe, 2009). The estimation of mortality in childhood traditionally focused on mortality below one year of age because mortality at early ages is highest among infants and also because measures of mortality for the age range 0 to 1 year can be obtained solely from registration data when those data are reliable (Nations, 1989).

There is still a big gap in infant mortality between developing and developed countries. Research shows that 1 in 36 children dies during the first month of life (neonates) in Sub-Saharan Africa, compared with 1 in 333 in developed countries (Ouatarra, 2018). The United Nation's Sustainable Development Goal 3 (SDGs) seeks to put an end to avoidable new-born deaths before 2030, which contributes to infant mortality. Over 60 countries will fail to meet the United Nation's Sustainable Development Goal 3 if considerable progress is not made (Ouatarra, 2018).

Research shows that there was a 2.5 percent annual decline in global child mortality between 1960 and 1990, with Sub-Saharan Africa (SSA) having the slowest decline, (Jahn et al, 2010) of about 1.0 percent in the 1960s, 2.0 percent between 1970 and 1985, and 1.0 percent between 1985 and 1990. Resulting to an annual average decline of about 1.3 percent between 1960 and 1990, (Hill, Kenneth, Amouzou, 2006).Children born in Sub-Saharan Africa today have a life expectancy of 51 years and almost 10.0 percent of them die in the first year of life. Approximately 4.1 million deaths occurred globally within the first year of life in 2017, accounting for 75 percent (of all under-five deaths (Tesfa et al., 2021). The contribution of infant deaths to overall child mortality has increased over the years and has reached 75.0 percent. It is important to target children under the age of 1 year (infants), and call for urgent and concerted action to further improve the survival chances of world's children (WHO, 2016).

Progress in the reduction of infant and child mortality accelerated in the period 2000-2017 as compared to the 1990s period with an annual rate reduction in the global under five mortality rates having increased from 1.9 percent in 1990-2000 to 4.0 percent in 2000-2017.An estimated 5.4 million children under 5 years died in 2017, (Roser, Max, Ritchie, Hannah, Dadonaite, 2013) and half of all these deaths, about 2.7 million, took place in sub Saharan Africa, (Hug, Lucia, Sharrow, David, Zhong, 2018). Even though infant mortality significantly declined worldwide, the decline in SSA was unsatisfactory which was, 92/1000 live births in 2000 to 53/1000 live births in 2018 (Tesfa et al., 2021).

1.2 Problem Statement

Malawi, like many countries in the Sub-Saharan Africa suffered from the highest rate of infant mortality. Malawi's infant mortality rate was at 37.828 deaths per 1000 live births in 2020 (Plecher, 2020). In 2020, the mortality rate among children under the age of 1 year in Africa was around 41.6 deaths per thousand live births. Although infant mortality rate of

Malawi fell gradually from 164.75 deaths per thousand live births in 1971 to 36.08 deaths per thousand births in 2021, this infant mortality rate is still high.

Infants are particularly vulnerable to their immediate living conditions and suffer the highest consequences of negative health outcomes from socio-economic issues and social disadvantages. As such, it is important that infant mortality be a focal point in societies to ensure that infant mortality levels are kept low. Studies have been conducted on infant mortality in Malawi but infant mortality rate is still high. To mention a few, (Madise & Diamond, 1995) used a logistic binomial model to analyze 1988 child spacing survey data to identify determinants of infant mortality and another study conducted by (Kalipeni & Moise, 2015) that used 1990-2010 Malawi demographic and health survey (MDHS) data to assess the reduction of infant mortality. The aim of this study was to apply survival analysis techniques to identify factors associated with infant mortality in Malawi using the recent demographic and health survey dataset.

1.3 Objectives

i. To examine the association between infant mortality and determinants in Malawi
ii. To examine the effects of unobserved heterogeneity (frailty) on infant mortality both at family and community level
iii. To find the best fit model for infant mortality data

CHAPTER 2

LITERATURE REVIEW

2.1 Chapter Overview

This chapter provides an overview of previous research on infant mortality which includes the methods used to conduct the studies, their findings and recommendations made. The chapter further provides an overview on survival analysis, frailty modeling and their different estimation methods.

As discussed in chapter 1, many studies have been conducted on infant mortality and this is because infant mortality rate is considered to be one of the key health indicators in an economy (Nasejje, 2015).This chapter summarizes the literature on non-statistical issues of infant mortality and statistical methods used when analyzing infant mortality data.

There are many classical modelling methods used in the examining of factors associated with infant mortality which include Bayesian analysis, logistic regression, Cox proportional hazard model (CPH) and simple correlation method, just to mention a few. However, Logistic regression analysis and CHP are the most commonly used method and that is why the two have been discussed further in this chapter. Logistic regression models the probabilities for classification problems with two possible outcomes i.e. this method requires a binary response variable. It is an extension of the linear regression model for classification problems, some of the basic assumptions that must be met for logistic regression include independence of errors, linearity in the logit for continuous variables, absence of multicollinearity, and lack of strongly influential outliers. Logistic regression however, does not determine the causal relationship between an independent variable and the outcome variable, but rather will allow for the describing of the variables associated to infant mortality (Dube, 2012).The logistic regression model was used by (Lemani, 2013) to study the survival of infant children in Malawi where it was claimed that the logistic method has low statistical power on censored children compared to the Cox proportional hazard model, as such it is problematic to use it when the time to exposure is short and when the risk of experiencing an event of interest vary with time.

Another method which is commonly used is the standard Cox proportional hazards model which is applicable when the interest is in time-to-event data and the data is assumed to be independent. Studies which use DHS data which is obtained from a cluster survey and assumed to be correlated violates the statistical assumption of interdependence when the standard Cox proportional hazard model is used and it does not adjust for unobserved confounders. In order to adjust for unobserved covariates there is need to use the Cox-frailty model method to examine factors associated with infant mortality since this model assumes that the risk of death of an individual is a function of measured factors and a random term on the baseline hazard due to the unobserved cluster effect (Khan & Awan, 2017) . It is important to use frailty modelling because of its capabilities in accounting for unobserved or unobservable risk factor effects in survival data analysis, (Niragire *et al*., 2011).

High infant mortality rates may increase fertility rates because families want to replace the lost children and high fertility rates pose a health risk to women and children. Research showed that in every year, an estimated 529 000 women die in pregnancy or childbirth (WHO, 2004) and the timing of births has an impact on child health in such a way that when a woman doesn’t have adequate child space, the new baby is often born underweight or premature, develops too slowly, and has an increased risk of dying before reaching the age of 1 year (Rustein, 2005). It was discovered that children born less than 2 years after the previous birth are about 2.5 times as likely to die before age 5 than children who are born 3–5 years apart (Setty-Venugopal, Upadhyay, 2002).

Malawi is a country in the sub-Saharan region and is characterized by high infant and child mortality which was estimated to be 104 deaths per 1000 live births and 95 deaths per 1000 live births respectively using data from the 2000 Malawi Demographic and Health Survey (Kalipeni & Moise, 2015).

Afeez *et al*. (2018) conducted a retrospective study to find out the risk factors responsible for infant mortality in Nigeria. Kaplan Meier curve was plotted to describe the rate of survival of some of the factors responsible for infant mortality. The log-rank test was used to test the null hypothesis that there is no difference between populations in the probability of an infant dying. The Cox-Proportional Hazard Model was fitted to assess the importance of various covariates in the survival times of infant through the hazard ratio. It was found that religion, sex of child, area of residence, economic status of the family and age of mother at birth are factors associated with infant mortality.

A national time series of NMR (Neonatal Mortality Rate) and neonatal deaths, was estimated using the UN IGME (United Nations Inter-agency Group for Child Mortality Estimation) multilevel statistical model with random effects parameters for level regression parameters at country level. UN IGME used an abridged life table approach and to calculate the absolute number of deaths among infants and children under-five in a given year and country (Walker et al., 2012). It was reviewed how relevant data from civil registration, sample registration, population censuses, and household surveys are compiled and assessed for United Nations member states. It was also reviewed how time series regression models are fitted to all points of acceptable quality to establish the trends in U5MR (under-five mortality rate) from which infant and neonatal mortality rates are generally derived. The application of this methodology indicated that, between 1990 and 2010, the global U5MR fell from 88 to 57 deaths per 1,000 live births, and the annual number of under-five deaths fell from 12.0 to 7.6 million(Walker *et al*., 2012). Although the annual rate of reduction in the U5MR accelerated from 1.9 percent for the period 1990-2000 to 2.5 percent for the period 2000-2010, it remains well below the 4.4 percent annual rate of reduction required to achieve the MDG 4 goal of a two-thirds reduction in U5MR from its 1990 value by 2015 (Walker *et al*., 2012).

A study on infant mortality and causes of infant deaths in rural Ethiopia conducted by Weldearegawi *et al* (Weldearegawi *et al*., 2015), used multiple Cox proportional Hazards regression model to investigate risk factors for infant death and causes of infant death were identified using physician review verbal autopsy method. It was found that mother's level of education, mother's age, pre-mature birth, respiratory infections and sepsis were the common causes of infant death.

A study conducted by (Nutiye, 2009) aimed to examine factors that are correlated with infant mortality in Turkey. The study used survival analysis and logistic regression to analyze data from the 2003-2004 Turkey demographic and health survey. The results of the study showed that birth interval is associated with infant mortality, breastfeeding is important for the survival chance of the infants under the age 3 months. Place of delivery and source of water the family uses were also found to be correlated with infant mortality. There was also a curvilinear relation between maternal age at birth and infant mortality which indicated high risk for infants born from teenage mothers and old age mothers (Nutiye, 2009).

(Bolstad & Manda, 2001) conducted a study where they investigated child mortality in Malawi using family and community effects and applied a Bayesian analysis method. The study found that early succeeding conception and short breastfeeding duration are the factors that have the highest in-creased risk for a child. (Bolstad & Manda, 2001)also found that there was more variability due to family effects than community effects in child mortality. They further learned from the random effects model that infant and early child deaths tend to cluster in some families and, to a lesser extent, in some communities. The family variation summarizes the effects of biological, genetic, parental competence whilst community variation summarizes the effects of differing community cultures and customs which were not accounted for in the fixed effect model.

2.2 Logistic Regression

Logistic regression is one of the most commonly used model for applied statistics and discrete data analysis. The logistic regression is a statistical analysis method used to explain the relationship between a dependent variable and one or more independent variables. This regression method is one of the generalized linear models with a logit link and works very similar to linear regression, with the exception that the response (dependent) variable is binomial taking the value of 1 if the event of interest occurred and 0 otherwise (Sperandei, 2016). The Logistic regression model is one of the most popularly used model in child mortality studies because it assumes that child survival is a binary response, child is dead or alive (Kazembe et al., 2012). Log-odds play an important role in logistic regression as it converts the logistic regression model from probability based to a likelihood based model.

The Logistic regression model equates the logit transform, the log-odds of the probability of a success, to the linear component as shown below;

$$\log\left(\frac{p_i}{1-p_i}\right) = \sum_{k=0}^{k} x_{ik}\beta_k \qquad i = 1,2,\dots,n \qquad 2.1$$

Where k is the number of independent variables specified in the model, p is the probability, x is an independent variable, β is the regression coefficient, i is the subject and n is the sample size. The probability p in terms of the explanatory variable x is given by;

$$p_i = \frac{\exp(\beta_0+\beta_i x_i)}{1+\exp(\beta_0+\beta_i x_i)} \qquad 2.2$$

Where:

p_i is the probability that the event of interest will occur

β_0 is the intercept

β_i is the regression coefficient for the explanatory variable x_i

2.2.1 Maximum Likelihood function for Logistic Regression

Maximum likelihood estimation is a probabilistic framework for estimating the parameters of a model. The logistic regression goal is to estimate the k+1 unknown parameters in Eq. 2.2. The maximum likelihood equation is derived from the probability of the dependent variable. Each y_i represents a binomial count in the i^{th} population as such the joint probability density function of Y is given by;

$$f(y|\beta) = \prod_{i=1}^{n} \frac{n_i!}{y_i!(n_i - y_i)!} p_i^{y_i} (1 - p_i)^{n_i - y_i} \qquad 2.3$$

And the likelihood function is given as;

$$L(\beta|y) = \prod_{i}^{n} \frac{n_i!}{y_i!(n_i - y_i)!} p_i^{y_i} (1 - p_i)^{n_i - y_i} \qquad 2.4$$

The values for β that maximize the likelihood function in Eq. 2.4 are called the maximum likelihood estimates and these estimates are found by computing the first and second derivative of the likelihood function since critical points occur when the first derivative is a zero and if the second derivative is less than zero, then the critical point is a maximum.

Note that the factorial terms do not contain any of the p_i as a result, they are essentially constants that can be ignored. Also,$a^{x-y} = \frac{a^x}{a^y}$, therefore the likelihood function can be rearranged and written as;

$$\prod_{i}^{n} \left(\frac{p_i}{1-p_i}\right)^{y_i} (1 - p_i)^{n_i}$$

Exponentiating both sides of Eq. 2.1 it becomes;

$$\frac{p_i}{1 - p_i} = e^{\sum_{k=0}^{k} x_{ik}\beta_k}$$

And solving for p_i result it;

$$p_i = \frac{e^{\sum_{k=0}^{k} x_{ik}\beta_k}}{1 + e^{\sum_{k=0}^{k} x_{ik}\beta_k}}$$

After substituting for$p_i = \frac{e^{\sum_{k=0}^{k} x_{ik}\beta_k}}{1+e^{\sum_{k=0}^{k} x_{ik}\beta_k}}$in equation 2.4 and simplifying its yield the log likelihood function;

$$l(\beta) = \sum_{i=1}^{N} y_i \left(\sum_{k=0}^{k} x_{ik}\beta_k \right) - n_i . \log(1 + e^{\sum_{k=0}^{k} x_{ik}\beta_k})$$

Differentiating the log likelihood function with respect to each β_k it becomes;

$$\frac{\partial l(\beta)}{\partial \beta_k} = \sum_{i=1}^{N} y_i x_{ik} - n_i p_i x_{ik} \qquad 2.5$$

Critical points can be found by setting each of the k+1 equations in 2.5 equal to zero and solving for each β_k. The critical point will be a maximum if the second partial derivatives is a negative definite. The second partial derivative is;

$$\frac{\partial^2 l(\beta)}{\partial \beta_k \partial \beta_{k'}} = -\sum_{i=1}^{N} n_i x_{ik} p_i (1 - p_i) x_{ik'}$$

2.2.2 Interpreting Parameters

Recall the logistic model;

$$\log\left(\frac{p_i}{1 - p_i}\right) = \sum_{k=0}^{k} x_{ik}\beta_k \qquad i = 1,2, \dots, N$$

Where $\frac{p_i}{1-p_i}$ is the odds of an event occurring. The regression coefficient in the population model is the log odds ratio, log(OR) , which is the difference between two log odds and can be used to compare the odds between two groups. The OR is obtained by exponentiating β;

$$e^{\beta} = e^{\log OR} = OR$$

We interpret OR > 1 as indicating a risk factor and OR < 1 indicating a protective factor.

2.3 Survival Analysis

Survival analysis is one of the highly active areas of research and is applied in many fields of study which include engineering, physical, biological and social sciences (Nasejje, 2015). Survival analysis is a statistical method or tool which is used to analyze time to events data and the time variable is usually referred to as survival time, because it gives the time that an individual has “survived" over some follow up period. An outcome that is of scientific interest is called an event and this event is observed in different studies like sociology, biology, demography, medicine and employment (Nasejje, 2015). An event is typically referred to as a failure. This is so because the event of interest is usually death, disease incidence, or some other negative individual experience. However, survival time may be something other than a failure such as “time to return to work after an elective surgical procedure," "time of courtship to wedding," in which case failure is a positive event. In survival analysis the interest lies in the time for an event of interest to occur from a given baseline and in this paper we are interested in the time it takes for an infant to die from birth. This technique allows one to depict the pattern of experiencing a survival event over time. There are four fundamental functions in survival analysis which include the cumulative probability function F(t), the survival function S(t), the hazard function h(t) and the cumulative hazard function H(t) (Nyinawajambo, 2018).

2.3.1 The survival Function

Assuming T is a continuous random variable with probability density function (p.d.f.) f(t) and cumulative survival/failure function, F(t) = Pr(T < t), the survival function S(t) is given by;

$$S(t) = \Pr(T \geq t) = 1 - F(t) = \int_t^\infty f(x)dx \qquad 2.6$$

The expression above is the probability of surviving beyond time t. The survival function is usually a downward sloping curve with time at the x-axis and survival probability at y axis (Lemani, 2013), as indicated in the figure 1 below. The survival function S(t) is a non-increasing function, S(0) =1 and S(∞) = 0.

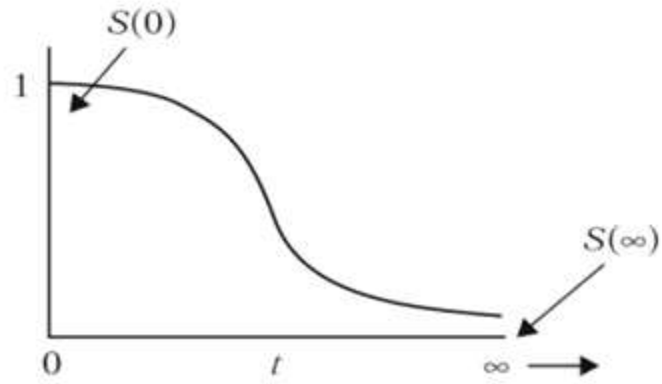

Figure 1: Survival Function

2.3.2 The Hazard Function

Given a set containing individuals who are at a risk of experiencing a certain event denoted by R(t) (risk set) where t represents the time, then the probability of an individual in the risk set experiencing the event in the small time interval[t; t + Δt] is defined as h(t) Δt and the hazard rate is given as;

$$h(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t | T \geq t) \qquad 2.7$$

The hazard function characterizes the risk of dying that is changing over time and it takes on any shape of a non-negative function and it varies depending on the type of survival data given unlike the survival function which is a downward sloping curve for any type of given survival data (Nyinawajambo, 2018). Some hazard functions, e.g. the exponential survival function has a constant hazard rate, meaning it does not change with time.

Non-parametric, semi-parametric and parametric methods exist for survival methods and these methods are discussed in the next section.

2.3.3 Non-Parametric Methods

Non-parametric methods are methods that do not make assumptions about a population's parameters, they are sometimes called "distribution free methods". Non-parametric methods are used to summarize survival data through estimates of the hazard and survival function. The aim of non-parametric estimation of the survival function is to come up with graphical summaries of the survival times for a given group of individuals considered in a study. Non-parametric methods for survival analysis include Life-table, log-rank and Kaplan Meier.

(i) Kaplan Meier Estimator

The Kaplan Meier estimator was originally derived as a non-parametric maximum likelihood estimator of a function and as a limit of the actuarial estimator as the time axis is partitioned into fine intervals. It is a non-parametric statistic, which is also known as the product limit estimator which is used to estimate the probability of dying (the hazard probability), the probability of surviving and median survival time. The Kaplan Meier estimator uses the exact failure time to give a simple and quick estimate of the survival function in presence of censoring (Lemani, 2013). It is denoted as;

$$S(\hat{t}) = \prod_{t_i<t} \left(1 - \frac{m_i}{r_i - 1}\right) \qquad 2.7$$

Where m_i represents the number of deaths at time i, and $r_i - 1$ represent the number of subjects at the start of the study. Kaplan Meier technique takes into account both censored and uncensored observations and assumes that censored times are independent to survival times while estimating survival probabilities. This technique divides the follow-up period into a number of small intervals and number of cases for each interval is determined and a probability of surviving to the end of that time period is obtained when the surviving proportion is multiplied by the surviving proportions for each of the preceding time periods (Damato et al., 2011).The survival probability is then plotted against time. Thus, survival function could be estimated using the Kaplan Meier graph. Under this graphical technique, bivariate analysis of infant survival was depicted within and across districts using the DHS data (Lemani, 2013).

(ii) Life Tables

The life table procedure is a conventional approach used since the 18^{th} century to analyze the distribution of mortality in a population. It takes into account information from censored cases whose full observation period will not have elapsed at the time of interview and whose survival outcome cannot therefore be recorded. The life table will allow depicting survival ratios and failure rate at every time interval (WHO et al., 2013). This method is an alternative method of Kaplan-Meier method with particularity of being able to assess the survivorship function of groups of individuals even though there is no survival information at individual level.

$$Let\ \ I_t = number\ alive\ at\ the\ beginning\ of\ time\ t,$$
$$d_t = number\ of\ deaths\ during\ the\ time\ interval$$

Then the probability of dying during the time interval is given by;

$$q_t = \frac{d_t}{I_t}$$

And the probability of dying during the time interval;

$$p_t = 1 - q_t$$

(iii) Log Rank Test

The log rank test is a large-sample chi-square test that provides an overall comparison of the Kaplan Meier curves being compared by using a statistic as a test criterion. Just like many other statistics used in other kinds of chi-square tests, this log rank statistic makes use of observed versus expected cell counts over categories of outcomes and the categories for this log-rank statistic are defined by each of the ordered failure times for the entire data being analyzed. The null hypothesis being tested is that there is no overall difference between two survival curves. The log-rank statistic is approximately chi-square with one degree of freedom under this null hypothesis. The log rank test however, does not provide an estimate of the size of the difference between groups or a confidence interval because it is purely a test of significance.

2.3.4 Semi-Parametric Methods

(i) Cox Proportional Hazard Regression Model

The Cox-proportional hazard model is essentially a commonly used survival regression model in medical research for investigating the association between the survival time of patients and one or more predictor variables. It is called a semi-parametric method because the distribution for the baseline hazard function is not specified. The Cox model was introduced in 1972 and has the form;

$$\lambda(t|X) = \lambda_0(t)\exp(X^T\beta) \qquad 2.8$$

Where $\lambda(t|X)$ is the hazard at time t, for an individual with covariate X, $\lambda_0(t)$ denotes the baseline hazard function and assumed to be unique for all individuals in the study population, X is the vector of observed covariates and βthe respective vector of regression parameters to be estimated(Cox, 1972). There are several important assumptions for appropriate use of the Cox proportional hazard regression model which include, independence of survival times between distinct individuals in the sample, a constant hazard ratio over time and a multiplicative relationship between the predictors and the hazard i.e. proportional hazard. The Cox model cannot be used in a situation where the assumption of proportional hazard is violated because it assumes hazard proportionality. The hazard ratio is the measure of the effect of the given covariates on survival time. For Example, given a categorical variable with two levels say X = 1 and X = 0 , where group 1 have chemo before surgery and group 0 have chemo after surgery to compare the hazard of death from cancer, the hazard ratio for the two groups is given as;

$$HR = \frac{h(t|X=1)}{h(t|X=0)} = \exp(\beta) \qquad 2.9$$

When HR = 1, it implies that the individuals in the two categories are at the same hazard risk of dying, when HR > 1, it implies that the individuals in the first category (X = 1) are at a higher hazard of dying and if HR < 1, the individuals in the second category (X = 0) are at a higher hazard of dying.

2.4 Cox-Fraitly Models

Proportional hazards (PH), and in particular the semi-parametric Cox model play a major role in the modelling of continuous event times (Wienke, 2003).The Cox model assumes the semi-parametric hazard;

$$\lambda(t|x_i) = \lambda_0(t)\exp(x_i^T\beta) \qquad 3.0$$

Frailty models aim at modelling the heterogeneity in the population, they can be used to account for the influence of unobserved covariates (Vaupel & Manton, 1979). The Parameter θ provides information on the variability (dependency) of the population in the same family or community. However, parameter estimation in frailty models is more challenging than in the Cox model since the corresponding profile likelihood has no closed form solution. In the Cox PH frailty model also known as the mixed PH model, the hazard rate of subject j belonging to cluster i with n_i subjects, conditionally on the covariates x_{ij} and the shared frailty b_i is given by;

$$\lambda_{ij}(t|x_{ij}, b_i) = b_i\lambda_0(t)\exp\left(x_{ij}^T\beta\right) \qquad i = 1, \dots, n,\ j = 1, \dots, n_i$$

Where bi is the frailty term and frequently assumed to follow a gamma distribution because of its mathematical convenience. There are two categories of frailty models which are the univariate frailty models that consider univariate survival times and the multivariate frailty models that take into account multivariate survival times (Wienke, 2003).The frailty, bi , is an unobservable random variable varying over the sample which increases the individual risk if b > 1 or decreases if b < 1. When b > 1 (frailty is greater than one) an individual is said to be at an increased hazard of failure therefore more frail than an average individual in a cluster whilst when b < 1 the individual has a lower risk or is less frail therefore tends to survive longer (Nasejje, 2015).

2.4.1 Univariate Frailty Models

Univariate frailty models take into account the non-homogeneity of a population. Unobserved heterogeneity comes about when important covariates have not been observed even though heterogeneity maybe explained by covariates. (Vaupel & Manton, 1979) introduced univariate frailty models (with a gamma distribution) into survival analysis to account for unobserved heterogeneity or missing covariates in the study population. This idea assumed that different patients possess different frailties and that the patients who are more "frail" or "prone" tend to have the event earlier that those who are less frail. Assuming that an individual child under the age of i where i = 1, 2,…, n has a survival time denoted as t_i and the covariate vector Xi has a frailty term denoted as b_i, then the survival function of individual i conditional on the frailty is given by;

$$S_i(t_i, X_i|b_i) = exp\left(-b_i e^{x_i^T \beta} \int_0^{t_i} h_0\,(s, X_i|b_i) ds\right) = exp(-b_i H_0(t_i) exp(x_i^T \beta) \qquad 3.1$$

Where $H_0(t_i) = \sum_0^{t_i} h_0\,(s)ds$ is the cumulative baseline hazard function (Nasejje, 2015). Also assuming that the frailty follows a gamma distribution (α= β), with mean E(B)=1, variance V (B) = $\frac{1}{\alpha}$ and the variance of b, the frailty term is denoted as θ, then V(B) =$\frac{1}{\alpha}$ = θ. The probability of a one parameter Gamma distribution f(b):

$$f(b) = \frac{\alpha^{\alpha} b^{\alpha-1} \exp(-\alpha b)}{\Gamma(\alpha)}$$

Substituting $\frac{1}{\alpha} = \theta$ we obtain;

$$f(b) = \frac{b^{\frac{1}{\theta}-1} e^{\frac{-b}{\theta}} \exp(-\alpha b)}{\Gamma(\frac{1}{\theta})\theta^{\frac{1}{\theta}}} \qquad 3.2$$

Now, letting T denote the random variable representing the survival times and b denote the frailty with the Gamma distribution, then the conditional survival function is given by:

$$S_i(t|b) = \exp(-bH_0(t))$$

The z is then integrated out from the conditional survival function which gives the unconditional survival function below;

$$S_i(t) = E[S(t|b)] = \int_0^\infty e^{-bH_0 \exp\left(X_i^t\beta\right)(t)f(b)db = L(H(t))} \qquad 3.3$$

Where L denotes the Laplace transform. The likelihood is then calculated as;

$$L(t, X_i, \beta, \theta) = \prod_{i=0}^{G} \prod_{j=0}^{n_i} b_i^{\delta_i} h_0\left(t_{ij}\right)^{\delta_i} exp\left(\delta_i X_{ij}^T \beta\right), [S\left(t_{ij}\right)]^{1-\delta_i}, \qquad 3.4$$

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

2.4.2 Multivariate Frailty Models

Multivariate frailty models have been used frequently for modelling dependence in multivariate time-to-event data. The aim of the frailty is to take into account the presence of the correlation between the multivariate survival times. Multivariate models with dependent random hazards provide a multivariate extension of the traditional univariate frailty model. Application of frailty models in the field of multivariate survival data is important because such kind of data occurs for example if lifetimes (or times of onset of a disease) of relatives (twins, parent-child) or recurrent events like infections in the same individual are considered. The independence between the clustered survival times cannot be assumed in such cases which where Multivariate models come in since they are able to account for the presence of dependence between these event times (Wienke, 2003). The dependence structure in the multivariate model arises from a latent variable in the conditional models for multiple observed survival times.

2.4.2.1 Shared Frailty Model

The shared frailty model is relevant to event times of related individuals, similar organs and repeated measurements. This model is called a shared frailty model because individuals in a cluster are assumed to share the same frailty (Wienke, 2003).Frailty is assumed to be independent across the groups or clusters while the survival times of individuals within the same group are conditionally dependent (Nasejje, 2015). A shared frailty model in survival analysis is defined as follows;

Let b_i denote the shared frailty that are assumed to be identically and independently distributed random variables, T_{ij} denote the survival time of the j_{th} individual in the i_{th} cluster given n clusters with n_i individuals and vector X_{ij} associated with the survival time(Wienke, 2003), then the hazard function of the j_{th} individual of the i_{th} cluster is given as:

$$h_{ij}(t) = b_i h_0(t) \exp(X_{ij}^T \beta) \quad 3.5$$

2.4.3 Parameter Estimation

To obtain the derived estimates of the parameters, the likelihood function is differentiated with respect to the parameters in the model and the resulting equations are then solved simultaneously. Due to the presence of latent variables it is not usually possible to solve the equations simultaneously with frailty models (Nasejje, 2015) and this requires us to use a more advanced method. Some of these advanced methods include;

. The Expectation-Maximisation Algorithm (EM-Algorithm);
. The Markov Chain Monte Carlo (MCMC) methods;
. The Monte Carlo EM (MCEM) approach;
. The penalised partial likelihood (PPL).

The EM algorithm and the Penalised Partial likelihood methods are mainly used when the survival data is right censored. With complicated forms of censoring like interval and left censoring, more advanced Frailty modelling methods have to be used to estimate the parameters of the model. Markov Chain Monte Carlo methods (MCMC) are the alternative methods that can be used to estimate parameters of parametric frailty models in circumstances where there exists left and interval censored data points in the data set (Nasejje, 2015).

(i) The Expectation-Maximisation Algorithm (EM-Algorithm)

Given the full likelihood $L_{full}(t_i, h_0, \beta, \theta)$ and assuming that the frailties follow a gamma distribution then the full likelihood of a shared frailty model of a cohort consisting of N individuals, where each individual is assigned to a cluster and with a total number G clusters where each cluster consists of n_i number of individuals is given by:

$$L_{full}(t_i, h_0, \beta, \theta) = \prod_{i=0}^{G} \prod_{j=1}^{n_i} b_i^{\delta_i} h_0(t_{ij})^{\delta_i} exp(\delta_i X_{ij}^T \beta) \exp(\delta_i H_0(t_{ij}))$$

$$= \prod_{i=1}^{G} \frac{b_i^{\frac{1}{\theta}+D_i-1} exp\left(\frac{-b_i}{\theta}\right)}{\Gamma(\frac{1}{\theta})\theta^{\frac{1}{\theta}}}$$

Where Di is the total number of events in the cluster i. The EM-algorithm method requires the initial estimates for β, $H_0(t_{ij})$ and $\theta(\ \widehat{\beta}, \widehat{H}_0(t_{ij})$ and $\widehat{\theta}$ respectively) be found. The model with no frailties is used to get the initial estimates $\hat{\beta} and\ \widehat{H}_0(t_{ij})$ and then use the obtained estimates together with β = 0 to get the expected values of the frailty terms ($b_i s$) (Nasejje, 2015).For the Gamma frailty model, (Hanagal, 2011) argues that the distribution of the frailty terms bi is a Gamma with the shape and scale parameters $\widehat{\alpha} = \frac{1}{\theta + D_i}$ and $\hat{\theta} = \frac{1}{\theta} + \sum_j H_0\left(t_{ij}\right) exp X_{ij}^T \beta$. Therefore the expected value of the frailties is given by;

$$E(b_i) = \frac{\hat{\alpha}_i}{\hat{\theta}_i}$$

$$E(\ln(b_i)) = \psi(\hat{\alpha}_i) - \ln(\hat{\theta}_i)$$

Where $\psi(.)$is a Di-gamma function given by;

$$\psi = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$$

To obtain the estimates of β and h_0 at the M-step, the expected values of the frailty terms are plugged into the modified partial likelihood (Nasejje, 2015) and the modified partial likelihood is given by:

$$L(\beta) = \prod_{k=1}^{M} \frac{\exp(\hat{b}\beta s_k)}{\left(\sum_{l\epsilon R(t_k)} blexp(X_l^T \beta)\right)^{d_k}} \qquad 3.6$$

Where t_k is the smallest failure time, d_k is the number of failures at time t_k, D_k is the set of all individuals who fail at time t_k and $S_k = \sum_{j\epsilon D_k} x_j$. The maximum likelihood estimate for the baseline hazard function is obtained from the expression below:

$$\widehat{\lambda_{0k}} = \frac{d_k}{\sum_{l\epsilon R(t_k)} blexp(X_l^T \beta)} \qquad 3.7$$

To find the estimates of θ, the estimates of $\hat{\beta},\ \hat{b}_i\ and\ \hat{\lambda}_0$ are plugged in.

(ii) The Markov Chain Monte Carlo Methods (MCMC)

Another method in parameter estimation for frailty methods is the MCMC method. These methods are statistical simulation techniques, Instead of writing down complex system of equations, a process is directly simulated given the probability density functions that describe it. Given the probability density functions (p.d.f's), its initial defining parameters and current values of the other parameters, the simulation process begins by iteratively re-sampling each parameter (Nasejje, 2015). They are known as Markov Chain Monte Carlo (MCMC) methods because one uses the previous sample values to generate randomly the next sample values which results into a Markov chain. An MCMC method consists of generating a set of Markov chains whose joint stationary distribution corresponds to the joint posterior of the model (Wienke, 2003). The posterior distribution is often very difficult to work with in a hierarchical model and almost always impossible to integrate out in order to find the marginal posterior of each random parameter but the MCMC methods enable us to overcome this problem (Wienke, 2003) .

The Gibbs sampling is one of the algorithms that have been created in order to obtain Markov chains with the desired stationary distribution. Gibbs sampling is used to fit frailty models on clustered failure time data with right censored observations, by sampling iteratively from the full conditional distribution of the parameters in the model (Nasejje, 2015). The basic idea behind the Gibbs sampling is to successively sample from the conditional distribution of each random node, given all the others in the model (Wienke, 2003).

2.5 Survival Analysis Parametric Methods

A parametric survival model is a well-recognized statistical technique for exploring the relationship between the survival of a person, a parametric distribution and several explanatory variables. It allows us to estimate the parameters of the distribution. Parametric survival analysis models typically require a non-negative distribution and the distributions that work well for survival data include the exponential, Weibull, gamma and lognormal distributions. There are two parametric methods which are commonly used in survival analysis which include the parametric proportional hazards model and the Accelerated failure time (AFT) model.

(a) Parametric Proportional hazards model

Parametric proportional hazard models are used to describe proportional hazards models in which the hazard function is specified i.e. when proportional hazard models are formulated with assuming a probability distribution for survival times, this leads to parametric models (Khosa, 2019). It was advocated by (Gong & Fang, 2013)to use parametric proportional hazard models for the analysis of interval censored data. Let consider the analysis of survival data when one is to assume a parametric form of distribution of survival time. Let T denote a continuous non-negative random variable representing survival time with p.d.f. (probability density function) f(t)and c.d.f. (cumulative density function) F(t) =$\Pr(T \leq t)$.

Let $\Lambda(t) = \int_0^t \lambda(u)du$ denote the cumulative hazard. Recall that $S(t) = \exp\{-\Lambda(t)\}$. Any distribution defined for $t\epsilon[0, \infty)$ can serve as a survival distribution. Some of the distributions that are commonly used include Exponential, Weibull, Geompertz-Makeham and Gamma.

All parametric models can be fit by maximizing the appropriate likelihood function. Let data consist of pairs (t_i, d_i) where;

t_i is the survival or censoring time $and\ d_i$ is a death indicator

The likelihood function under general non-informative censoring has the form;

$$L(\theta) = \prod_{i=1}^{n} \lambda(t_i|x_i)^{d_i} S(t_i|x_i) \qquad 3.8$$

And in general must be maximized numerically.

(b) Accelerated failure time (AFT) model

The Acceleration failure time model is a parametric model which was introduced by (Cox, 1972) and it is known as accelerated failure time model because of the term "failure" which indicates the event of interest which could be death, disease etc. The term "Accelerated" indicates the responsible factor for which the rate of failure is increased and that factor is referred to as the "Acceleration factor" (Saikia & Barman, 2017).

If the appropriate parametric form of AFT model is used then it offers a potential statistical approach in case of survival data which is based upon the survival curve rather than the hazard function. In AFT model, the dependent variable is log of the survival time T, and the assumption is that the effect of covariates act multiplicatively (proportionally) with respect to the survival time.

A semi-parametric model is a statistical model that has parametric and non-parametric components. Survival semi-parametric methods are called semi-parametric because while the hazard function is estimated non-parametrically, the functional form of the covariates is parametric and this is a strength because the non-parametric estimate of the hazard function offers much greater flexibility than most parametric approaches. Semi-parametric models have few assumptions which makes them a popular choice, however, parametric model provide greater efficiency in such a way that only a few parameters are estimated and this model is comparatively easy to interpret. It also provides the ability to extrapolate beyond the range of the data. Parametric models do have challenges which include choosing a reasonable distribution to run the models.

The standard Cox proportional hazards model is applicable when time-to-event data are independent and our Study is going to use DHS data which is obtained from a cluster survey and assumed to be correlated as such the statistical assumption of interdependence is violated if the standard Cox proportional hazard model is used as such we are going to use the Cox-frailty model which accounts for both observed and unobserved effects. It is also important to consider the possibility that some children are frailer than others i.e. some children are more likely to experience the hazard than others as such there is need to use the Cox-frailty model that captures total effects of all factors that influence the child's risk of death that are not included in the standard Cox-proportional hazard model.

CHAPTER 3

RESEARCH METHODOLOGY

In this chapter the first section describes the data sources and the second section describes the methods that will be used for data analysis.

3.1 Data Sources

The data for this study is from the 2015-16 Malawi Demographic Healthy Survey (2015-16 MDHS) which was implemented by the National Statistical Office from 19 October 2015 to 17 February 2016. The funding for the 2015-16 MDHS was provided by the government of Malawi, the United States Agency for International Development (USAID), the United Nations Children's Fund (UNICEF), the Malawi National AIDS Commission (NAC), the United Nations Population Fund (UNFPA), UN WOMEN, Irish Aid, and the World Bank (National Statistical Office (NSO), 2017). The DHS program is a USAID project to assist developing countries worldwide in collecting and monitoring data to evaluate the population, health and nutrition programs. All official raw data and reports from all countries where DHS is application can be accessed through https://www.dhsprogram.com. The primary objective of the 2015-16 MDHS was to provide estimates of basic demographic and health indicators.

The survey was based on a nationally representative sample which provided estimates at the national and regional levels and for urban and rural areas with key indicator estimates at the district level. The survey included 26,361 households, 24,562 female respondents, and 7,478 male respondents. The 2015-16 MDHS included household and respondent characteristics, fertility and family planning, infant and child health and mortality, maternal health and maternal and adult mortality, child and adult nutrition, malaria, HIV/AIDS, domestic violence, orphans, and vulnerable children. The sampling frame used for the 2015-16 MDHS is the frame of the Malawi Population and Housing Census (MPHC), conducted in Malawi in 2008, and provided by the Malawi National Statistical Office (NSO). The 2015-16 MDHS sample was stratified and selected in two stages. Each district was stratified into urban and rural areas; this yielded 56 sampling strata. Samples of standard enumeration areas (SEAs) were selected independently in each stratum in two stages.

Infant and child mortality data was collected as part of a retrospective birth history in which female respondents listed the children they have born, child's date of birth, survivorship status, current age or age at death which they used to indirectly and directly estimate infant mortality rate.

Table 1: Selected Variables

Variables (DHS Codes)	**Label**
B7	Age at Death (Months, Imputed)
B5	Child is alive
B3	Date of birth (CMC**)
V008	Date of interview(CMC)
V106	Mother Highest education level
V190	Socio-Economic Status of the Family
B4	Sex of child
B0	Type of birth
V113	Source of drinking water
V151	Sex of household head
M15	Place of delivery
V102	Area of residence
M18	Size of child at birth
V130	Religion
V013	Mothers’ age group

3.1.2 Measurements of study variables

- Outcome variable

In this study the dependent variable was infant mortality was defined as the death of a child under the age of 1 year and this variable was measured as a binary response: yes or no. All births that occurred within 5 years before the date of interview were included. There was no specific variable for the survival time in the dataset as such the children survival times, in years, who were alive were calculated by subtracting the date of birth (CMC- century month code) from the date of interview (CMC) and then dividing by 12 to get survival time in months as shown below;

$$Survival\ time\ for\ alive\ children = \frac{date\ of\ Interview(V008) - date\ of\ birth(B3)}{12}$$

The survival time for children who are dead was simply the age at death divided by 12;

$$Survival\ time\ for\ dead\ children = \frac{Age\ at\ death(B7)}{12}$$

- Explanatory variables

The explanatory variables were grouped into four categories which are social demographic, social economic, biological and environmental factors.

The social-demographic factors include Mothers' age group, Religion and Sex of household head (SHH). Religion had the following categories; Catholic, Anglican, Muslim, CCAP, other Christian and no religion. Sex of household head had female and male categories. The mother's age group, in years, at birth of the child was calculated as the difference between the child's birth date and mothers' birthdate which was then categorized into 15-19, 20-24, 25-29, 30-34, 35-39, 40-44 and 45-49 years.

The social economic factors include mother's highest education and economic status of the family. Mother's highest education was categorized into no education, primary, secondary, and higher (tertiary). Economic status of the family had the following categories; poorest, poorer, average, richer and richest. The DHS program calculates the composite score wealth index which determines the family economic status by combining ownership of several household assets (televisions or bicycles), construction materials for the household in which participants live, as well as their accessibility to water and sanitation services (Gondwe *et al.*, 2021).

The biological factors include sex of child, type of birth and size of child at birth. Type of birth was categorized into single and multiple whilst the size of child at birth covariate was categorized into very small, smaller than average, average and larger than average.

The environmental factors include source of drinking water, area of residence and place of delivery. Area of residence was categorized into rural and urban whilst place of delivery had 4 categories which were respondent's home, Other home, Government hospital, Government health center, Government health post / outreach, Other public sector, Private hospital / clinic, Cham/mission hospital, Cham/mission health center and BLM(Banja la Mtsogolo). Globally, nearly a billion people still lack access to improved sources of drinking water and unimproved water and sanitation are major causes of diarrhea which globally accounts for approximately 1.4 million child deaths each year (Ezeh *et al.*, 2014). As such it is important to look at this environmental factor which is source of drinking water. For this covariate respondents were asked their main source of drinking water for members of their household was asked and the responses were categorized into twelve categories which were Piped into dwelling, Piped to yard/plot, Piped to neighbor, Public tap/standpipe, Tube well or borehole, Protected well, Unprotected well, Protected spring, Unprotected spring, River/dam/lake/ponds/stream/canal, Rainwater and Not a dejure resident.

3.2. Methods of Analysis

Data analysis was conducted in Stata 14 at univariate and multivariable levels. Kaplan-Meier and log-rank tests were used for the univariate analysis and since the Cox-frailty model is a modification of the Cox proportional hazard model, both models were fitted for the multivariable analysis. The significance level was taken as a p value <0.05.

Let t represent the time to death (survival time) of a child under 1year of age (age of the child) in the data. Assuming that the survival times or are identically and independently distributed, the Cox Proportional Hazard model with covariates was fitted as follows;

$$\begin{aligned}\lambda(t|Z) = \lambda_0(t)\exp(\beta_1 &* mothers' age\ group + \beta_2 * religion + \beta_3 * SHH + \beta_4 \\ &* size\ at\ birth + \beta_5 * Area\ of\ residence + \beta_6 * sex\ of\ child + \beta_7 \\ &* type\ of\ birth + \beta_8 * mothers' highest\ education + \beta_9 \\ &* family\ economic\ status + \beta_{10} * place\ of\ delivery + \beta_{11} \\ &* source\ of\ drinking\ water\end{aligned}$$

For the Cox frailty model which is an extension of the Cox proportional hazard model, family and community effects on infant mortality will be considered. The Frailty model has an unobserved multiplicative effect on the hazard rate for all individuals in the same group. This is why this model was used in this study since infants in the same family or community share the same nuisance (frailty) factor. Some infant deaths occur more in certain families than others and this variation could be due to effects of biological, genetic, parental competence, and other family specific factors that have not been accounted for in the Cox proportional hazard model (Bolstad & Manda, 2001). And also, some communities experience more infant deaths than others which could be due to different cultural practices and customs as such it is important to study these random effects in infant mortality.

Let t_{ik}be the time child k, in family or community i, leaves the study, either by death or by surviving to the end of the study. Let X_{ik} denote the design vector for child k, in family or community i, for the fixed effect explanatory variables and let β be the vector of fixed effect coefficient. Then the hazard function for child k, in family/community i, will be given by;

$h_{ik}(t|x_{ik}, b_{ik})$ =$b_{ik}h_0(t)\exp(x_{ik}^T\beta)$

Whereb_i is the random effects which will be interpreted as relative risk since it operates multiplicatively on the hazard function (Bolstad & Manda, 2001).

3.2.1 Model Diagnostics

Since the Cox proportional hazard model is semi-parametric and that it does not have an implied error, model checking is commonly implied checking whether the proportional hazards assumption is met. Testing the time dependent covariates is equivalent to testing for a non-zero slope in a generalized linear regression of the scaled Schoenfeld residuals on functions of time. A non-zero slope is an indication of a violation of the proportional hazard assumption.

CHAPTER 4

RESULTS

In this chapter, the results of the study have been presented and discussed with reference to the aim of the study.

This study aimed to identify factors associated with infant mortality in Malawi and to examine the effects of unobserved heterogeneity (frailty) on infant mortality both at family and community level. This chapter provides a description of the results that were found after analyzing the 2015-16 MDHS data.

4.1 Description of Study Population

This section gives a brief description of the study population based on the factors that were studied. A total number of 4232 infants was considered for this study and out of these, 721 (17.04%) died before their first birthday. The mean age in months for infants who died is 2.25 months which indicate that a lot of infants die in their early days afterbirth as presented in table 2. As can be seen from tables 3,4 and 5, out of the 721 dead infants, 405 (56.17%) were males and 316 (48.83%) were females, which clearly indicates that more male infants died than females. On the family economic status infants born from poorest families had the highest death percentage of 24.41 and 85.25 percent infant that died were from the rural area. The table also showed that the population had more households that are poorer and poorest. The table 4 also showed that a lot of households used boreholes as a source of drinking water, 2535 households out of the 4232 use boreholes as a source of drinking water. A high number of mothers' education level, 2798, was primary.

Table 2: Descriptive Statistics for age in months

status	N	Sd(months)	IQR(months)	Median(months)
alive	3511	3.64	3 to 9	6.00
dead	721	3.79	0 to 3	0.00
total	4232	3.96	2 to 9	5.00

Table 3: Distribution of deaths by survival determinants

COVARIATE	Number of deaths	Percentage (%)
Sex of child		
male	405	56.17
female	316	48.83
Family economic		
status	176	24.41
Poorest	174	24.13
Poorer	127	17.61
Middle	124	17.20
Richer	120	16.64
richest		
Place of delivery		
Respondents home	47	6.52
Other home	19	2.64
Govt hospital	246	34.12
Govt health center	306	42.44
Govt outreach	7	0.97
Other pub sector	0	0
Private hospital	20	2.77
Cham/mission hospital	35	4.85
Cham/mission health	32	4.44
center	9	1.25
Other		
Religion		
Catholic	702	17.48
CCAP	549	10.12
Anglican	187	2.77
Seventh day Adventist	280	8.32
Other Christian	1919	48.27
Muslim	573	12.76
Other	4	0.28

Table 4: Cont of distribution of deaths by survival determinants

Sex of household head		
Male	2354	74.48
female	978	25.52
Source of drinking water		
Piped into dwelling	95	2.91
Piped to yard	264	5.69
Piped to neighbor	148	2.64
Public tap	413	8.04
Tube well/borehore	2535	62.69
Protected well	125	2.91
Unprotected well	333	8.46
Unprotected spring	12	1.66
River/dam	200	4.02
Rainwater	3	0.28
other	0	0.00
Not a dejure resident	38	0.42
Mothers' highest education		
No education(ref)	495	12.21
Primary	2798	69.21
Secondary	858	17.48
higher	81	1.11
Mothers' age group		
15-19	657	11.37
20-24	1364	30.64
25-29	877	19.97
30-34	679	16.64
35-39	436	12.34
40-44	164	5.83
45-49	55	3.19

Table 5: Cont of distribution of deaths by survival determinants

Area of residence		
Urban	690	14.15
Rural	3542	85.85
Size of child		
Very large	329	7.77
Larger than average	1030	20.80
Average	2108	42.72
Smaller than average	523	16.78
Very small	184	7.63
Don't know	58	4.30
Type of birth		
Single birth	609	84.47
1st of multiple	54	7.49
2nd of multiple	58	8.04

4.2 Life Table Results

The life table summarized the mortality trend among infants and as can be seen from the table 6, there was a high number of deaths in infants between the ages of 0 and 1 month old, there were 458 deaths. Which is in agreement with the research that a lot of children die in their first month of life (neonatal period). Generally, as the months increase the number of deaths decrease which implies that age of child plays a role in infant mortality.

Table 6: summary of survival probabilities for infants

Interval /Months	Beginning total no. of children	deaths	lost	survival	Confidence interval
0 1	4232	458	128	0.8901	(0.8802, 0.8992)
1 2	3646	50	274	0.8774	(0.8671, 0.8871)
2 3	3322	21	299	0.8716	(0.8610, 0.8815)
3 4	3002	17	277	0.8664	(0.8556, 0.8765)
4 5	2708	16	273	0.8611	(0.8500, 0.8714)
5 6	2419	14	315	0.8557	(0.8443, 0.8664)
6 7	2090	23	275	0.8456	(0.8336, 0.8569)
7 8	1792	13	270	0.8390	(0.8265, 0.8507)
8 9	1509	21	256	0.8263	(0.8127, 0.8389)
9 10	1232	27	279	0.8058	(0.7905, 0.8202)
10 11	926	9	258	0.7967	(0.7804,0.8120)
11 12	659	12	297	0.7780	(0.7587, 0.7960)
12 13	350	40	310	0.6184	(0.5700, 0.6630)

4.3 Kaplan Meier Results

In this section the research used a non-parametric method, the Kaplan Meier, which is mainly graphical, to describe how the risk of death for the children under 1 year is distributed across the strata of some of the chosen covariates.

Only a few Kaplan Meier curves are presented in the thesis and the others are presented in appendix B. As can be seen in figures 2, 3 and 4, the curve shows that male infants had a high probability of death compared to female infants. Infants born in poorer families had a high probability of death compared to their counterparts and infants born in families headed by males had a higher probability of survival than infants born in families headed by females.

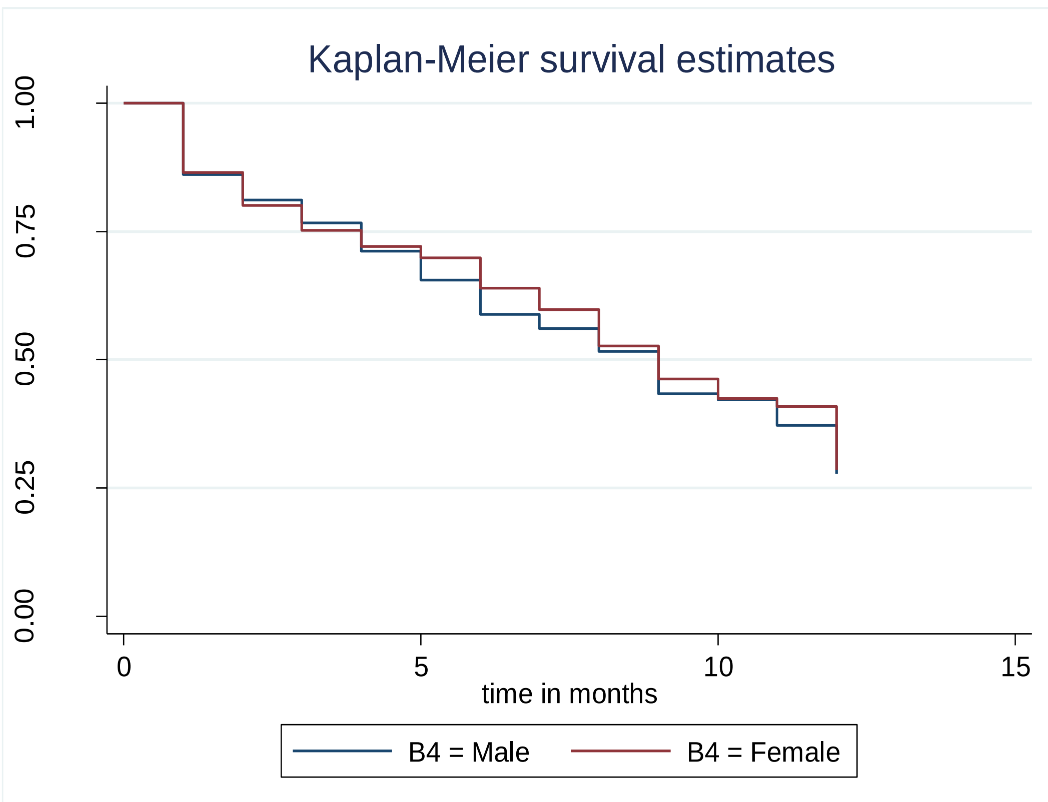

Figure 2: Survival estimates by sex of child

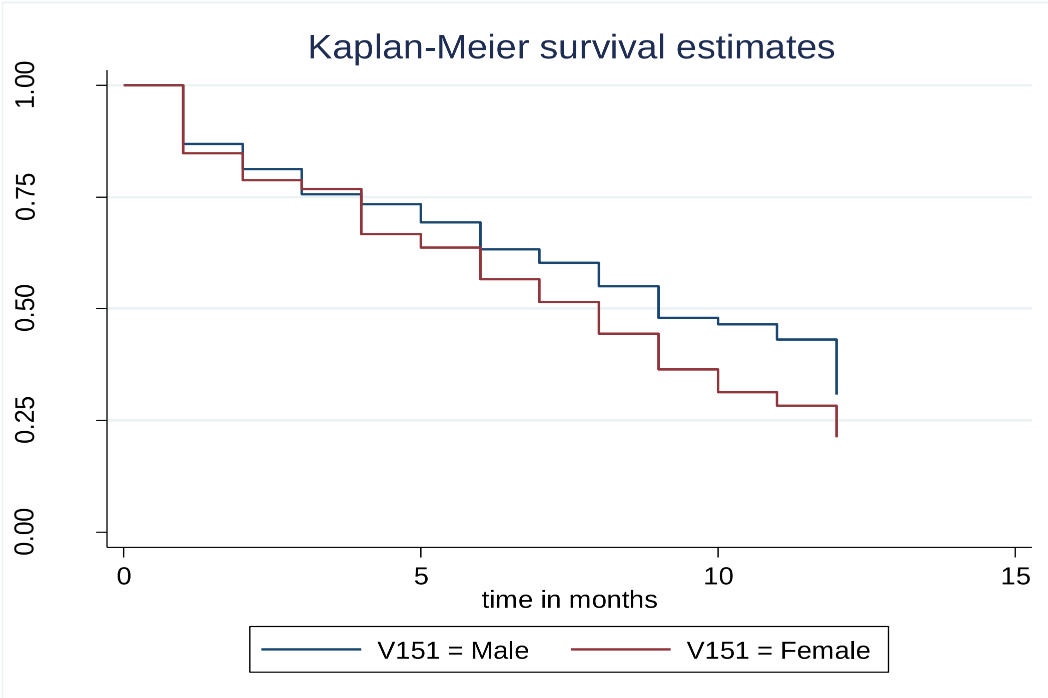

Figure 3: Survival estimates by sex of household head

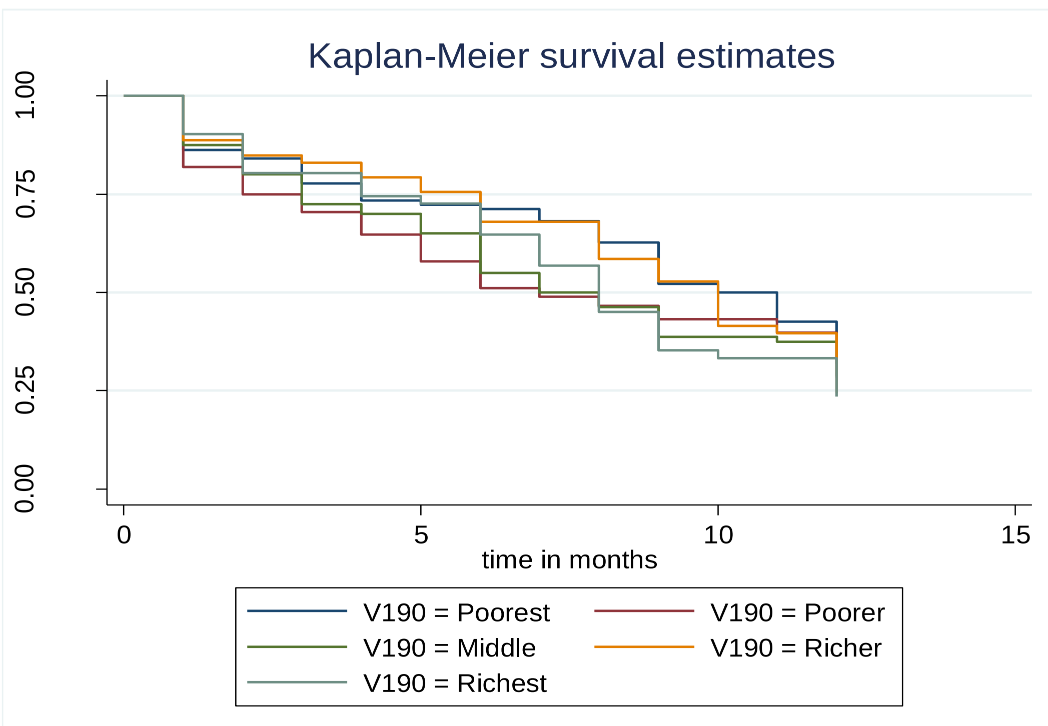

Figure 4: Survival estimates by family economic status

4.4 Log Rank Test Results

The log rank test is a popular test used to test the null hypothesis of no difference in survival between two or more independent groups. In this study, the test was used to compare the survival experience between/among groups of the variables at 5% significance level.

The results showed that size of child at birth was found to have a significant association with infant mortality with P-value = 0.003and chi2=18.20. The other covariates SHH, type of birth, area of residence and mothers' age group were also found to have a statistical significant association with infant mortality with P-value=0.014 and chi2=6.01, P-value<0.001 and chi2=21.68,P-value=0.002 and chi2=9.25 and P-value<0.001 and chi2= 47.00 respectively. The rest of the covariates were found to have no significant influence on infant mortality since the P-values were greater than 0.05 which implied that there was no statistically significant evidence that the survival distributions were not the same. Table 7 gives a summary of the log rank test results.

Table 7: Log Rank Test Results

COVARIATE	Chi2	P-VALUE
Sex of child	1.17	0.279
Family Economic Status	5.12	0.275
Size of child	18.20	0.003
Religion	10.98	0.140
Type of birth	21.68	<0.001
Source of drinking water	10.55	0.568
Mothers' Age group	47.00	<0.001
Mothers' highest education	7.07	0.069
Place of delivery	16.39	0.059
Sex of household head	6.05	0.014
Area of residence	9.25	0.002

4.5. Cox Proportional Hazard Model

The Cox hazard model tests the hypothesis that hazard ratio is equal to 1 (HR=1), meaning that there is no difference in the relative risk of death between the group of interest and the reference group.If the hazard ratio is > 1, this indicates that the treatment group has a shorter survival than the control referenced group, and if it is < 1, it indicates that the group of interest is less likely to have a shorter time to the event than the reference group.

The results for the Cox proportional hazard analysis are presented in tables 8, 9 and 10. The findings of the study showed that the overall model was highly significant with a p-value of 0.000, indicating that at least one of the covariates exerts effects on infant mortality in Malawi.

According to the results, infants who were the second multiple babies to be born and first multiple babies had a higher risk of dying before reaching the age of one than children who were born single. This covariate, type of birth, was highly significant with p-values (95% CI) of <0.001(1.639, 5.700) and 0.049 (1.004, 4.326) for 2nd and 1st multiple babies subcategories respectively. It was also found that sex of household head was significant in infant mortality. The results indicated that children born in households whose head was a female had a higher risk of dying before reaching the age of 1 year than children born in households with a male head, P-value=0.026, HR=1.37 and 95%CI= (1.037, 1.803).

For place of delivery covariate, private hospital subcategory was significant with P-value=0.020 and hazard ratio HR= 2.97 (95%CI=1.175, 7.524). This indicated that children who are born in private hospitals had a higher risk of dying 2.97 times more than children born in respondent's home before reaching the age of 1 year, whereas the hazard ratios for other groups were not statistically significant.

It was also found that infants born from families whose religion is other had a higher risk of dying compared to infants born from catholic families with P-value=0.024, HR=10.95 and 95%CI =(1.374, 87.240), whereas the hazard ratios for other groups were not statistically significant.

As can be observed from the results, two groups were highly significant from the covariate mothers' age group with p-values of <0.001. Age group 40-44 and 45-49 had hazard ratios of 3.86, 95%CI = (2.148, 6.930) and 5.22, 95%CI= (2.630, 10.337) respectively, which indicated that infants whose mother's age group was 40-44 and 45-49 had a higher risk of death compared to infants whose mother's age group was 15-20. It was also found that the confidence interval for the 45-49 age category was wide, this could be because the sample size used for the analysis was small.

As can be observed from the results table, five groups from type of drinking water covariate were significant. It was found that infants born in families whose source of drinking water was piped to neighbor, tube well/bore hole, protected well, unprotected well and river/dam had a lower risk of death compared to infants born in families whose source of drinking water was piped into dwelling.

Sex of child, size of child, mother's education, family economic status and area of residence were not statistically significant. The coefficient for female group was positive which indicated that females had a higher hazard rate and shorter survival time. For the economic status and area of residence, all groups had positive coefficients which indicated that the variables of interest had higher hazard rates and shorter survival time compared to the reference variables. The size of child covariate groups had both negative and positive coefficients. Larger than average and average infants had negative coefficients which indicated that they had lower risk of dying and had longer survival time. Whilst smaller than average and very small infants had positive coefficients which indicated that they had higher hazard rates. For mother's education, secondary and higher groups had negative coefficients which indicated that infants born from mothers of the two groups had a lower risk of death whilst infants born from primary school mothers had a higher risk of death.

Table 8: COX PH MODEL RESULTS

Covariate	Coefficient	Hazard ratio (95% conf. Interval)	Std Error	P-value
Sex of child				
male(ref)		1.000		
female	0.120	1.13(0.881 1.443)	1.14	0.339
Family economic status		1.000		
Poorest (ref)	0.091	1.13(0.764, 1.569)	0.21	0.620
Poorer	0.196	1.20(0.839, 1.764)	0.23	0.301
Middle	-0.115	0.89(0.585, 1.357)	0.19	0.593
Richer	-0.085	0.90(0.534 , 1.579)	0.25	0.759
richest				
Place of delivery				
Respondents		1.000		
home(ref)	0.459	1.41(0.521 , 4.809)	0.79	0.417
Other home	0.214	1.11(0.604 , 2.537)	0.40	0.560
Govt hospital	0.364	1.32(0.722 , 2.871)	0.46	0.301
Govt health center	0.660	1.71(0.584, 6.415)	1.04	0.280
Govt outreach	1.089	2.90(1.175 , 7.524)	1.36	0.021
Private hospital	-0.087	0.81(0.377 , 2.227)	0.36	0.847
Cham/mission	0.702	1.76(0.874 , 4.653)	0.74	0.100
hospital	0.036	1.03(0. .276 , 3.887)	0.69	0.958
Cham/mission health				
center				
Other				

Table 9:CONT OF COX PH MODEL RESULTS

Religion				
Catholic(ref)		1.000		
CCAP	-0.181	0.85(0. 519 , 1.343)	0.21	0.457
Anglican	-0.558	0.59(0.251 , 1.300)	0.25	0.183
Seventh day	-0.009	1.01(0. .558 , 1.757)	0.29	0.974
Adventist	-0.024	0.99(0. 689 , 1.384)	0.18	0.895
Other Christian	-0.173	0.85(0 .533 , 1.327)	0.20	0.457
Muslim	2.393	10.22(1.374 ,87.240)	10.84	0.024
Other				
Type of birth				
Single(ref)		1.00		
1st of Multiple	0.734	2.27(1.004 , 4.326)	0.84	0.049
2nd of multiple	1.117	3.26(1.639 , 5.700)	1.03	<0.001
Source of drinking				
water		1.00		
Piped into dwelling	-0.332	0.64(0.281 , 1.829)	0.30	0.487
Piped to yard	-1.654	0.17(0.046 , 0.795)	0.12	0.023
Piped to neighbor	-0.949	0.33(0 .145 , 1.029)	0.16	0.057
Public tap	-1.058	0.30(0.135 , 0.892)	0.14	0.028
Tube well/borehore	-1.426	0.22(0.067 , 0.856)	0.14	0.028
Protected well	-1.061	0.29(0.124 , 0.964)	0.15	0.042
Unprotected well	-0.971	0.35(0.111 , 1.297)	0.22	0.122
Unprotected spring	-1.303	0.23(0.088 , 0.829)	0.13	0.022
River/dam	-1.738	0.16(0.019 , 1.620)	0.18	0.125
Not a dejure resident				
Mothers' highest				
education		1.00		
No education(ref)	0.157	1.17(0.790 , 1.729)	0.23	0.443
Primary	-0.057	0.93(0 .559 , 1.594)	0.25	0.830
Secondary	-1.589	1.85(0.025 , 1.604)	0.20	0.131
higher				

Table 10 :CONT OF COX PH MODEL RESULTS

Sex of household head				
Male(ref)		1.00		
Female	0.313	1.37(1.037 , 1.803)	0.19	0.026
Mothers' age group				
15-19(ref)		1.00		
20-24	0.349	1.36(0.910 , 2.204)	0.31	0.122
25-29	0.307	1.31(0.842 , 2.193)	0.32	0.209
30-34	0.300	1.31(0 .822 , 2.217)	0.33	0.235
35-39	0.427	1.40(0 .900 , 2.611)	0.38	0.116
40-44	1.350	3.49(2.148 , 6.930)	1.04	<0.001
45-49	1.652	4.63(2.630 , 10.337)	1.60	<0.001
Area of residence				
Urban(ref)		1.00		
Rural	0.455	1.60(0.923 , 2.686)	0.43	0.095
Size of child				
Very large(ref)		1.00		
Larger than average	-0.115	0.89(0.527 , 1.504)	0.24	0.667
Average	-0.030	0.97(0.594 , 1.582)	0.24	0.904
Smaller than average	0.532	1.70(0.985 , 2.941)	0.48	0.056
Very small	0.009	1.01(0.497 , 2.049)	0.37	0.979
Don't know	0.900	2.46(0.815 , 7.413)	1.39	0.110

4.6 Multivariable Model Development

Likelihood ratio tests were used for the multivariable model development. The covariates which were significant in the Cox proportional hazard model are the ones which were included in the final model. Firstly, type of birth was adjusted in the final model by fitting a Cox proportional hazard model with just one variable and then fitting another model with the five variables from the final model to perform a likelihood ratio test. After adjustment for type of birth, the null hypothesis, the smaller model provides as good a fit for the data as the larger model, was rejected since the P-value was 0.0002 which was less than 0.05 and this indicated that including type of birth creates a statistically significant improvement in the fit of the final model. Two factors were then adjusted, type of birth and place of delivery.

After adjustment for type of birth and place of delivery, the P-value=0.0005 was also less than 0.05 as such the null hypothesis was rejected and conclude that the two covariates create a statistically significant improvement in the fit of the final model. However, after adjusting for type of birth, place of delivery and mothers' age group the P-value=0.436 was greater than 0.05 as such we failed to reject the null hypothesis and conclude that mothers' age group does not create an improvement to the fit of the model.

Finally, type of birth, place of delivery, sex of household head and source of drinking water covariates were adjusted for and the P-value was less than 0.001 as such we rejected the null hypothesis which indicated that the covariates create an improvement to the fit of the final model. The models were compared by displaying betas and summary statistics for each model which are presented in table 11. The BIC is Schwarz' Bayesian Information Criterion, which is a function of the log-likelihood. Smaller values indicate a better fit and as can be seen from table 4.6 model 2, with covariates type of birth and place of delivery had the lowest BIC value.

4.6.1 Model Diagnostics

The Schoenfeld residuals goodness of fit test was used to check the Cox proportional hazard model fitted in section 4.4. The null hypothesis for this test was that there are no violations of the proportional hazards assumption among the variables in the model. The results in table 12 for the global p-value for the test was found to be 0.571 and all the covariates had p-values greater than 0.05, as such the study failed to reject the null hypothesis. The results indicated that all the explanatory variables are constant over time since they are not statistically significant, hence they were no violations of the proportional hazards assumption.

Graphs of the scaled Schoenfeld residuals were then obtained for each covariate to test proportionality of each predictor. As discussed in chapter 3, a non-zero slope indicates a violation of the proportional hazard assumption. As can be seen from figures 5, 6, and 7, there was a zero slope on the graphs as such we concluded that there was no violation of the proportional hazards assumption for religion, size of child and type of birth predictors respectively. The scaled Schoenfeld residuals plots for the other variables presented in appendix B also had a zero slope which indicated that they did not violate the proportional hazards assumption.

Table 11:Test of proportional-hazards assumption

Covariate	**rho**	**Chi2**	**df**	**p-value**
Sex of child	0.040	0.43	1	0.512
Family economic status	0.097	2.39	1	0.122
Place of delivery	0.007	0.01	1	0.915
religion	0.008	0.02	1	0.896
Mother's education	-0.115	2.97	1	0.085
Type of birth	0.047	0.58	1	0.447
Source of drinking water	-0.018	0.07	1	0.785
Sex of household head	-0.029	0.22	1	0.641
Mother's age group	0.033	0.33		0.567
Area of residence	0.096	2.34	1	0.126
Size of child at birth	-0.043	0.52	1	0.473
Global test		9.55	11	0.571

Table 12: Coefficients (Log Of Hazard Ratios) And Summary Statistics

	Model 2 b	Model 3 b	Model 4 b	Model 5 b	Final model b
Type of birth single	0	0	0	0	0
1st of multiple	0 .845	0.894	0.778	0.782	0.835
2nd of multiple	1.215	1.303	1.168	1.160	1.255
Place of delivery Respondents home	0	0	0	0	0
Other home	0.410	0.281	0 .430	0.469	0 .303
Govt hospital	-0.1284	-0.035	-0.118	-0.088	-0.009
Govt health center	0.168	0.239	0.170	0.209	0.272
Govt outreach	0.328	0.475	0 .416	0.476	0.609
Other pub sector	-40.026	-39.781	-36.073	-34.959	-33.729
Private hospital	0.909	1.015	0.864	0.898	1.015
Cham/mission hospital	-0.316	-0.264	-0.279	-0.248	-0.214
Cham/mission health center	0.413	0.521	0 .435	0.489	0 .573
other	-0.166	-0.079	-0.134	-0.065	-0.0015
Mothers' age group15-19		0			0
20-24		0.281			0.287
25-29		0 .215			0.208
30-34		0.239			0.231
35-39		0.322			0.295
40-44		1.270			1.234
45-49		1.597			1.523
Source of drinking water Piped into dwelling			0	0	0
Piped to yard			-0.011	-0.018	-0.06
Piped to neighbor			-1.253	-1.268	-1.255
Public tap			-0.264	-0.282	-0.34
Tubewell/bore hole			-0.16	-0.192	-0.273
Protected well			-0.640	-0.651	-0.654
Unprotected well			-0.102	-0.145	-0.243
Unprotected spring River/dam			-36.419 0 .049	-35.491 -0.036	-34.628 -0.041
rainwater			-0.424	-0.430	-0.537
Cart with small tank			-36.237	-35.152	-35.237
other			-36.239	-35.507	-34.627
Not a dejure resident			-0.757	-0.800	-0.751
Sex of household head female				0	0
male				0.333	0.282
Chi2	29.11681**	63.08413***	39.94171*	45.66825**	76.89593***
BIC	3942.396	3957.637	4038.19	4040.664	4058.645

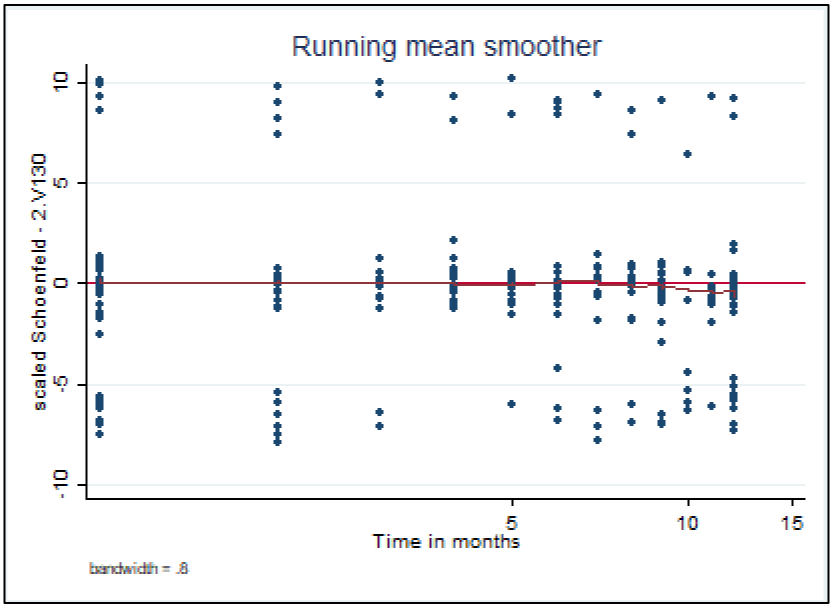

Figure 5: Religion Schoenfeld residuals plot

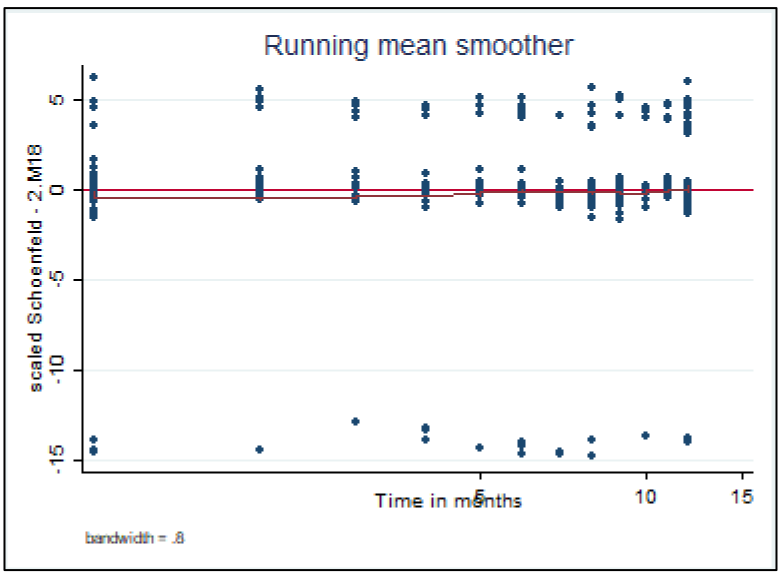

Figure 6: Size of child Schoenfeld residuals plot

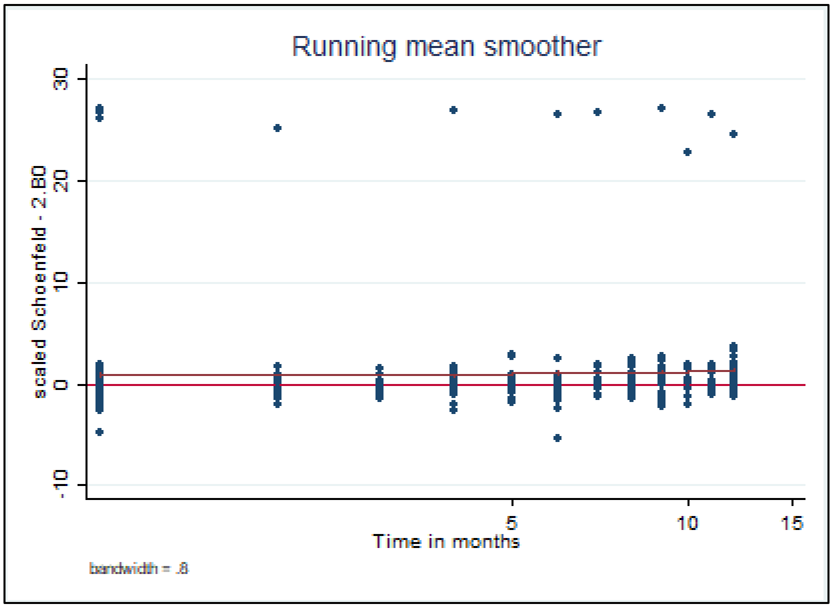

Figure 7: Type of birth Schoenfeld residuals plot

4.7. Cox Frailty Model

Two models were fitted in this section one was the Cox frailty model with community effects as the frailty term and a Cox frailty model with household effects as the frailty term.296 households and 845 communities (clusters) were considered in this study. Results for the household effects Cox frailty model are presented in tables 13, 14 and 15. By using the likelihood ratio test with a null hypothesis that the variance of the frailty term is zero (θ = 0), the chi-square test statistic (χ = 2.90) with a p-value of 0.04 at 0.05 level of significance, there was enough evidence to show the existence of unobserved heterogeneity at household level which suggests that some households were associated to a higher risk of children dying before reaching the age of one year than others. However, for the community effects, the model did not converge.

The factors that were strongly associated with infant mortality after controlling for household effects were identified by looking at the p-values and the 95% confidence intervals. The factors whose p-values were greater than 0.05 level of significance and 95%CI spanned a 1 implied that the factors were not significant and those with p-value<0.05 and 95%CI did not span a 1 implied significance. The children born in households headed by women were at a high risk of death before reaching the age of one year than those born in households where the man is the head, (P-value=0.027, HR=1.37, 95%CI =1.036, 1.803). The other factors which were found to be associated with infant mortality after controlling for household effects included place of delivery, religion, type of birth, source of drinking water, mothers' age group and size of child at birth.

Table 13: Cox Frailty Model Results

Family /household effects			
covariate	**Hazard ratio (95% confidence interval)**	**Std Error**	**P-value**
Sex of child			
male(ref)			
female	1.16(0.906, 1.486)	0.14	0.238
Family economic status			
Poorest (ref)			
Poorer	1.000		
Middle	1.14(0.792, 1.636)	0.21	0.483
Richer	1.20(0.824, 1.752)	0.23	0.339
richest	0.88(0.577, 1.352)	0.19	0.569
	0.89(0.516, 1.547)	0.25	0.689
Place of delivery			
Respondents home(ref)	1.000		
Other home	1.52(0.494, 4.670)	0.87	0.466
Govt hospital	1.12(0.543, 2.305)	0.41	0.760
Govt health center	1.35(0.067, 2.725)	0.48	0.390
Govt outreach	1.76(0.524, 5.936)	1.09	0.359
Private hospital	3.12(1.225, 7.954)	1.49	0.017
Cham/mission hospital	0.81(0.330, 1.980)	0.37	0.643
Cham/mission health center	1.81(0.077, 4.201)	0.78	0.168
other	1.00(0.265, 3.796)	0.68	0.996

Table 14: Cont Of Cox Frailty Model Results

Religion			
Catholic(ref)			
CCAP	0.85(0.526, 1.373)	0.21	0.508
Anglican	0.58(0.253, 1.323)	0.24	0.195
Seventh day Adventist	1.00(0.559, 1.768)	0.29	0.985
Other Christian	0.98(0.686, 1.389)	0.18	0.894
Muslim	0.84(0.529, 1.327)	0.20	0.452
other	1.00(1.192, 83.778)	10.84	0.034
Type of birth			
Single(ref)			
1st of Multiple	2.13(1.121, 4.844)	0.81	0.023
2nd of multiple	3.14(1.765, 6.238)	1.02	<0.001
Source of drinking water			
Piped into dwelling(ref)	1.00		
Piped to yard	0.63(0.247, 1.618)	0.30	0.430
Piped to neighbor	0.16(0.038 , 0.675)	0.12	0.013
Public tap	0.31(0.116, 0.831)	0.16	0.020
Tube well/borehore	0.28(0.107, 0.718)	0.13	0.008
Protected well	0.20(0. 0547, 0.701)	0.13	0.012
Unprotected well	0.27(0.097, 0.770)	0.15	0.014
Unprotected spring	0.33(0.094, 1.156)	0.21	0.111
River/dam	0.21(0.0699, 0.669)	0.13	0.008
Not a dejure resident	0.16(0.016, 1.453)	0.18	0.103
Mothers' highest education	1.00		
No education(ref)	1.18(0. 787, 1.730)	0.24	0.442
Primary	0.93(0. 545, 1.563)	0.25	0.766
Secondary	0.20(0 .022, 1.452)	0.21	0.108
higher			

Table 15: Cont Of Cox Frailty Model Results

Sex of household head	1.00		
Male(ref)	1.37(1.033, 1.807)	0.19	0.028
female			
Mothers' age group			
15-19(ref)	1.00		
20-24	1.39(0 .852, 2.074)	0.32	0.209
25-29	1.37(0. 808, 2.116)	0.34	0.274
30-34	1.34(0.782, 2.121)	0.33	0.320
35-39	1.55(0.821, 2.383)	0.43	0.216
40-44	3.80(1.868, 6.075)	1.16	<0.001
45-49	5.21(2.279, 9.171)	1.87	<0.001
Area of residence			
Urban(ref)	1.00		
rural	1.60(0. 981, 2.904)	0.44	0.058
Size of child			
Very large(ref)	1.00		
Larger than average	0.88(0. 517, 1.493)	0.209	0.633
Average	0.98(0.593, 1.598)	0.24	0.918
Smaller than average	1.78(1.00, 3.048)	0.50	0.046
Very small	1.01(0.484, 2.044)	0.37	0.991
Don't know	2.27(0.745, 7.038)	1.30	0.148
Frailty Variance	0.17	0.12	

4.8 Parametric Frailty Models

Parametric frailty models were fit to select the best fit model. Two parametric frailty models werefit for both the household and community effects using two different distributions making a total of four models. The first parametric frailty models to be fitted were the Weibull distribution model to assess if there exists unobserved heterogeneity at household and community level. The results are presented in tables 16, 17 and 18. The other models were fit using the log normal distribution and the results are presented in appendix C. For the Weibull/gamma models, the results indicated that there were no unobserved heterogeneity at household level with P-value=0.056 and θ=2.54 and no unobserved heterogeneity at community level with P-value=1.000 and θ=0.00. Whilst for the log normal distribution models, it was found that there were no unobserved heterogeneity at household level with P-value=0.090 and $\theta = 1.80$ and also that there were no unobserved heterogeneity at community level with P-value=0.499 and $\theta =$ 2.6e-06.

Table 16: Parametric Frailty Models

	Household effects (weibull)			Community effects (weibull)		
COVARIATE	**Hazard ratio(95% confidence interval)**	**SE**	**P-value**	**Hazard ratio(95% confidence interval)**	**SE**	**P-value**
Sex of child						
male(ref)	1.000					
female	1.14(0.891 , 1.467)	0.146	0.292	1.139(0.891 , 1.469)	0.1143	0.300
Family economic status						
Poorest (ref)						
Poorer	1.096(0.762, 1.575)	0.203	0.622	1.091(0.762 ,1.563)	0.200	0.632
Middle	1.202(0.824, 1.754)	0.232	0.339	1.210(0.835,1.755)	0.229	0.313
Richer	0.861(0.563, 1.319)	0.187	0.492	0.868(0.579,1.322)	0.186	0.511
richest	0.904(0.523, 1.561)	0.252	0.717	0.916(0.534, 1.568)	0.251	0.750
Place of delivery						
Respondents home(ref)	1.000					
Other home	1.662(0.536, 5.153)	0.959	0.397	1.543(0.509, 4.676)	0.91	0.442
Govt hospital	1.182(0.573, 2.439)	0.436	0.651	1.174(0.575, 2.398)	0.46	0.659
Govt health center	1.434(0.713, 2.884)	0.511	0.312	1.405(0.706, 2.795)	0.51	0.332
Govt outreach	1.921(0.571, 6.470)	1.190	0.292	1.865(0.565, 6.166)	1.21	0.307
Private hospital	3.475(1.358 , 8.894)	1.666	0.009	3.229(1.288, 8.098)	1.43	0.012
Cham/mission hospital	0.859(0.351, 2.102)	0.392	0.739	0.865(0.35, 2.096)	0.42	0.749
Cham/mission healthcenter	1.992(0.855, 4.643)	0.860	0.110	1.139(0.843, 4.462)	0.87	0.119
other	0.988(0.258, 3.737)	0.669	0.979	0.955(0.265, 3.732)	0.70	0.995

Table 17: CONT OF PARAMETRIC FRAILTY MODELS

Religion						
Catholic(ref)	1.000					
CCAP	0.854(0.528, 1.382)	0.209	0.521	0.854(0.531, 1.376)	0.208	0.518
Anglican	0.572(0.249, 1.312)	0.242	0.187	0.591(0.260, 1.341)	0.247	0.209
Seventh day	1.037(0.582, 1.851)	0.306	0.900	1.051(0.594 , 1.863)	0.306	0.862
Adventist	0.979(0.688, 1.394)	0.177	0.907	0.986(0.696, 1.398)	0.175	0.939
Other Christian	0.851(0.536, 1.353)	0.201	0.496	0.871(0.552, 1.375)	0.202	0.554
Muslim	11.671(1.362, 99.929)	12.784	0.025	11.238(1.410, 89.554)	11.901	0.022
other						
Type of birth						
Single(ref)	1.000					
1st of Multiple	2.126(1.017, 4.443)	0.799	0.045	2.071(1.003, 4.278)	0.766	0.049
2nd of multiple	2.957(1.564, 5.591)	0.960	<0.001	2.881(1.545, 5.374)	0.916	0.001
Source of drinking water						
Piped into	1.00					
dwelling(ref)	0.671(0.261, 1.724)	0.323	0.407	0.670(0.265, 1.695)	0.317	0.399
Piped to yard	0.177(0.042 , 0.743)	0.129	0.018	0.186(0.045, 0.770)	0.134	0.020
Piped to neighbor	0.354(0.132, 0.952)	0.178	0.040	0.381(0.145, 1.001)	0.187	0.050
Public tap	0.315(0.121, 0.821)	0.154	0.018	0.339(0.133, 0.864)	0.161	0.023
Tube well/borehore	0.206(0.057, 0.746)	0.135	0.016	0.227(0.064, 0.808)	0.147	0.022
Protected well	0.291(0.103, 0.826)	0.155	0.020	0.314(0.113, 0.869)	0.163	0.026
Unprotected well	0.350(0.100, 1.226)	0.223	0.101	0.372(0.109 , 1.26)	0.232	0.114
Unprotected spring	0.248(0.080, 0.773)	0.144	0.016	0.270(0.089, 0.819)	0.153	0.021
River/dam	0.160(0.017, 1.496)	0.183	0.108	0.166(0.018, 1.518)	0.187	0.112
Not a dejure resident						
Mothers' highest education						
No education(ref)	1.00					
Primary	1.133(0.764, 1.681)	0.227	0.535	1.132(0.766, 1.673)	0.225	0.532
Secondary	0.897(0.529, 1.524)	0.242	0.690	0.910(0.535, 1.537)	0.243	0.727
higher	0.169(0.021 , 1.349)	0.179	0.094	0.175(0.022, 1.379)	0.184	0.098

Table 18: CONT OF PARAMETRIC FRAILTY MODELS

Sex of household head						
Male(ref)	1.00					
female	1.37(1.028, 1.801)	0.194	0.032	1.366(1.035, 1.802)	0.193	0.027
Mothers' age group						
15-19(ref)	1.00					
20-24	1.431(0.917, 2.232)	0.324	0.114	1.457(0.938, 2.264)	0.327	0.094
25-29	1.375(0.849, 2.227)	0.338	0.196	1.378(0.854, 2.223)	0.336	0.189
30-34	1.336(0.809, 2.204)	0.341	0.257	1.348(0.821, 2.214)	0.341	0.237
35-39	1.607(0.943, 2.746)	0.438	0.081	1.602(0.942, 2.723)	0.411	0.082
40-44	3.692(2.034, 6.703)	1.123	<0.001	3.798(2.114, 6.825)	1.135	<0.001
45-49	5.538(2.699, 10.955)	1.943	<0.001	5.415(2.736, 10.719)	1.886	<0.001
Area of residence						
Urban(ref)	1.00					
rural	1.525(0.887, 2.623)	0.421	0.127	1.499(0.878, 2.558)	0.408	0.137
Size of child						
Very large(ref)	1.00					
Larger than average	0.916(0.540, 1.555)	0.24	0.747	0.929(50.51, 1.567)	0.24	0.783
Average	1.00(0.610, 1.640)	0.24	0.998	1.001(0.614, 1.632)	0.24	0.997
Smaller than average	1.799(1.036, 3.126)	0.50	0.037	1.748(1.013, 3.016)	0.48	0.045
Very small	1.096(0.537, 2.239)	0.37	0.801	1.123(0.555, 2.270)	0.37	0.746
Don't know	2.177(0.709, 6.679)	1.30	0.174	2.369(0.789, 7.116)	1.38	0.124
constant	0.003(0.000, 0.011)	0.002	<0.001	0.002(0.000, 0 .010)	0.001	<0.001
ln(p)	0.459(0.357, 0.562)	0.052	<0.001	0.452(0.3498, 0.555)	0.520	<0.001
ln(theta)	-1.836(-3.228, -0.445)	0.709	0.010	-15.430(-1261.713, 1230.852)	635.780	0.981
p	1.154(1.429, 1.755)	0.833		1.572(1.419, 1.743)	0.827	
1/p	0.631(0.569, 0.699)	0.033		0.635(0.573, 0.704)	0.334	
theta	0.159(.0039, 0.640)	0.113	0.056	5.32e-07	0.000	1.000

4.9 Best Fitting Model Selection

The best fit model was selected using Akakian Information Criteria (AIC) and the Log-likelihood ratio test. The lowest Akakian Information Criteria and the highest Log-likelihood ratio value indicates the best fit model. For both the household and community effects models, the Weibull-gamma frailty models were found to be the best fit models since they had the lowest AIC values as shown in table 19.

Table 19: Model comparison with different distributional assumptions

Model	Baseline Hazard distribution	Frailty distribution	Frailty variance (p-value)	AIC	BIC	LRR
Cox model-household effects	N/A	gamma	0.174 (0.043)	3873.20	4189.47	-1885.60
Cox model-community effects	N/A	gamma	Did not converge			
Shared frailty-household effects	Weibull	gamma	0.159 (0.056)	2027.597	2362.472	-959.7985
Shared frailty-community effects	Weibull	gamma	5.32e-07 (1.000)	2030.136	2363.001	-961.068
Shared frailty-household effects	Log normal	gamma	0.130 (0.90)	2061.138	2396.012	-976.569
Shared frailty-community effects	Log normal	gamma	2.11e-07 (1.000)	2062.941	2397.470	-977.470

CHAPTER 5

DISCUSSION, CONCLUSION AND RECOMMENDATIONS

The first section of this chapter provides a discussion of the results presented in the previous chapter, the conclusion is provided in the second section and finally recommendations in the last section.

5.1 Discussion

This study used survival analysis and frailty modelling to examine the factors that are associated with infant mortality in Malawi. The descriptive statistics for age of child in months indicate that the mean for infants who died is 2.25 months which is in agreement with what was reported by (UNICEF DATA, 2020) that a child's survival is most vulnerable within the first 28 days of life.

Two groups of mother's age at child birth were found to have a significant association with infant mortality in both the log rank test and Cox proportional hazard model. The results showed that children born from mothers aged 40-44 and 45-49 years had a higher chance of dying before reaching the age of 1 year compared to children born from mothers aged 15-19 years. This then indicated that women who have children at older age, 40-49 years, had significantly increased risk of infant mortality in comparison to the women who had their child at younger ages, less than 18 years. It is expected that children born to young mothers (aged less than 20 years) and those born to older mothers (aged 40-49 years) should have higher mortality than those born to mothers aged 20-39 years (Kembo, 2009) which is similar to what has been found in this study. Young mothers are said to be at higher risk of experiencing infant mortality due to their emotional and psychological immaturity (Dube, 2012) and also because young girls/adolescents delay reacting to pregnancies might lack knowledge of the correct health-seeking methods with regard to their pregnancies (Phipps et al., 2002). The results from this study however, are in contradiction with what Lemani (Lemani, 2013) found when covariates of infant mortality in Malawi were modelled where it was observed that Children born to mothers aged less than 20 years had an increased risk of death compared to those born when their mothers were aged above 20 (Lemani, 2013) and (Dube, 2012) found that women who had children at a young age had significantly increased chances of infant mortality (59%) in comparison to the women who had their children at older ages. This contradiction could be because pregnancy after 40 years of age has a risk of complications such as high blood pressure, preeclampsia, gestational diabetes and birth abnormalities(Kay & Villines, 2020). This is why pregnancy after 40 years of age requires quality prenatal care, healthy lifestyle maintenance and health center delivery which are things that some women in rural Malawian residence don’t have access to and this might lead to high infant mortality rates among older women in Malawi.

An interesting finding of this study is that mothers’ highest education had no significant association with infant mortality which contradicts the studies of (Omariba *et al*., 2007), which found that mothers with secondary education had a 20% lower chance of experiencing infant mortality compared to mothers’ who were not educated. Oftenly, maternal education is viewed as an indication of level of skills and knowledge of the mother which help in the effective use of available child care resources such as health services which is assumed to lower child and infant mortality, however, the findings from this study indicated that there was no significant association between infant mortality and mother's education level. This was in agreement with studies such as (Dube, 2012) which found that education was an insignificant determinant of infant mortality in Zimbabwe and this is expected because mother's education is more strongly associated with child mortality than infant mortality (Lemani, 2013). These results were also similar to the (Makoena, 2011) study on risk factors associated with high infant and child mortality in Lesotho where it was found that there was no significant association between mother's education and childhood mortality.

Religion is a factor that is assumed to play a part in infant mortality because of the different beliefs that religions have. For example Pentecostalism is usually characterized by less trust in conventional medicine which make these groups particularly sensitive to the threats to the welfare of young children which might account for higher infant mortality rates in communities with a large proportion of Pentecostal churches (Garcia *et al.*, 2012).This factor was found to be insignificant by the log rank test but after adjusting for other covariates some of its subcategories were found to be significant in the Cox proportional hazard model. The results indicated that Children born from families with other religion category have a higher chance of dying before reaching the age of one year compared to children born from catholic families. The results from this study are similar to what was observed in Zimbabwe where religion was significantly associated with infant mortality in such a way that members of Zionist and Apostolic churches showed a historical higher infant mortality than members of mission churches(Gregson *et al.*, 1999).

Both the log rank test and Cox proportional hazard model results showed that sex of household head was a significant covariate. Children born from female headed household (FHH) were more likely to die before the age of one year compared to children from male headed households (MHH) headed by males. This is the case because there is a difference in economic conditions of FHH and MHH in such a way that FHH are generally poorer than MHH and families with low economic status(poor) usually have higher percentage of infant deaths due to lack of better access to health services (Gupta, Ashish Kumar, Borkotoky, 2015).

For the environmental factors, place of delivery and source of drinking water are the covariates whose subcategories were found to have a significant association with infant mortality. The results were in agreement with studies such as (Folasade, 2000), which found that source of drinking water and child mortality were significantly associated in Nigeria. The (Ezeh, Osita,Agho, 2014) study also had consistent results where it was found that the mortality from unimproved water and sanitation was significantly higher by 38% compared to improved water and sanitation.

The source of drinking water has an impact on infant mortality in a way that children are more vulnerable to the health hazards associated with unimproved water supply and sanitation because their immune, respiratory and digestive systems are still developing (Ezeh, Osita,Agho, 2014). The place of delivery results of this study are in agreement with the (Ajaari et al., 2012) study which found that there were more neonatal deaths among deliveries outside health facilities than among deliveries within health facilities. Place of delivery is a significant predictor of infant mortality because children delivered at a health facility are likely to experience lower mortality than children delivered at home because health facilities provide sanitary environment and medically correct birth assistance (Ajaari *et al.*, 2012).

Infant mortality is higher in boys than girls in most parts of the world and this has been explained by sex differences in genetic and biological makeup in such a way that boys are biologically weaker and more susceptible to diseases and premature death (Pongou, 2013). This study however, had another interesting finding, the Cox proportional hazard model results indicated that males had a lower chance of dying before reaching the age of one year than females. These results are in contrast with the (Ashorn *et al.*, 2002) and (Lemani, 2013) studies which found that males had a higher risk of infant mortality compared to females.

Type of birth was also found to have a highly significant association with infant mortality in both the log rank analysis and Cox proportional hazard model. It was found that women who have single births had a lower risk of experiencing infant mortality compared to mothers who have multiple births. These results are consistent with the results from (Uthman et al., 2008) study which found that children born multiple births were more than twice as likely to die during infancy as infants born singleton. A possible reason for this observation is that multiple births are high-risk births that require special and expensive care (Uthman et al., 2008). Multiple-birth children are also at a greater risk of birth defects and disabilities (Uthman *et al.*, 2008).

Frailty modelling of infant mortality was done using the semi-parametric methods and parametric methods. For the semi-parametric frailty models to account for family/household effects and community effects results indicated that infant deaths are likely to cluster in families due unmeasured factors which is in agreement with studies such as (Madise & Diamond, 1995) and (Bolstad & Manda, 2001) which found that infant mortality tend to cluster in families in Malawi. This indicates that there are differences in infant mortality risks between families and the family effect may be a result of biological factors, such as hereditary diseases, or different child care practices, immunization and nutrition (Madise & Diamond, 1995).Whilst for the community effects there was not enough evidence to show the existence of unobserved heterogeneity at community level because the model did not converge. This would be expected, because during the infant period, the child is mostly kept in the house and does not interact much with the community outside the family. These results are in contrast with the (Omariba *et al*., 2007) study where it was found that effects of unmeasured environmental factors and community factors are important for child mortality.

For the parametric frailty models, the first part, Weibull distribution and gamma frailty were used to fit the models. It was found that there was not enough evidence that both household and community effects play a role in infant mortality. The second part used log normal distribution and gamma frailty, and there was not enough evidence to conclude for both household and community effects.

5.2 Conclusion

This study examined the factors associated with infant mortality whilst controlling for household and community effects. It was found that SHH, mothers' age group, source of drinking water, religion, type of birth and place of delivery had a significant association with infant mortality. Particularly, the results indicated that FHH are at a higher risk of experiencing infant mortality and mothers who had single births had a lower chance of experiencing infant mortality compared to mothers who had multiple births. They also indicated that mothers who had children at older ages were at a higher risk of experiencing infant mortality compared to women who birthed children at younger ages. Furthermore, this study found that there existed some unobservable family/household effects which tend to make infant deaths cluster in some families.

Although Malawi has managed to achieve a significant reduction in infant and child mortality rates, the rates remain high compared to most African countries as such there is need for more effort to reduce these mortality rates. This study provided insights into the risk factors of infant and child mortality in Malawi, which contains vital information for health policy makers in government and non-governmental organizations.

In conclusion, this study revealed that SHH, mothers' age group, source of drinking water, religion, type of birth and place of delivery are associated infant mortality and that there are unobservable family effects which make infant deaths to cluster in some families. These factors need to be considered when planning and developing policies against infant mortality in order to successfully work towards reducing infant mortality rate in Malawi.

5.3 Recommendations

Recommendations from this study are that, mothers should be sensitized on the importance of having child deliveries in health facilities to avoid birth complications and loss of child if birthed at home. Women who have a family history of multiple births should be educated on the high risks that multiple births have and should be closely monitored during pregnancy. There is also need to identify and educate the religions that deny their members of traditional and medical medicine about the importance of medicine.

Women should be encouraged to have children before they turn the age of 40 years to avoid the birth complications which occur due to old age. Families need to be sensitized on the importance of using clean water as a source of drinking water and teach them ways of cleaning water such as boiling the water before use. Families which have experienced multiple infant deaths need to be identified as vulnerable households and studied to find out what makes them vulnerable and help reduce the infant mortality rate.

This study has been conducted using nationally representative data with a large sample size, although this sample is a fragment of the population of the Malawian women who experienced childbearing in the last five years before data collected, it gives a representative picture of the population at the time of the MDHS 2015-2016 survey. Thus, the sample can be understood to be a true reflection of the Malawian women who had given birth in the last five years prior to the survey.

However, the study had a number of limitations in both data sources and methodology which might have affected the results. There are some critics that the DHS survey is associated with which include collection of data from women aged 15-49 who are alive in a given household which means that no information is collected for mothers who have died and this creates a bias in the results. It was a challenge to control for community effects because the MDHS data doesn’t have specific community characteristics.

REFERENCES

Afeez, B. M., Maxwell, O., Osuji, G. A., & Chinedu, I. U. (2018). Reprospective Analysis of Some Factors Responsible for Infant Mortality in Nigeria : Evidence from Nigeria Demographic and Health Survy (NDHS). *American Journal of Mathematics and Statistics*, *8*(6), 184–189. https://doi.org/10.5923/j.ajms.20180806.04

Ajaari, J., Masanja, H., & Owusu-Agyei, S. (2012). Impact of Place of delivery on Neonatal Mortality in Rural Tanzania. *Global Health*, 49–59.

Ashorn, P., Maleta, K., & Espo, M. (2002). Male biased mortality among 1-2 year old children in rural Malawi. *BMJ Open*.

Bolstad, W. M., & Manda, S. O. (2001). Investigating child mortality in Malawi Using Family and community Random Effects : A Bayesian Analysis. *Journal of the American Statistical Association*, *96*(453), 12–19.

Cox, D. R. (1972). Regresson models and Life-Tables. *Journal of the Royal Statistics Society*, *34(2)*, 187–220.

Damato, B., Eleuteri, A., & Taktak, A. (2011). Estimating prognosis for survival after treatment of choroidal melanoma. *PUBMED*, 285–295.

Dube, Z. B. (2012). *The relationship between mothers' maternal age and infant mortality in Zimbabwe*. University Of Witwatersrand.

Ezeh, Osita,Agho, K. (2014). The impact of water and sanitation on childhood mortality in nigeria: Evidence from demographic and health surveys 2003-2013. *International Journal of Environmental Research and Public Health*, 9256–9272.

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

Garcia, G., Bartkowski, J. P., & Xu, X. (2012). Religion and Infant Mortality in the United States : A Community- Level Investigation of Denominational Variations. *Journal for the Scientific Study of Religion*, *May*.

Gondwe, K. W., Walker, R. J., Mkandawire-Valhmu, L., Dressel, A., Ngui, E. M., Kako, P. M., & Egede, L. (2021). Predictors of wealth index in Malawi – Analysis of Malawi demographic Health Survey 2004–2015/16. *Public Health in Practice*, *2*(November 2020), 100059. https://doi.org/10.1016/j.puhip.2020.100059

Gong, Q., & Fang, L. (2013). Comparison of different parametric proportional hazards models for interval-censored data:a simulation study. *PUB MED*, *36(1)*, 276–283.

Gregson, S., Zhuwau, Z., Anderson, R. M., & Chandiwana, S. K. (1999). Apostles and Zionists:the influence of religon on demographic change in rural zimbabwe. *Population Studies(Camb)*.

Gupta , Ashish Kumar, Borkotoky, M. (2015). Household headship and infant mortality in India: Evaluating the determinants and differentials. *Global Heath Journal*, 44–52.

Hanagal, D. (2011). *Modelling survival data using frailty models*.

Hill, Kenneth, Amouzou, A. (2006). Trends in child mortality, 19960-2000. *The World Bank*.

Hug, Lucia, Sharrow, David, Zhong, K. (2018). Levels and trends in child mortality. *World Health Organization*.

Kalipeni, E., & Moise, I. (2015). Assessing the reduction in infant mortality rates in Malawi over the 1990-2010 decades. *Global Public Health*, 757–779.

Kay, C., & Villines, Z. (2020). what to kow about pregnancy after 40. *Medical News Today*.

Kazembe, L., Clarke, A., & Kandala, N.-B. (2012). The effects of Education, income, and child mortality on fertility in South Africa. *BMJ Open*.

Kembo, J. (2009). Determinants of infant and chld mortality in Zimbabwe: Results of multivariate hazard analysis. *Demographic Research*.

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

Khosa, S. K. (2019). *Parametric proportional hazard models with applications in survival analysis*. University of Saskatchewan.

Lemani, C. (2013). *Modelling Covariates of infant and child mortality in Malawi*. university of capetown.

Madise, N. J., & Diamond, I. A. N. (1995). *Determinants Of Infant Mortality In Malawi : An Analysis To Control For Death Clustering Within Families*. 95–106.

Makoena, M. (2011). *Risk factors associated with high infant and child mortality in Lesotho*.

Nasejje, J. B. (2015). *University Of Kwazulu-Natal Under-Five Child Mortality* (Issue September 2013). University of Kwazulu-Natal.

National Statistical Office (NSO). (2017). *Malawi Demographic and Health Survey 2015-16.*

Nations, U. (1989). *Step By Step Guide To The Estimation Of Child Mortality*.

Ndawala, J. (2015). Infant and Child Mortality. *DHS Program*, 97–104.

Niragire, F., Wangombe, A., & Achia, T. N. O. (2011). Use of the shared frailty model to identify the determinants of child mortality in Rwanda. *Rwanda Journal*, *20*, 89–105.

Nyinawajambo, M. R. (2018). *Survival Analysis of Time to Event Data An application to child mortality in Sub-Saharan Africa Region using Their Demographic and Health Surveys*. Orebro University.

Omariba, W., Beaujot, R., & Rajulton, F. (2007). Determinants of infant and child mortality in Kenya:An analysis controlling for frailty effects. *Population Research and Policy Review*, 299–321.

Ouatarra, D. (2018). *Combatting Infant Mortality: A Priority for Africa*. International Policy Digest. https://intpolicydigest.org/combatting-infant-mortality-a-priority-for-africa/

Phipps, M., Blume, J., & DeMonner, S. (2002). Young maternal age associated with increased risk of postneonatal death. *PUBMED*, 481–486.

Pongou, R. (2013). Why is infant mortality higher in boys than in girls? A new hypothesis based on preconception environment and evidence from a large sample of twins. *PUB MED*.

Roser, Max, Ritchie, Hannah, Dadonaite, B. (2013). Population health metrics. *Our World in Data*.

S.Nutiye. (2009). *Determinants of Infant Mortality in Turkey*. *4*(September), 113–122.

Saikia, R., & Barman, M. P. (2017). A Review on Accelerated Failure Time Models. *International Journal of Statistics and Systems*, *12*(2), 311–322.

Sartorius, Kurt, Sartorius, B. (2014). Global infant mortality trends and attributable determinants-an ecological study using data from 192 countries for the period 1990-2011. *Population Health Metrics*.

Sperandei, S. (2016). *Lessons in biostatistics Understanding logistic regression analysis*. *February 2014*. https://doi.org/10.11613/BM.2014.003

Tesfa, D., Tiruneh, S. A., Azanaw, M. M., Gebremariam, A. D., Engdaw, M. T., Kefale, B., Abebe, B., & Dessalegn, T. (2021). Time to death and its determinants among under-five children in Sub-Saharan Africa using the recent (2010–2018) demographic and health survey data: country-based shared frailty analyses. *BMC Pediatrics*, *21*(1), 1–12. https://doi.org/10.1186/s12887-021-02950-3

Treibe, L. A. (2009). Infant mortality rate. *Publications of Encyclopedia of Gender and Society*.

UNICEF DATA. (2020). *Neonatal mortality*.

Uthman, O., Uthman, M., & Yahaya, I. (2008). A population-based study of effect of multiple birth on infant mortality in Nigeria. *BMC Pregnancy and Childbirth*.

Vaupel, J., & Manton, K. (1979). The Impact of heterogeneity in individual frailty on te dynamics of mortality. *JSTORI*, 439–454.

Walker, N., Hill, K., & Zhao, F. (2012). Child Mortality Estimation: Methods Used to Adjust for Bias due to AIDS in Estimating Trends in Under-Five Mortality. *PUBMED*.

Weldearegawi, B., Adama, Y., & Spigt, M. (2015). Infant Mortality and causes of infant deaths in rural Ethiopia:a population-based cohort of 3684 births. *BMC Public Health*.

Wienke, A. (2003). Frailty Models. *Max Planck Institute for Demographic Research*, *49*(0), 0–13.

APPENDICES

APPENDIX A

Table 20: STATA CODES

1	use "C:\Users\Esther.Khundi\Desktop\school\graphs\dhs2.dta"
2	gen hypage=(V008-B3)/12
3	label varhypage "Age of child at interview"
4	gentimeyears=.
5	replace timeyears=hypage
6	replace timeyears=(B7/12) if B5==0
7	gen dead=(B5==0)
8	label vartimeyears "survival time of the child in years"
9	label var dead "the child is dead"
10	gen hypmonth=V008-B3
11	gentimemonths=.
12	replace timemonths=hypmonth
13	replace timemonths=B7 if(B5==0)
14	label varhypmonth "age of child in months"
15	gen status=.
16	replace status=0 if B5==1
17	replace status=1 if B5==0

18	label var status "The survival status of the child"
19	label define status 1"dead" 0"alive"
20	label values status status
21	stsettimemonths, failure(status==1) exit(timemonths==12)
22	tabstattimemonths, by(dead) statistics(n mean sd min q max) columns(statistics) format(%8.2f)
23	tabulate dead B4, row
24	tabulate dead V190, row
25	tabulate dead M18, row
26	tabulate dead V130, row
27	tabulate dead V151, row

Table 21: STATA CODES CONTI…

28	tabulate dead V102, row
29	tabulate dead M15, row
30	tabulate dead V113, row
31	tabulate dead V106, row
32	tabulate dead V013, row
33	tabulate dead B0, row
34	ltabletimemonths status
35	sts graph, by(B4) risktablegraphregion(fcolor(white))
36	sts graph, by(V190) risktablegraphregion(fcolor(white))
37	sts graph, by(V013) risktablegraphregion(fcolor(white))
38	sts graph, by(V113) risktablegraphregion(fcolor(white))
39	sts graph, by(V106) risktablegraphregion(fcolor(white))
40	sts graph, by(V102) risktablegraphregion(fcolor(white))
41	sts graph, by(V151) risktablegraphregion(fcolor(white))
54	sts test V130
55	sts test V151
56	sts test V190
57	stCox i.B4 i.V190 i.M15 i.V130 i.V106 i.B0 i.V113 i.V151 i.V013 i.V102 i.M18
58	stCox i.B0

59	eststo model_1
60	stCox i.V013 i.V113 i.B0 i.M15 i.V151
61	eststomodel_final
62	lrtest model_1 model_final
63	stCox i.B0 i.M15
64	eststo model_2
65	lrtest model_2 model_final
66	stCox i.B0 i.M15 i.V013
67	eststo model_3

Table 22: STATA CODES CONTI…

68	lrtest model_3 model_final
69	stCox i.B0 i.M15 i.V113
70	eststo model_4
71	lrtest model_4 model_final
72	stCox i.B0 i.M15 i.V113 i.V151
73	eststo model_5
74	lrtest model_5 model_final
75	estout model_2 model_3 model_4 model_5 model_final, stats(n chi2 bic, star(chi2)) prehead("Betas")
76	estatphtest, detail
77	stphtest, log plot(2.B0) yline(0) graphregion(fcolor(white))
78	stphtest, log plot(2.B4) yline(0) graphregion(fcolor(white))
79	stphtest, log plot(12.M15) yline(0) graphregion(fcolor(white))
80	stphtest, log plot(2.M18) yline(0) graphregion(fcolor(white))
81	stphtest, log plot(2.V102) yline(0) graphregion(fcolor(white))
82	stphtest, log plot(2.V106) yline(0) graphregion(fcolor(white))
83	stphtest, log plot(2.V013) yline(0) graphregion(fcolor(white))
84	stphtest, log plot(12.V113) yline(0) graphregion(fcolor(white))
85	stCox i.B4 i.V190 i.M15 i.V130 i.B0 i.V113 i.V106 i.V151 i.V013 i.V102 i.M18,shared(V002)
86	estatic
87	stCox i.B4 i.V190 i.M15 i.V130 i.B0 i.V113 i.V106 i.V151 i.V013 i.V102 i.M18,shared(V001)

88	streg i.B4 i.V190 i.M15 i.V130 i.B0 i.V113 i.V106 i.V151 i.V013 i.V102 i.M18, dist(weib) frailty(gamma) shared(V002)
89	estatic
90	streg i.B4 i.V190 i.M15 i.V130 i.B0 i.V113 i.V106 i.V151 i.V013 i.V102 i.M18, dist(weib) frailty(gamma) shared(V001)
91	estatic
92	streg i.B4 i.V190 i.M15 i.V130 i.B0 i.V113 i.V106 i.V151 i.V013 i.V102 i.M18, dist(lnormal) frailty(gamma) shared(V002)
93	estatic
94	streg i.B4 i.V190 i.M15 i.V130 i.B0 i.V113 i.V106 i.V151 i.V013 i.V102 i.M18, dist(lnormal) frailty(gamma) shared(V001)
95	estatic

APPENDIX B

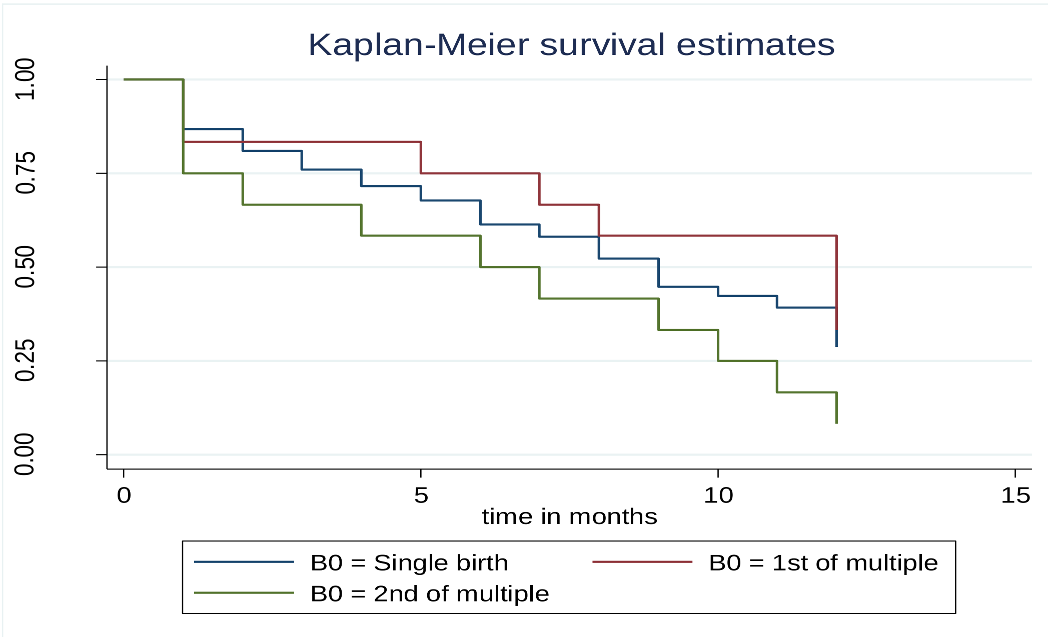

Figure 8: Survival estimates by type of birth

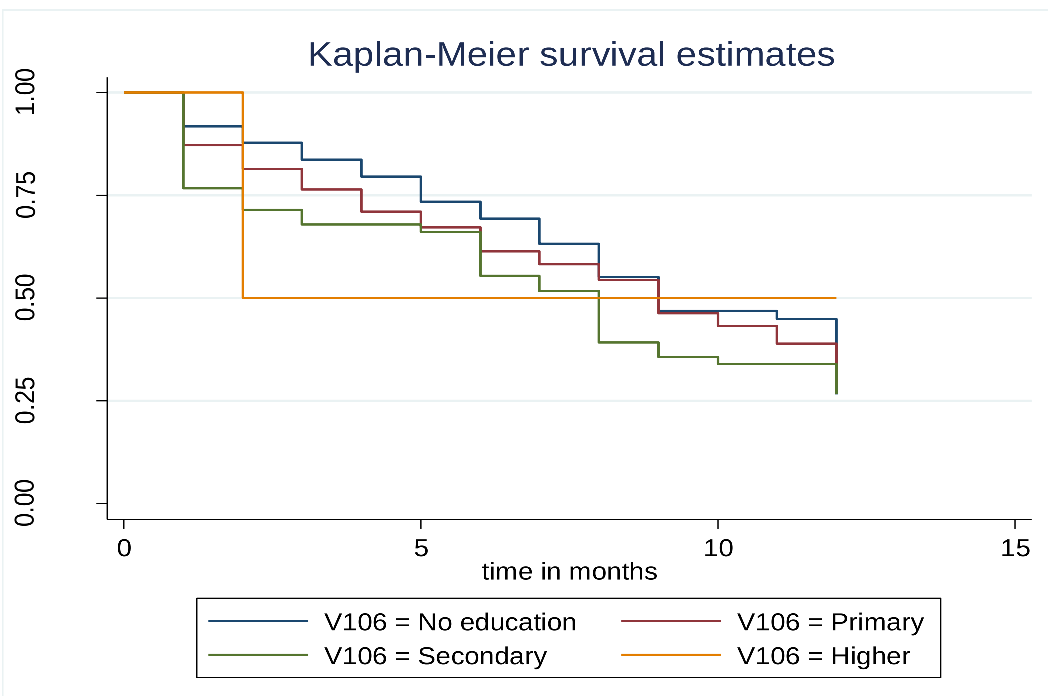

Figure 9: Survival estimates by mother's education

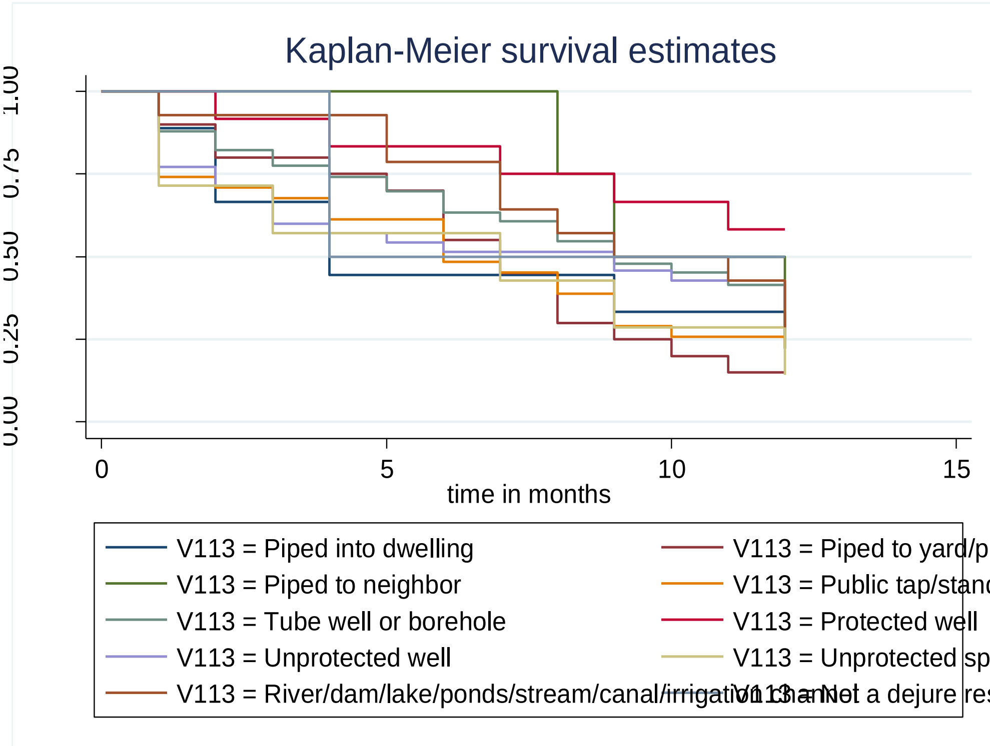

Figure 10: Survival estimates by source of drinking water

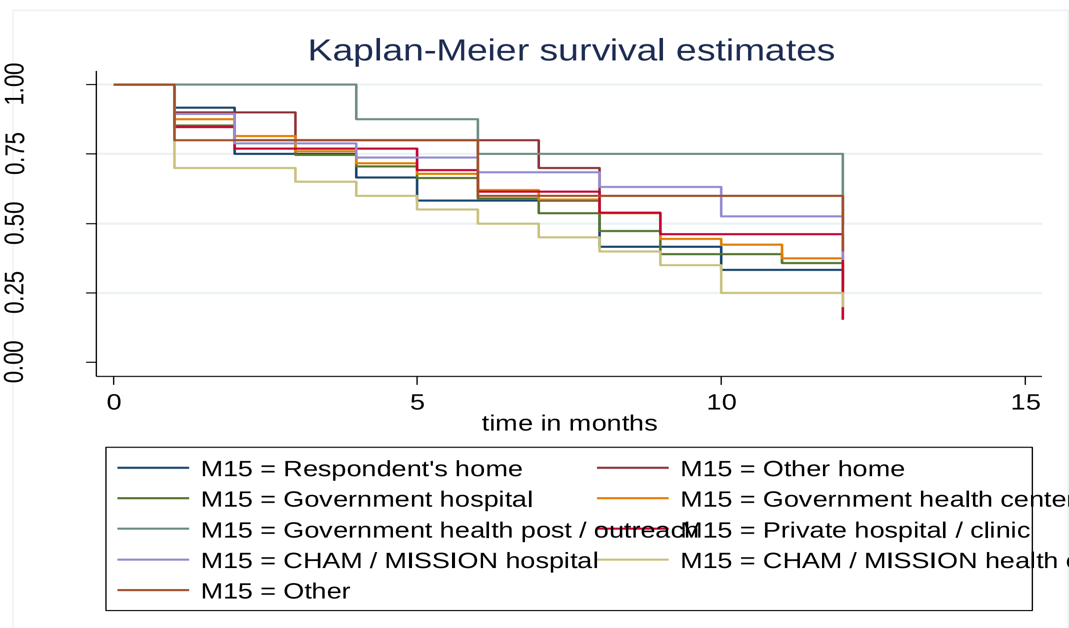

Figure 11: Survival estimates by place of delivery

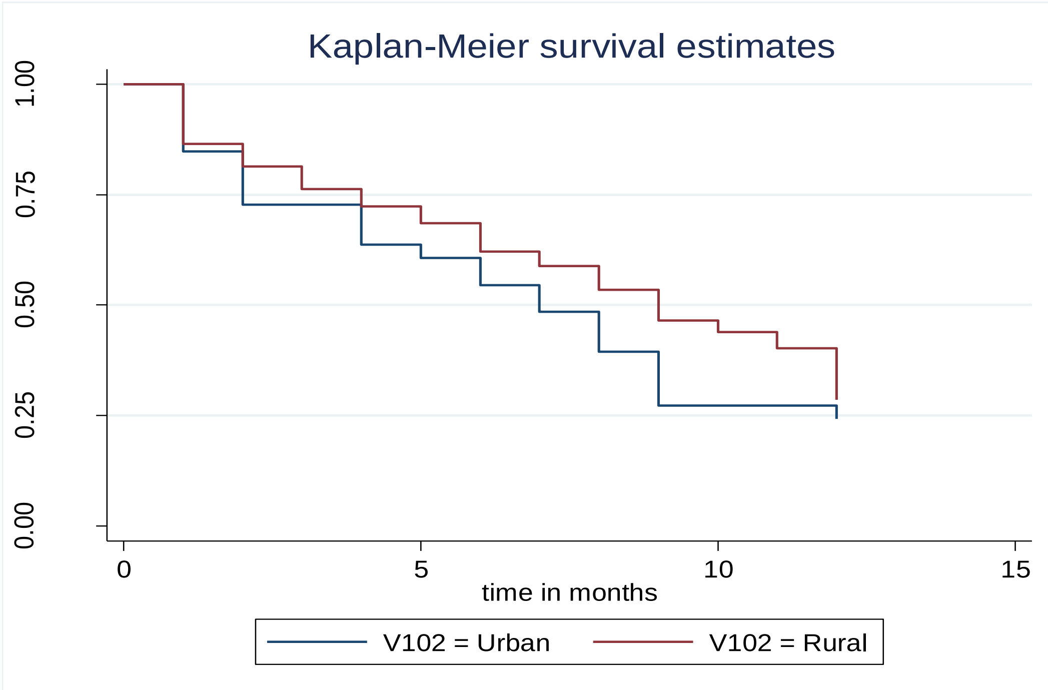

Figure 12: Survival estimates by area of residence

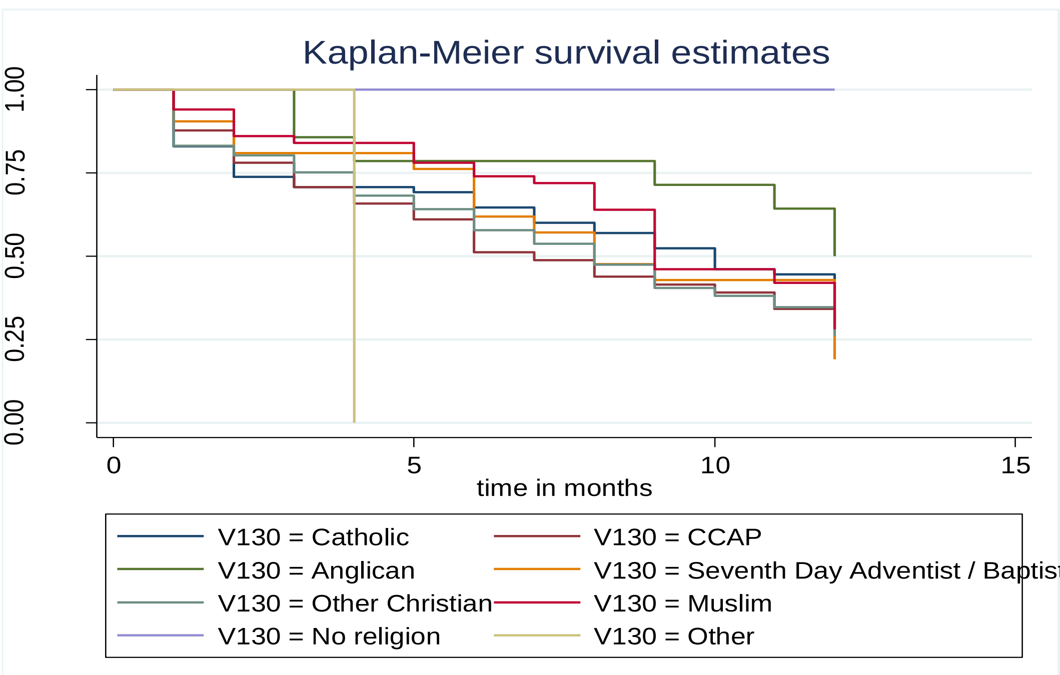

Figure 13: Survival estimates by religion

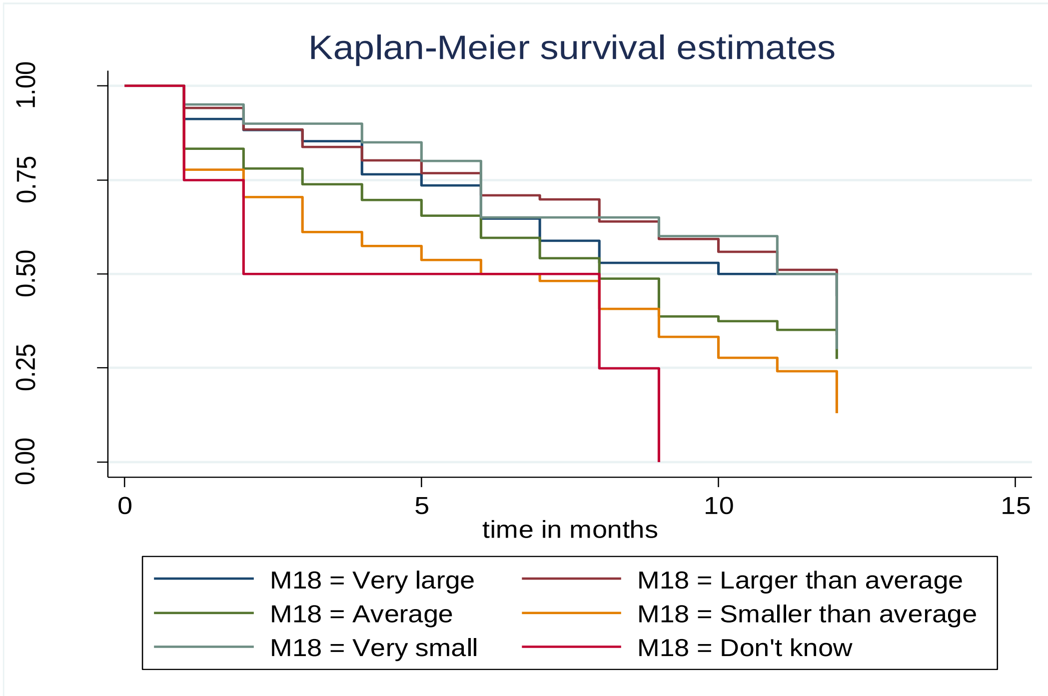

Figure 14: Survival estimates by size at birth

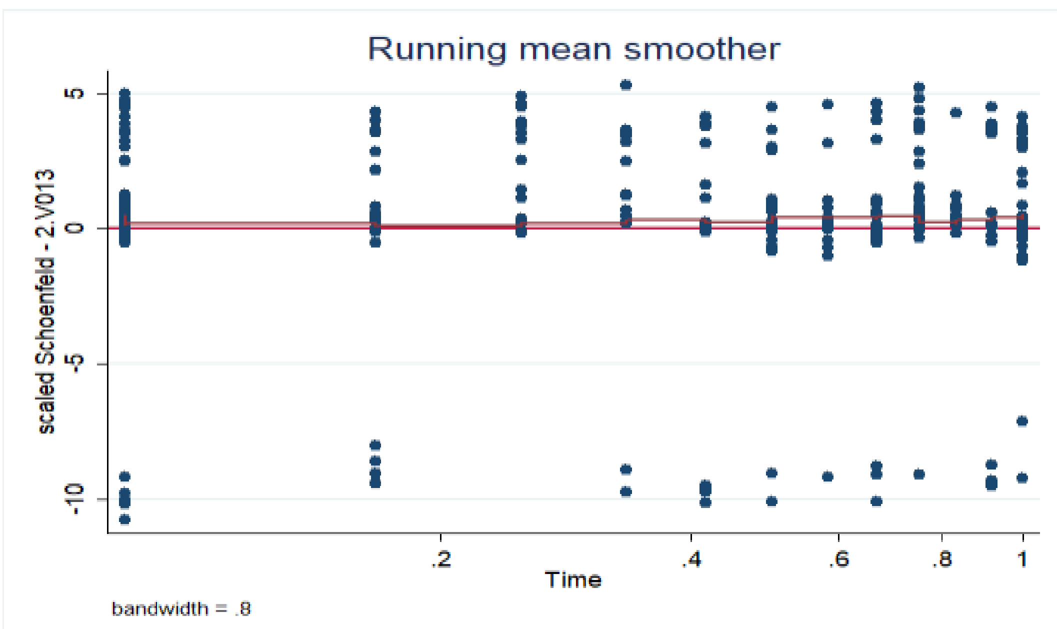

Figure 15: Residual plot for Mother's age group

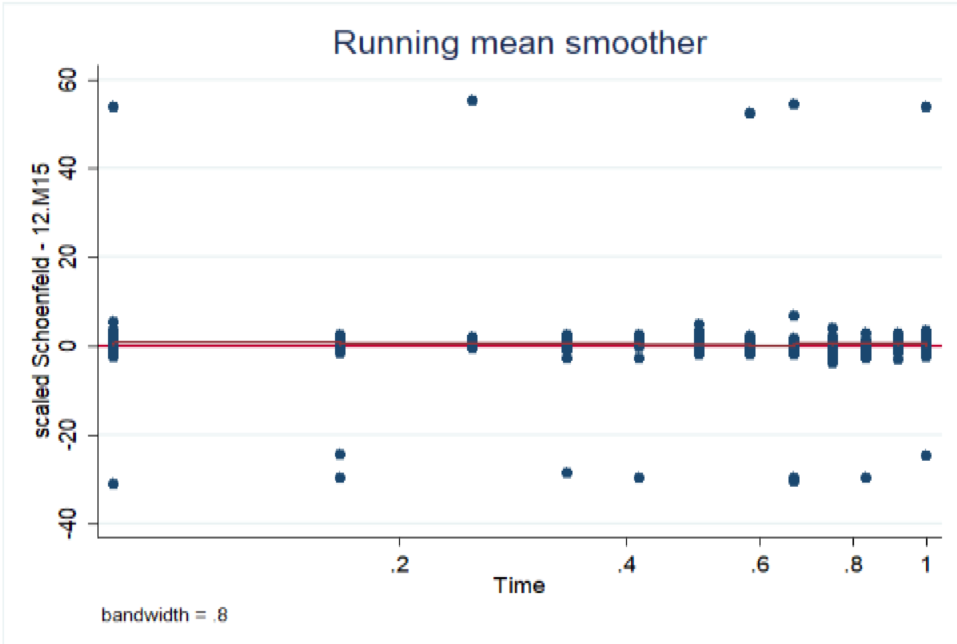

Figure 16: Residual plot for Place of Delivery

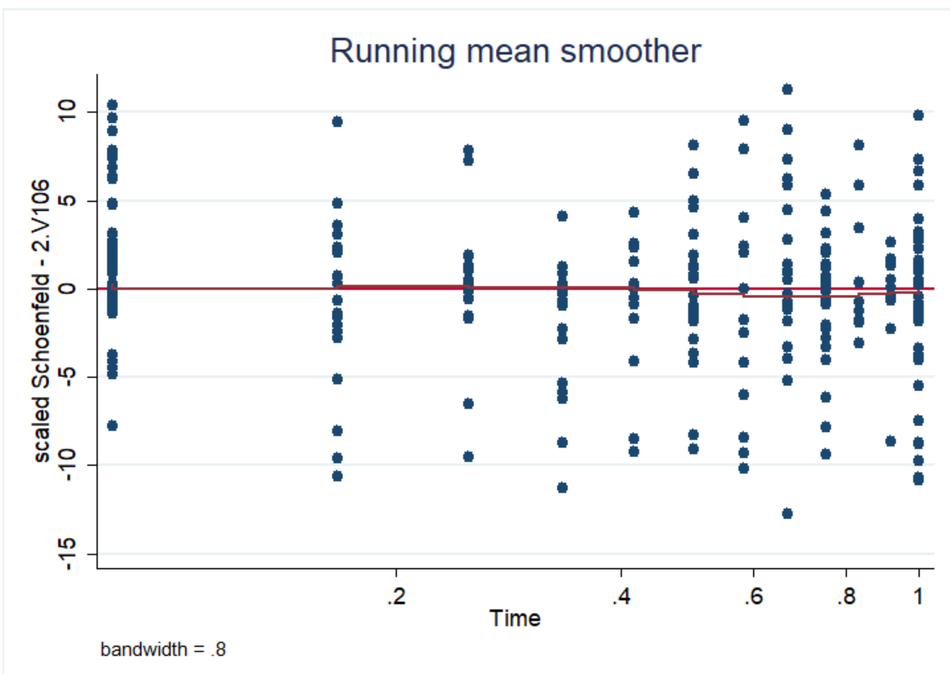

Figure 17: Residual Plot for Mother's Education level

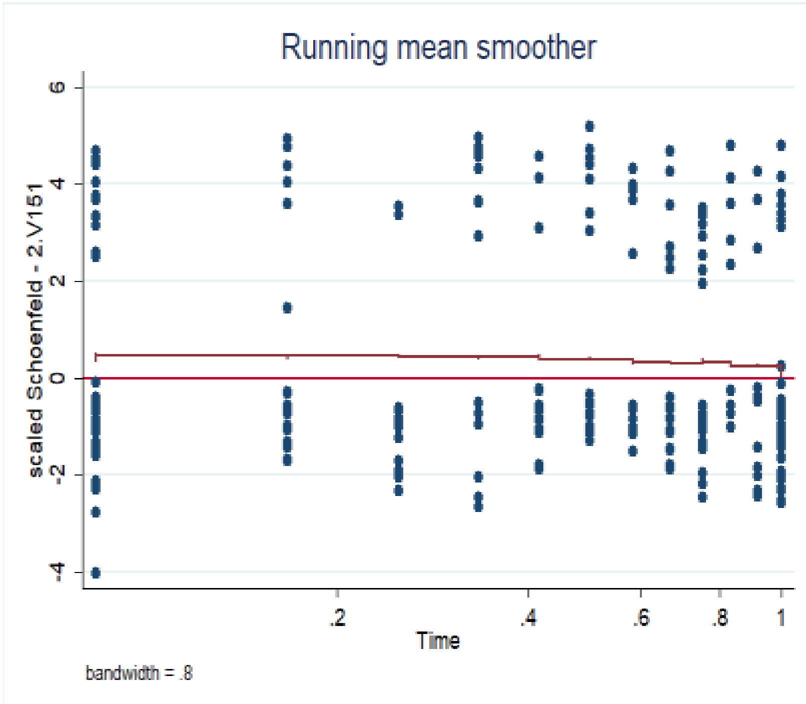

Figure 18: Residual plot for Sex of household head

APPENDIX C

Table 23: Log normal frailty model for household effects

Covariate	Coefficient	Std Error	z	P>z	[95% Conf. Interval]
B4					
Female	-.082	0.096	-0.86	0.391	(-0.272,0.106)
V190					
Poorer	-0.134	0.141	-0.95	0.341	(-0.411,0.142)
Middle	-0.190	0.149-1.28		0.202	(-0.483, 0.102)
Richer	0.130	0.164	0.80	0.427	(-0.191, 0.452)
Richest	0.145	0.220	0.66	0.511	(-0.288,0.578)
M15					
Other home	-0.404	0.460-0.88		0.380	(-1.308,0.498)
Government..	-0.098	0.281	-0.35	0.725	(-0.651, 0.453)
Government..	-0.257	0.272	-0.94	0.346	(-0.792,0.277)
Government..	-0.304	0.512	-0.59	0.553	(-1.307,0.699)
Other publ..	5.191	36029.33	0.00	1.000	(-70610.99, 70621.37)
Private ho..	-0.946	0.393-2.40		0.016	(-1.717, -0.175)
CHAM / MIS..	0.135	0.342	0.39	0.693	(-0.535,0.805)
CHAM / MIS..	-0.530	0.332	-1.60	0.110	(-1.183,0.121)
Other	-0.011	0.506	-0.02	0.981	(-1.005,0.981)
V130					
CCAP	0.106	0.180	0.59	0.557	(-0.248,0.460)
Anglican	0.566	0.314	1.80	0.071	(-0.049, 1.182)
Seventh Da..	0.088	0.227	0.39	0.699	(-0.358,0.534)

Other Chri..	0.045 0.137 0.33	0.742 (-0.223,0.313)
Muslim	0.200 0.179 1.12	0.264 (-0.151,0.552)
No religion	7.8633370.088 0.00	0.998 (-6597.388, 6613.115)
Other	-1.877 1.132 -1.66	0.097(-4.096,0.341)
B0		
1st of mul..	-0.518 0.318-1.63	0.104 (-1.144, 0.106)
2nd of mul..	-0.814 0.286 -2.84	0.005 (-1.376, -0.251)
V113		
Piped to y..	0.319 0.3720.86	0.390(-0.409.1.049)
Piped to n..	1.426 0.526 2.71	0.007 (0.395, 2.458)
Public tap..	0.812 0.389 2.08	0.037 (0.048, 1.575)
Tube well ..	0.995 0.388 2.56	0.010 (0.233, 1.757)
Protected ..	1.395 0.506 2.76	0.006 (0.403, 2.388)
Unprotecte..	0.891 0.421 2.12	0.034 (0.065, 1.716)
Protected ..	8.118 4764.516 0.00	0.999 (-9330.161, 9346.398)
Unprotecte..	0.835 0.510 1.64	0.102 (-0.164, 1.835)
River/dam/..	1.147 0.450 2.55	0.011 (0.264, 2.029)
Rainwater	6.860 84065.43 0.00	1.000 (-164758.4, 164772.1)
Other	7.847 6785.051 0.00	0.999 (-13290.61, 13306.3)
Not a deju..	1.883 0.970 1.94	0.052 (-0.019, 3.786)
V106		
Primary	-0.137 0.157 -0.88	0.380 (-0.445,0.169)

Secondary	-0.095	0.201	-0.48	0.634	(-0.490,0.299)
Higher	0.940	0.593	1.58	0.113	(-0.222, 2.104)
V151					
Female	-0.221	0.111	-1.99	0.046	(-0.440, -0.003)
V013					
20-24	-0.219	0.162	-1.36	0.175	(-0.537, 0.097)
25-29	-0.278	0.175	-1.59	0.112	(-0.621,0.064)
30-34	-0.184	0.185	-0.99	0.321	(-0.547,0.179)
35-39	-0.376	0.200	-1.88	0.060	(-0.770,0.016)
40-44	-0.896	0.250	-3.58	0.000	(-1.387, -0.405)
45-49	-1.421	0.326	-4.35	0.000	(-2.061, -0.781)
V102					
Rural	-0.370	0.208	-1.77	0.076	(-0.779,0.038)
M18					
Larger tha..	0.106	0.202	0.52	0.600	(-0.290, 0.503)
Average	-0.058	0.188	-0.31	0.754	(-0.427, 0.309)
Smaller th..	-0.504	0.215	-2.34	0.019	(-0.926, -0.082)
Very small	-0.027	0.290	-0.09	0.924	(-0.597, 0.542)
Don't know	-0.713	0.459	-1.55	0.120	(-1.612, 0.186)
_cons	1.660	0.548	3.03	0.002	(0.585, 2.736)

/ln_sig	0.354 0.047 7.49	0.000 (0.262, 0.447)
/ln_the	-2.035 0.827 -2.46	0.014 (-3.657, -0.412)
sigma	1.426 0.067	1.299 1.564
theta	0.130 0.108	0.0258 0.661
Likelihood-ratio	test of theta=0: chibar2(01) =1.80	Prob>=chibar2 = 0.090

Table 24: Log normal frailty model for community effects

_t	Coef. Std. Err z	P>z	[95% Conf. Interval]
B4			
Female	-.0804336 .0963455 -0.83	0.404	-.2692673 .1084001
V190			
Poorer	-.1338625 .140941 -0.95	0.342	-.4101018 .1423769
Middle	-.1978521 .1485672 -1.33	0.183	-.4890384 .0933342
Richer	.1255289 .1637349 0.77	0.443	-.1953856 .4464435
Richest	.1352733 .2196273 0.62	0.538	-.2951883 .5657349
M15			
Other home	-.3775234 .4577643 -0.82	0.410	-1.274725 .5196781
Government..	-.1002638 .2813505 -0.36	0.722	-.6517007 .4511732
Government..	-.2538095 .2724146 -0.93	0.351	-.7877323 .2801134
Government..	-.2956098 .5091125 -0.58	0.561	-1.293452 .7022324
Other publ..	7.310202 6246195 0.00	1.000	-1.22e+07 1.22e+07
Private ho..	-.9175255 .3900044 -2.35	0.019	-1.68192 -.1531311
CHAM / MIS..	.1264909 .3415179 0.37	0.711	-.5428718 .7958537
CHAM / MIS..	-.5240594 .3316949 -1.58	0.114	-1.17417 .1260506
Other	-.0268422 .5077041 -0.05	0.958	-1.021924 .9682396
V130			
CCAP	.1037142 .1805522 0.57	0.566	-.2501616 .4575901
Anglican	.5505175 .3135294 1.76	0.079	-.0639888 1.165024
Seventh Da..	.0737437 .2272455 0.32	0.746	-.3716492 .5191367
Other Chri..	.0373721 .13636 0.27	0.784	-.2298886 .3046328
Muslim	.1876383 .1788422 1.05	0.294	-.1628861 .5381626

No religion	10.27827 921345.3 0.00	1.000	-1805793 1805814
Other	-1.921634 1.132423 -1.70	0.090	-4.141142 .2978731
B0			
1st of mul..	-.5049252 .3164683 -1.60	0.111	-1.125192 .1153412
2nd of mul..	-.8030826 .2841468 -2.83	0.005	-1.36 -.2461651
V113			
Piped to y..	.324351 .3690399 0.88	0.379	-.3989538 1.047656
Piped to n..	1.4074 .5240106 2.69	0.007	.3803577 2.434441
Public tap..	.7804939 .3857908 2.02	0.043	.0243578 1.53663
Tube well ..	.9598354 .3840331 2.50	0.012	.2071443 1.712527
Protected ..	1.348098 .5037817 2.68	0.007	.3607036 2.335492
Unprotecte..	.8564605 .4161914 2.06	0.040	.0407403 1.672181
Protected ..	10.21098 653828.6 0.00	1.000	-1281470 1281491
Unprotecte..	.8114336 .5050576 1.61	0.108	-.178461 1.801328
River/dam/..	1.103921 .4461117 2.47	0.013	.2295579 1.978284
Rainwater	8.760713 9609659 0.00	1.000	-1.88e+07 1.88e+07
Other	10.10654 1506550 0.00	1.000	-2952774 2952794
Not a deju..	1.880139 .9698175 1.94	0.053	-.0206681 3.780947
V106			
Primary	-.1365481 .1571049 -0.87	0.385	-.4444681 .1713719
Secondary	-.0988768 .2014273 -0.49	0.624	-.4936671 .2959134
Higher	.9296382 .5929107 1.57	0.117	-.2324455 2.091722
V151			
Female	-.2279418 .1110751 -2.05	0.040	-.445645 -.0102386

V013			
20-24	-.2320593 .1618395 -1.43	0.152	-.549259 .0851403
25-29	-.281676 .1747884 -1.61	0.107	-.6242551 .060903
30-34	-.1919769 .1852599 -1.04	0.300	-.5550797 .1711259
35-39	-.3763069 .2003535 -1.88	0.060	-.7689927 .0163788
40-44	-.9216202 .2488654 -3.70	0.000	-1.409387 -.4338529
45-49	-1.431121 .3225633 -4.44	0.000	-2.063334 -.798909
V102			
Rural	-.3599173 .2078115 -1.73	0.083	-.7672204 .0473858
M18			
Larger tha..	.0990319 .2024025 0.49	0.625	-.2976698 .4957336
Average	-.0570779 .1877803 -0.30	0.761	-.4251204 .3109647
Smaller th..	-.4919836 .2145574 -2.29	0.022	-.9125085 -.0714588
Very small	-.0403739 .2903094 -0.14	0.889	-.6093698 .528622
Don't know	-.7720519 .4597998 -1.68	0.093	-1.673243 .1291393
_cons	1.712176 .5450584 3.14	0.002	.643881 2.780471
/ln_sig	.3601193 .0471199 7.64	0.000	.267766 .4524726
/ln_the	-15.37195 557.3202 -0.03	0.978	-1107.699 1076.956
sigma	1.4335 .0675464	1.307041 1.572195	
theta	2.11e-07 .0001175		
Likelihood-ratio	test of theta=0: chibar2(01) =	0.00	Prob>=chibar2 = 1.000