UNIVERSITY OF KWAZULU NATAL

Statistical Models to Understand Factors Associated with Under-five Child Mortality in Tanzania

By

Welcome Jabulani Dlamini

A Thesis Submitted in Fulfilment of

The Requirements of The Degree for

MASTER OF SCIENCE

in

STATISTICS

under the supervision of

Dr SF Melesse

and

Prof HG Mwambi

School of Mathematics, Statistics and Computer Science

Pietermaritzburg Campus

July 13, 2016

Declaration

I, Welcome Dlamini, declare that this thesis titled, 'Statistical Models to Understand Factors Associated with Under-five Child Mortality in Tanzania' and the work presented is my own. I confirm that:

- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

I hereby confirm that all passages which are literally or in general matter taken out of publications or other sources are marked as such.

Mr. Welcome J. Dlamini — Date

Dr. Sileshi F. Melesse — Date

Prof. Henry G. Mwambi — Date

Acknowledgements

First and foremost, praises and thanks to the God, the Almighty, for His showers of blessings throughout my research work to complete the research successfully.

The research included in this thesis could not have been performed if not for the assistance, patience, and support of many individuals. I would like to extend my gratitude first and foremost to my Supervisor Dr. Sileshi Melesse for mentoring me over the course of my graduate studies. He has helped me through extremely difficult times over the course of the analysis and the writing of the Thesis and for that I sincerely thank him for his confidence in me. I would additionally like to thank Prof. Henry Mwambi for his support in both the research and especially the helpful advises that has led to this document. His knowledge and understanding of research work has allowed me to fully express the concepts behind this research. They have taught me the methodology to carry out the research and to present the research works as clearly as possible. It was a great privilege and honor to work and study under their guidance. I am extremely grateful for what they have offered me. I would also like to thank both of them for their friendship, empathy, and great sense of humor.

Finally I would like to extend my deepest gratitude to my family without their love, support and understanding I could never have completed this Master's degree. This research was supported financially by the South African Centre for Epidemiological Modelling and Analysis (SACEMA). SACEMA is known as a national research centre established under the Centre of Excellence programme of the Department of Science and Technology and the National Research Foundation.

Abstract

The risk or probability of dying between birth and five years of age expressed per 1000 live births is known as Under-five mortality. The well-being of a child reflects household, community and national involvement on family health. This will have an immense future contribution towards the development of a country. Globally, a substantial progress in improving child survival since 1990 has been made. The decline globally in under-five mortality from approximately 12.7 million in 1990 to approximately 6.3 million in 2013 had been observed. Notably, all regions except Sub-Saharan Africa, Central Asia, Southern Asia and Oceania had reduced the rate by 52% or more in 2013. This study aims to identify factors that are associated with the under-five mortality in Tanzania. In order to robustly identify these factors, the study utilized different statistical models that accommodate a response which is dichotomous. Models studied include Logistic Regression (LR), Survey Logistic Regression (SLR), Generalized Linear Mixed Model (GLMM) and Generalized Additive Model (GAM). The result revealed that HIV status of the mother is associated with the under-five mortality. Furthermore, the results revealed that childbirth order number, breastfeeding and a total number of children alive affects the survival status of the child. The study shows that there is a need to intensify child health interventions to reduce the under-five mortality rate even more and to be in line with the millennium development goal 4(MDG4).

Keywords: Survey Logistic Regression, Generalized Linear Mixed Models(GLMMs), Generalized Additive Models (GAMs) and Cubic spline smoother.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Tanzania is a relatively large country in the East African region sharing borders with Kenya, Uganda, Rwanda, Zambia, Malawi, Mozambique, Burundi and the Democratic Republic of Congo (DRC). This nation now is considered one of the oldest known (continuously inhabited) areas on the planet. Tanzania is also bordered by the Indian Ocean on the east. North-East of Tanzania is mountainous with the famous mountains including Meru and Kilimanjaro, the highest peak in Africa. Mount Kilimanjaro is covered with snow even though it is so close to the equator thus, its natural beauty attracts thousands of tourists each year. This mountain stands at 5,895 m tall. The mainland in Tanzania dominated by a large central plateau, one covered with grasslands, plains, and rolling hills. Figure 1.1 shows the location of Tanzania in Africa with several tourism attractions such as Africa's largest lakes, including Lake Nyasa (Lake Malawi), Lake Victoria (Africa's largest lake), and Lake Tanganyika (Africa's deepest lake)(Dagne, 2011).

The name Tanzania itself derives from the country's two states, Zanzibar, and Tanganyika. Zanzibar is an archipelago off the coast of Tanzania and a semi-autonomous part of the country (UNICEF, 2012). With 947,300 square kilometers of land, Tanzania is the 31th largest country in the world and the 14th largest in Africa. The last official census recording the population of Tanzania occurred in 2012 and showed that there were 49,639,138 people living in the country. However, currently, there are 50,757,459 living in Tanzania and of this total population approximately 1.3 million reside on the islands of Zanzibar (TACAIDS and NBS OCGS, 2013; Agwanda and Amani, 2014). This equates

(a) Location of Tanzania in Africa

(b) Tanzania regions

Figure 1.1: Maps showing Tanzania regions and location in Africa

Source:`http://www.worldatlas.com/webimage/countrys/africa/tz.htm`

to a population density of about 47.5 people per square kilometer. Figure 1.2 shows the most populated cities in Tanzania.

Tanzania is one of the country with the highest birth rates in the world and more than 44% of the population is under the age of 15. The total fertility rate is 5.2 children born per woman. Tanzania has the 18th highest population growth rate in the world and birth rate and does not show any signs of changing soon. Tanzania's population growth rate continues to climb with a current rate of around 3.0% annually. If this trend continues, it is projected that Tanzania will have a population of 138 million by 2050, making it the 13th most populated country by then compared to its current rank of 26th. Over the past decade, studies show that more than 30% of households are headed by females in Tanzania (Factbook, 2015).

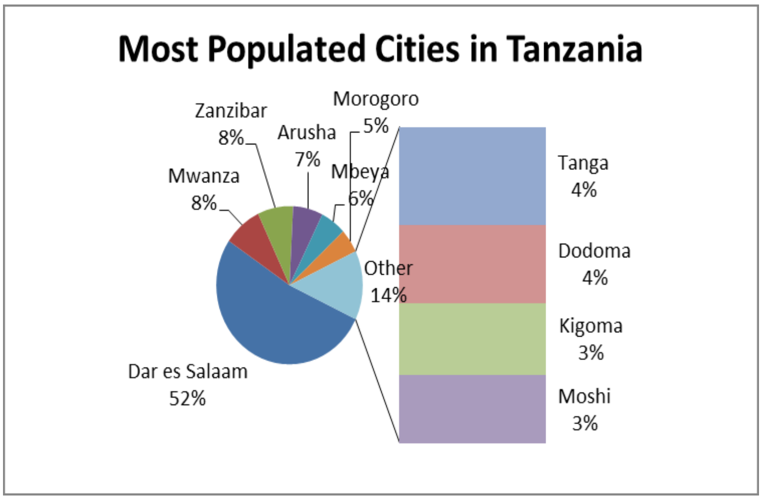

Figure 1.2: Most populated cities in Tanzania

Source of data:`http://www.worldatlas.com/webimage/countrys/africa/tz.htm`

The 2012 population house census (PHC) showed that in Tanzania 37% of private households had access to piped water as the main source of drinking water with urban areas having the majority of household using piped water (TACAIDS and NBS OCGS, 2013). The low median age of Tanzania is attributed to a generalized human immunodeficiency virus (HIV) epidemic in the country. It's estimated that there are over 1.6 million Tanzanians currently living with HIV/AIDS and the epidemic has resulted in an estimated 1.3 million orphans. The overall HIV prevalence rate in Tanzania is 5.1% although this reaches as high as 15.4% among women in some areas. This epidemic may result into; lower life expectancy, a higher infant mortality rate, under-five mortality rate, higher death rate, changes in age and sex distribution in the population as well as a lower population growth.

The under-five mortality rate is an important indicator of child well-being, including health and nutritional status of the child. It can also be an indicator of the coverage of social and economic development and child survival interventions. Tanzania is one of the countries in the Sub-Saharan Africa region with high under-five mortality rate (Factbook, 2015). Prior to 2011-2012 Tanzania HIV/AIDS and Malaria Indicator Survey (THMIS) the under-five mortality rate was 76 per 1000 live births. Furthermore, the same survey reveals that there was a higher number of deaths among males compared to females and in rural areas compared to urban areas. It also reveals that Tanzania's infant mortality has declined from 115 deaths per 1,000 live births in 1988 to 45 deaths per 1,000 live births in 2012 (UNICEF, 2012). These statistics suggest that Tanzania could achieve the Millennium Development Goal 4 (MDG4) which is to reduce under-five mortality rate by two-thirds by 2015 with intensified child health interventions in both rural and urban areas (UNICEF, 2012).

1.2 Health Issues and Health Sector Budget

The health of the Tanzanian population is generally poor and the underlying cause of poor health includes sanitation and under-nutrition. As part of the mitigating strategies, family planning is required to lower the birth rate and control the population growth (Kwesigabo et al., 2012). Despite existing control measures, the population has continued to witness massive growth. The number of deaths of children under-five years is the results of poor hygiene and about 20 percent of the child mortality are the results of preventable health issue such as diarrhea. It is known that sanitation has an impact in reducing diarrhea but access to drinking water and having better toilet facilities is still a challenge in Tanzania (Kwesigabo et al., 2012). Lack of water supply and sanitation facilities plays a huge role in contributing to poor school attendance which affects educational performance than early dropouts.

There is a link between health and agriculture and it's well known that food security at the household level is important to good health. Furthermore, agriculture is the source of livelihood among the poor and could be directly or indirectly linked to poor health including the incidence of malaria and livestock diseases (Hawkes and Ruel, 2006). The health of individuals also affects agriculture indirectly since people's health status affects efforts for agricultural production. The wealth index has an impact on how people will perform at work which may affect income and production (Hawkes and Ruel, 2006). In 2001, the United Republic of Tanzania signed the Abuja declaration which was to allocate 15% of the government budget to the health sector. However, Tanzania has only managed to allocate 8.9% of its total budget (Kwesigabo et al., 2012). The budget has to be met by using domestic resource since donors have shown not to be sustainable. It is also well known that health is one of the basic services that government has to provide (TACAIDS and NBS OCGS, 2013).

1.3 Literature Review

The probability per 1000 live births that newborn child will die before reaching age five is known as under-five mortality rate. Most deaths of children under the age five are as a result of nutritional conditions which may lead to a weak immune system. These occur within the first year of life. A child's risk of dying is higher in the first week of life thus safe childbirth and effective early nutritional care are essential in order to prevent such deaths. More than half the number of child death under the age of five is due to conditions that could be prevented or treated with an access to affordable interventions. The leading causes of the under-five mortality include malaria, pneumonia, diarrhea and birth complications (UNICEF, 2012). This section intends to review studies which have been done and are related to this study.

The world has made a substantial progress towards achieving MDG4 as the under-five deaths had declined from 12.6 million on average in 1990 to 6.6 million on averages in 2012 worldwide (UNICEF, 2012). All regions in the world except for Sub-Saharan Africa and Oceania countries had reduced their under-five mortality by more than 49 percent in 2012. The average annual rate of reduction in under-five mortality had increased from 1.2% a year over 1990-1995 to 3.9% over 2005 to 2012. However, this remains insufficient to reach MDG4, mainly in Sub-Saharan Africa, Oceania, Central Asia and Southern Asia (UNICEF, 2012). Sub-Saharan Africa still has the highest rates of child mortality with an under-five mortality rate of 98 deaths per 1000 live births which are high compared to developed regions (Manda, 1999; Susuman, 2012; UNICEF, 2012).

There are many studies which have been done concerning under-five mortality rate, yet not many have considered HIV/AIDS as one of the risk factors for under-five mortality. Lemani (2013) considered HIV/AIDS as a risk factor of under-five mortality rate. Lemani (2013) found that there is a significant relationship between infant and child mortality. The statistical method mostly used known as logistic regression was adopted in this study with other modeling techniques such as survival analysis. The assumption made for logistic regression model is that data was obtained from the finite population using simple random sampling. However, since that used is from a complex survey with clustering the method was weighted to account for correlated data. The study noted that HIV/AIDS status of a mother has an impact on child survival as a result of the child being infected with HIV. The child whose mother was HIV-positive has an increased risk of death than a child whose mother was HIV-negative. The study by Lemani (2013) focused on the impact of a mother's HIV status on the under-five mortality rate in Malawi. The HIV status of the mother was found to be associated with increased risk of under-five mortality rate after controlling other factors. The MDG4 in Malawi, that is reducing childhood death is likely to be achieved. Provided that the development and planning policies consider improving factors that enhance education, health provision among others (Lemani, 2013).

Coovadia et al. (2007) study had an objective of reviewing the available data related to child mortality in Africa and its association with the HIV infections status of a mother and child. In this study, it was shown that survival of the child is indeed influenced by the HIV epidemic in different ways, such as mother to child breastfeeding. The study further revealed that child mortality is closely linked to the maternal health status. The mortality rate of HIV-negative children of HIV-positive mothers was 166 per 1000 live births, but the mortality rate of HIV-negative children for the HIV-negative mother was 128 per 1000 live births. Nevertheless, there is a need for studies on control strategies.

This will give improved overall effect of the HIV epidemic on child mortality.

Ettarh and Kimani (2012) investigated the influence of geographical location and maternal factors on the likelihood of mortality. The Multivariate analysis was used to compare the risk factors in urban and rural areas in Kenya. Kenya is also found in East Africa within the Sub-Saharan Africa region in which there are still some concerns with regard to achieving the Millennium Development Goal 4. The study was based on the national cross-sectional demographic and Health survey from 2008-2009. In this study, deaths among the under-five children were found to be more frequent in rural areas for mothers age 21 years compared to mothers of the same age in urban areas. One of the factors found to be significant was wealth index of the household, where in rural areas household with greater wealth were less likely to experience under-five deaths compared to the poor household. This study also focused on understanding of the drivers of the under-five mortality in rural and urban areas. The under-five mortality is associated with young mothers, poor households, inadequate breastfeeding and is limited to certain specific geographic areas. Some studies have shown that for developing countries, maternal education and age of the mother are important determinants of the under-five mortality. The mortality rates are higher among less educated mothers compared to mothers with higher education level hence the maternal education is important because it increases a mother's knowledge and skills. This leads to effective understanding and using the available information and resources for the survival of the child (Ettarh and Kimani, 2012).

Among other studies which considered diarrhea as one of the risk factor for under-five mortality, Sukaina (2009); Walker et al. (2012) studied the incidence in low-income and middle-income countries between 1990 and 2010. Despite diarrhea showing a slight decline, additional efforts were required to improve both prevention and treatment. This study suggested that diarrhea is one of the leading causes of mortality among children under the age of five around in the world. A study by Masanja et al. (2008) also revealed the same findings. The possible causes of diarrhea include; inadequate water, and sanitation and nutritional risk factors. Masanja et al. (2008) had an objective of identifying what might have contributed to the reductions in mortality yearly, and further investigate prospect for meeting the Millennium Development Goal 4(MDG4) target by 2015. This study used different demographic and health survey (DHS) done in Tanzania since 1990. The analysis was done for each data set to generate estimates of mortality in children younger than 5 years old (Masanja et al., 2008). The estimates for trends in mortality between 1990 and 2004 were fitted using regression models instead of forecasting which was done for 2005 to 2015. The main aim of this was to investigate if Tanzania health system could affect child mortality or not. In 2000-2004, an accelerated reduction in mortality was observed with point estimates of 141.5 (95% CI: 141.5-141.5) death rate per 1000 live births. During this period, there was a 40 percent reduction to reach a point estimates 83.2 (95% CI: 70.1-96.3) death per 1000 live births. During this period, an improvement in health systems and an increase in child survival interventions, that is to say, good management of childhood illness and an insecticide-treated net was observed. There was no change in other determinants of child survival, except for the slow increase in the HIV/AIDS. Tanzania could well achieve the Millennium Development Goal 4 (MDG4) provided such trends of improved child survival are to be sustained going forward (Masanja et al., 2008). This study aims at identifying factors associated with the under-five mortality in Tanzania by utilizing different statistical models. The adopted model in identifying risk factors for under-five child mortality. However, we may not always assume that data was obtained from a finite population using simple random sampling. If data was obtained from the finite population using stratified, in order to make valid statistical inference we may have to account for survey design features by considering survey logistic regression. They may also be a problem of correlation between observation that we need to account for by considering the generalized linear mixed model. The linearity assumption is made when using generalized linear model and generalized linear mixed model. The alternative model that can be used is the generalized additive model that does not make linearity assumption between outcome and predictors. We first describe the data and outline methods to be used in this study.

1.4 Data and Methods

1.4.1 Description of the Data

This study uses part of the data from the Tanzania demographic healthy survey of the period 2011-2012 as part of HIV/AIDS and Malaria Indicator Survey. This was the third population-based survey of this nature conducted in Tanzania. The objective of 2011-2012 THMIS was to provide up to date information on key indicators needed to keep track of progress in Tanzania health program including knowledge, attitude and behaviors relating to HIV/AIDS, plus other sexual transmitted disease and malaria. THMIS also provides data on the prevalence of anemia, the prevalence of malaria among children 6-59 months, and prevalence of HIV among the general population for men and women between the ages 15 to 49 (Commission, 2013; TACAIDS and NBS OCGS, 2013). THMIS data which was obtained on request on `http://www.dhsprogram.com` was considered in this study. THMIS sample was selected using stratified, two-stage cluster design. In stage 1 a total of 563 clusters were selected (clusters consisted of enumeration areas). In stage 2 approximately 18 households were selected from each cluster which yielded a sample size of 10496.

1.4.2 Methods

To summarize the main characteristics of the data, exploratory data analysis (EDA) will be carried out. The formalization of relations between the outcome and the other variables will be done by utilizing modeling techniques in increasing complexity. This are:

- Logistic regression models without and with design effects,
- Generalized Linear Mixed Models and
- Generalized Additive Models

Statistical Package for the Social Science (SPSS) version 21 and Statistical Analysis System (SAS) 9.3 are used to fit these statistical models . Thereafter these results are discussed and interpreted.

1.5 Problem Statement and Justification of the Study

Tanzania has been undergoing the unpredicted decline in mortality particularly among children under-five years of age. The decline also includes mortality among old age groups, notably among adults in the most productive years. With the mortality rate among children under-five years of age declining, the MDG4 will be achieved given that the country intensities child health interventions. This unpredicted dramatic mortality decline could results in accelerated population growth unless birth rate also declines. Urban areas of Tanzania has shown signs of fertility decline. Nevertheless, it is not the case with rural areas. The inertia of demographic change will lead to the significant change in age-structure of the population. When the children reach reproductive age, then the population will increase further even though fertility declines. An increase in the survival of adults will result in fast-growing aged population possibly leading to implications of raising the importance of chronic diseases and demands on the health system. In this study ordinary logistic regression and other methods are used to achieve the objectives of the study.

Ordinary Least square regression models and Logistic linear regression models both assumes a linear form of predictor variables to the response. The right-hand side of the generalized linear models (GLMs) has a linear relationship with left-hand side. It is known that logistic regression falls under the generalized linear model. It's possible to have predictor variables that have a non-linear relationship with the response, so in this case, logistic regression could give unrealistic results. The alternative to the linear predictor method is the Generalized Additive Models (GAM). Where the linear form $\sum_{j=1}^{n} \beta_j X_j$ is replaced by the general smooth function $\sum_{j=1}^{n} S_j(X_j)$ in which the smooth functions are unspecified and can be estimated. We can estimate the smooth function using non-parametric approach such as loess (Hastie and Tibshirani, 1986).

1.6 Objectives of the Study

The children are the future and economic assets of the world. Their future development may be affected by the factors associated with the under-five mortality. Furthermore, child well-being reflects household, community and national involvement in the family health in Tanzania. This contributes both directly and indirectly in a country's development. The main objective of this study is to use a series of statistical methods to determine and understand factors that significantly affect under-five mortality in Tanzania. The findings from this study can be used to evaluate the progress Tanzania has made towards achieving the Millennium Development Goal 4 programs and develop new health strategies based on the findings of this study. The identified factors may be used to guide policy and decision making to speed up the provision of a better life for all.

1.7 Outline of the Study

This study is organized into six chapters. Chapter 2 consists of exploratory data analysis carried out using SPSS. A chapter 3 introduces the Generalized Linear Models (GLMs), describes the logistic regression, and provides the statistical model to be used. Chapter 3 also introduces the Survey logistic regression and the model will be fitted. The Generalized Linear Mixed Model (GLMM) is discussed and its application using part of 2011-2012 THMIS data in Chapter 4. In Chapter 5 the Generalized Additive Model and application using part of 2011-2012 THMIS data is introduced. Chapter 6 discusses the findings, presents recommendations, and conclusion.

Chapter 2

Exploratory Data Analysis

2.1 Introduction

The purpose of exploratory data analysis (EDA) is to help understand the data in detail before the modeling and inference tasks. In this section, a detailed and extensive exploratory data analysis is presented. The EDA focuses on the following:

- Assessing assumptions on which inference will be based,
- Determine association between the outcome variable and predictor variables,
- Providing a basis for further data collection through survey.

The simple descriptive statistics such as frequency distributions and percentages are computed to describe some of the variables and to check the variables that have missing values. We first describe variables from the data set of interest and then present results performed.

2.2 Study Variables

2.2.1 Dependent Variable

The response variable in this study is survival status of a child which is a dichotomous variable showing the status: of a child alive or not. The response variable is coded as "1" if the child is not alive and "0" if the child is alive at the time of the survey.

2.2.2 Independent Variables

The survey captured a vast range of variables. However, this study considers only 16 variables including HIV status of the respondent which were selected based on current literature. The lists of the explanatory or predictor variables in this study are indicated in Table 2.1:

Table 2.1: Description of predictor variables in the study.

	Variables	**Explanation**
	Socio-demographic characteristics	
1	Sex of child	male (1), female(2)
2	Mother's age	<20 years (0), 20-34 years (1), >34 years(2)
3	Birth order number	first birth (1), 2-4 births(2), > 4 births (3)
4	Current breastfeeding	Yes(1), No(0)
5	Current marital status	married(1), not married(0)
6	Age of the household head	less than 21 years (0), 21-34 years(1),>34 years(2)
7	Mother's HIV status	HIV negative(0), HIV positive(1)
	Socio-economic characteristics	
8	Type of place of residence	urban(1), rural (2)
9	Wealth index	Poor(0), Middle(1), Rich(2)
10	Number of living children	<2 children(0),2-4 children(1), >4 children(2)
11	Number of children ever born	< 2 children(1), 2-4 children(2), >4 children (3)
12	Number of children 5 years or under	less than 2 (1), 2-4 (2), more than 4 (3)
13	Respondent level of education	no education(0), primary (1), secondary and higher(2)
14	Mother currently working	No(0), Yes(1)
	Household environment characteristics	
15	Source of drinking water	safe water (1) and not safe water (0)
16	Main floor	unfinished(0) and finished(1)

2.3 Preliminary Analysis

The purpose of this study is to determine some of the risk factors for under-five child mortality in Tanzania. To perform this analysis, the baseline characteristics of the individuals need to be further explored, specifically the mother's working status, birth order number, age, mother's education level, type of place of residence, marital status and wealth index. These variables are categorical variables and we will now look at the analysis of the frequency tables which were obtained. The results in Table 2.2 show that the sample consisted of 67.3% (n=7416) of respondents aged from 20-34 years. The respondents with age less than 20 and above 34 years accounted for 4.5% and 28.1% of the sample respectively. Approximately two-thirds of the sample respondents were younger individuals aged 15-34. The result shows that child sex was almost equally distributed with males accounted for 50.3%, and females accounted for 49.7% of the sample thus the 1:1 sex ratio is closely exhibited in the sample. We also observe that more respondents were currently breastfeeding and accounted for 55.8% (n=6084) of the sample, and that the sample had 76% (n=8374) of the respondents that were married. Table 2.2 also shows that 38.2% respondents had less than two births and 34.6 % had more than 4 births. About 4.2% of the respondents were HIV-positive while about 4.9% were missing values.

Table 2.2: Socio-demographic characteristics distribution of the respondents.

Covariates	**Characteristics**	**Frequency**	**Percent (%)**
Sex of child	Male	5540	50.3
	Female	5473	49.7
Respondent age	Less than 20 years	498	4.5
	20 - 34 years	7416	67.3
	Over 34 years	3099	28.1
Birth order	Less than 2 births	4203	38.2
	2 - 4 births	2994	27.2
	Above 4 births	3816	34.6
Current breastfeeding	No	4929	44.8
	Yes	6084	55.8
Current marital status	Never in union	496	4.5
	Married	8374	76
	Living with partner	1022	9.3
	Divorced	624	2.3
	Widowed	258	5.7
	No longer living with partner	239	2.2
Household head age	Less than 20	33	0.3
	20-34 years old	3388	30.8
	Above 34 years	7592	68.9
Mother's HIV status	HIV negative	10007	90.9
	HIV positive	467	4.2
	missing	539	4.9
Total		11013	100

Table 2.3 displays Socio-economic characteristic distribution. From this table, we observed that individuals from rural areas were about 84.5%. The urban areas were less represented with only 15.5% of the sample and there were no missing values for this variable. The percentages in the rich and poor categories were 35.5% and 43.3% respectively. Most of the respondents were currently working on which they were 88%. There were 64.1% (n=7061) respondents with primary education and those with no education were 25.1% (n=2768). The variable for those currently working had nine missing value. The

percentage was 71.4% (n=7862) of respondents with less than 2 children under the age five in a household. Nearly 15.5% of the total sample was made up of individuals from the urban area. Table 2.4 displays household environment characteristic distribution. From

Table 2.3: Socio-economic characteristic distribution of the respondent in Tanzania.

Covariates	**Characteristics**	**Frequency**	**Percent (%)**
Type of place of residence	Urban	1707	15.5
	Rural	9306	84.5
Wealth index	Poor	4768	43.3
	Middle	2336	21.2
	Rich	3909	35.5
Number of living children	Less than 2 children	1457	12.9
	2 - 4 children	5613	51
	Above 4 children	3943	36.1
Number of children ever born	Below 2 children	3101	28.2
	2 - 4 children	3295	29.9
	Above 4 children	4617	41.9
Number of children 5 or under	Below 2 children	7862	71.4
	2 - 4 children	2574	23.1
	Above 4 children	577	5.2
Respondent level of education	No education	2768	25.1
	Primary education	7061	64.1
	Secondary or higher	1184	10.8
Mother currently working	No	1315	11.9
	Yes	9689	88
	missing	9	0.1
Total		11013	100

the table, 51.4% of the respondents reported that they had access to safe water. Most of the respondent 73.9% reported that their houses consisted of unfinished floor. This variable had three values that were missing.

Table 2.4: Household environment characteristic distribution of the respondents.

Covariates	**Characteristics**	**Frequency**	**Percent (%)**
Source of drinking water	Safe water	5665	51.4
	Not safe water	5348	48.6
Main floor	Unfinished	8133	73.9
	Finished	2877	26.1
	missing	3	0
Total		11013	100

2.4 Chi-Square Test of Association

It is important to find out if there is an association between the response variable and the categorical predictor variables with the use of a cross-tabulation techniques. From

Table 2.5, Table 2.7 and Table 2.6, we deduce that the variables with p-values less than 5% level of significant were significantly associated with the response variable. Table 2.5, Table 2.7 and Table 2.6 shows the proportion of each category of the covariates and results of chi-square of association. The proportion 47.29% of child dying is higher for respondents with first birth compared to the respondents with two or more births. The proportion 73.2% of children dying was higher for respondents aged from 20 to 34. We observed that mothers from rural areas had a higher proportion of children dying than the mother in the urban area. The child from a mother who does not breastfeed had a higher chance of dying and the child from a mother with less than two children under the age five in a household had a higher proportion 85.59% of dying. The child from a mother who was currently working had the higher proportion 85.14% than the child born from a mother who was not working. In summary we can say that Table 2.5, Table 2.7 and Table 2.6 shows that birth order, child sex, place of residence, marital status, wealth index, respondent's age, number of child ever born, number of children five or less, respondent education level, HIV status of the mother, age of household head, and number of living children are all univariately significantly associated with child survival status.

Table 2.5: Bivariate analysis of associations with under-five mortality scio-demographic characteristics.

Socio-demographic characteristics					
Covariate	**Sample Size**	**DF**	**proportion**	**Chi-square**	**p-value**
Birth order	11013	2		19.426	0.0001
First births			0.4729		
2 - 4 births			0.2635		
More than 4 births			0.2635		
HIV status of the mother	10474	1		23.933	0.0001
HIV negative			0.9071		
HIV positive			0.0929		
Sex of child	11013	1		2.299	0.129
Male			0.5383		
Female			0.4617		
Age of House hold head	11013	2		6.685	0.02
Less than 21 years			0.0135		
21-34 years			0.3423		
More than 34 years			0.6441		
Current breast feeding	11013	1		109.247	0.0001
Yes			0.3108		
No			0.6892		
Current marital status	11013	1		14.547	0.0001
Married			0.6847		
Not married			0.3153		
Respondents age	11013	2		12.687	0.0001
Less than 20 years			0.0586		
20 - 34 years			0.732		
35 and older			0.2095		
Place of residence	11013	1		8.039	0.005
Rural			0.7973		
Urban			0.2027		

Table 2.6: Bivariate analysis of associations with under-five mortality Household environment characteristics.

Covariate	**Sample Size**	**DF**	**proportion**	**Chi-square**	**p-value**
Household environment characteristics					
Main floor	11010	2		0.98	0.322
Unfinished			0.7185		
finished			0.2815		
Source of drinking water	11013	1		2.005	0.1570
Safe water			0.5473		
Not Safe water			0.4527		

Table 2.7: Bivariate analysis of associations with under-five mortality socio-economic characteristics.

Covariate	Sample Size	DF	proportion	Chi-square	p-value
Socio-economic characteristics					
Wealth index	11013	2		8.285	0.016
Poor			0.4617		
Middle			0.1577		
Rich			0.3851		
Number of Child ever born	11013	2		7.984	0.029
Less than 2 children			0.3018		
2 - 4 children			0.3423		
More than 4 children			0.3559		
Number of children 5 or under	11013	2		45.66	0.0001
Less than 2 children			0.8559		
2 - 4 children			0.1171		
More than 4 children			0.027		
Mother currently working	11013	1		3.735	0.053
No			0.1486		
Yes			0.8514		
Respondent level of education	11013	2		3.349	0.082
No education			0.2275		
Primary education			0.6419		
Secondary and higher			0.1306		
Number of living children	11013	2		153.251	0.0001
Less than 2 children			0.5812		
2 - 4 children			0.2568		
More than 4 children			0.1622		

2.5 Conclusion

Exploratory data analysis plays an important role to help get a preliminary understanding of trend and patterns before using model based approaches. According to frequency tables, it has been observed that most of the respondents were aged from 20 to 34 years and also by those with primary education. There were slightly more males than female. The sample had 84.5% of the respondents from rural areas. There were eleven out of sixteen factors found to be associated with a child survival among them we have: birth order number, marital status, current breastfeeding, HIV status of a mother, place of residence, the age of the respondent, number of children 5 or under in a household and number of a child ever born.

Chapter 3

Generalized Linear Models

3.1 Introduction

In Chapter 1 we stated that the main objective of this study is to identify factors associated with the under-five mortality using THMIS data in Tanzania. The outcome is dichotomous (child alive or child not alive) which can be assumed to follow the Bernoulli distribution which is a member of the exponential family. In order to make valid statistical inference all covariates which potentially affect the child survival status will be assumed to have fixed effects thus the Generalized Linear Model can be fitted to the data of interest. Firstly, we review the theory of the Generalized Linear Models.

3.2 Review of the Generalized Linear Models

The Generalized Linear Model (GLM) incorporates covariates in order to explain the dependence of an outcome variable on measured covariates values. The outcome variable is assumed to come from an exponential family of distribution. The GLM is also used to accommodate non-normal responses and provide a unified approach to modeling all type of response variable (Dobson and Barnett, 2008; McCullagh and Nelder, 1989; Olsson, 2002). One can describe the GLM as a unified mathematical way of describing the relationships between a response variable and a set of covariates. More specifically the generalized linear model is an extension of the linear model is given by.

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{3.1}$$

where, $\boldsymbol{X}$ is the design matrix of covariates, $\boldsymbol{\beta}$ is the vector of coefficients and $\boldsymbol{\epsilon}$ is the vector of error terms. Let $\eta = \boldsymbol{X\beta}$, here η is the linear predictor part of the model. Since a generalized linear model extends the general linear model by relaxing the assumption that dependent variable y is (independent) normally distributed with mean zero and constant variance, this allows the distribution to be part of the exponential family of distributions (Olsson, 2002). Instead of modeling the mean directly, the model is specified in term of some function $g(\boldsymbol{\mu})$, so the model becomes

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{X\beta} \tag{3.2}$$

where, $g(.)$ is the link function. We now look at the key properties of the exponential family of distributions.

3.2.1 Exponential Family of Distributions

The exponential family is known as a general class of distribution that includes the well known normal distribution as a special case (Olsson, 2002). One can show that a distribution belongs to the exponential family of distribution provided the probability distribution function (pdf) of an observation $y_i (i = 1, 2, \ldots, n)$ from the distribution can be expressed as

$$f(y_i, \theta_i, \phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right) \tag{3.3}$$

where, $a(\phi)$ and $b(\theta_i)$ are known functions and $c(y_i, \phi)$ is some function of y_i and ϕ. The parameter θ_i is called the canonical parameter, ϕ is the dispersion parameter. The mean, $\mu = E(y) = b^{'}(\theta)$, and the variance, $var(y) = \phi b^{''}(\theta)$, can be obtained as shown in Appendix B. We next study the components of the Generalized Linear Models (GLMs) and then consider parameter estimation for the model.

3.2.2 Components of Generalized Linear Models

GLM is comprised of three components namely, random component, link function and systematic component. The random component refers to the probability distribution of the response variable Y. The distribution may include the normal distribution and we say the random component is normally distributed. This leads us to the ordinary regression model. When the outcome observations have the value "0" and "1" then the most plausible distribution for a random variable is the Bernoulli distribution. The link function is the logit link. This component leads to the application of the logistic regression models. The systematic component: is a function of covariates $x_1, x_2, x_3, \ldots, x_p$ that leads to the linear predictor η given by $\eta = \alpha + \sum_{j=1}^{n} x_j \beta_j$.

3.2.3 Parameter Estimation

Maximum Likelihood can be used as a theoretical basis for the parameter estimation in generalized linear models. The likelihood function is given by

$$\begin{aligned} \mathrm{L}(y;\theta) &= \prod_{i=i}^{n} \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right) \\ &= \exp\left(\sum_{i=1}^{n}\left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right)\right). \end{aligned} \tag{3.4}$$

The log-likelihood function is given by

$$l(y,\theta) = \log \mathrm{L}(y,\theta) = \sum_{i=i}^{n}\left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right). \tag{3.5}$$

The parameters are obtained by taking the derivatives of the log-likelihood function with respect to $\beta_j (j = 0, 1, 2 \ldots, p)$ and equating to zero then solve the equations simultaneously. Here p is the number of parameters. We obtain the score vector function given by $(U_{\beta_1}, U_{\beta_2}, U_{\beta_3}, \ldots, U_{\beta_p})'$ where U_{β_j} is given by

$$U_{\beta_j} = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j}.$$

Using the chain rule we have

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}.$$

The first factor is given by $\frac{\partial l_i}{\partial \theta_i} = \frac{y_i - b^{'}(\theta i)}{a_i(\phi)} = \frac{y_i - \mu_i}{a_i(\phi)}$ since $\mu_i = E(y_i) = b_i^{'}(\theta_i)$. The second factor is $\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b^{''}(\theta_i)} = \frac{a_i(\phi)}{var(y_i)}$. The third factor depends on the link function $\frac{\partial \mu_i}{\partial \eta_i}$. The fourth factor is $\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$ where x_{ij} is the j^{th} element of the covariates vector $\mathbf{x_i}$ for the i^{th} observation. Substituting the factors for $\frac{\partial l_i}{\partial \theta_i}, \frac{\partial \theta_i}{\partial \mu_i}, \frac{\partial \mu_i}{\partial \eta_i}$ and $\frac{\partial \eta_i}{\partial \beta_j}$.

$$\frac{\partial l}{\partial \beta_j} = \frac{y_i - \mu_i}{var(y_i)} b^{''}(\theta_i) x_{ij} = \frac{y_i - \mu_i}{a_i(\phi)} x_{ij}.$$

The system of equations to be solved for $\beta_j^{'}s$ is given by the following

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \left(\frac{y_i - \mu_i}{a_i(\phi)} x_{ij} \right) = 0. \tag{3.6}$$

The system of equations can be solved iteratively using either Fisher's scoring (FS) or Newton-Raphson (NR) algorithms for maximum likelihood estimation (McCullagh and Nelder, 1989; Olsson, 2002). These algorithms are available in statistical software such as SAS and STATA. Many packages, including SAS, use Fisher scoring algorithm as a default iterative technique. Using this FS method is equivalent to using iterative reweighted least squares (IWLS). Both NR and FS gives similar parameter estimates. However, estimated covariance matrix parameters could be slightly different. This is due to the fact that FS is based on the expected information matrix while NR is based on the observed information matrix. In the case of the logistic regression model, both expected and observed information matrices yield identical covariance matrices for both models. The parameter estimates are used to assess the model adequacy and its fit. In the next section, we consider methods for model selection and diagnostics.

3.3 Model Selection and Diagnostics

3.3.1 Model Selection

Akaike's Information Criterion

One way to evaluate a model is to use the Information Criterion (IC). This criterion attempts to quantify how well the model has predicted the data. The Akaike's Information Criterion(AIC) is a useful statistic for comparing the relative fit of different models. This statistic was proposed by Akaike (1974) and is given by

$$\mathrm{AIC} = -2\mathrm{logLikelihood} + 2\mathrm{k} \tag{3.7}$$

where, k is the number of parameters in the model. This method penalizes the log-likelihood for the number of parameters estimated (Akaike, 1974). A model that minimizes the AIC is preferred. The method is particularly useful when comparing non-nested models.

Schwarz Criterion

An alternative to AIC for also comparing non-nested models is Schwarz Criterion (SC) also known as Bayesian Information Criterion (BIC) and was proposed by Schwarz et al. (1978). SC is given by

$$SC = -2\mathrm{logLikelihood} + \mathrm{k}\log(n). \tag{3.8}$$

Here, n is the sample size and k is the number of parameters estimated. SC produces more severe penalization on the likelihood for estimating more parameters (Allison, 2012). The model chosen is the one which leads to the minimum SC. While doing a model selection, we can narrow down the options before comparing models. This can be done by building the regression model step by step using selection procedure of variables that enters the model. These procedures are; forward, backward and stepwise selection. Forward selection starts with the null model and enters one covariate at a time, that is found to be significant at some level of significance(α) until all significant variables are added to the model. Backward selection begins with the model that contains all covariates and drops one at a time, that is, insignificant at some level of significance α. This is done until all non-significant variables are removed from the model. The stepwise selection works in the same way as the forward selection procedure. However, the advantage of stepwise over forward selection is that variables already in the model are considered to be excluded in the model each time the new covariate is added in the model. In the case where there are many covariates the stepwise procedure is a preferable since it minimizes the chance of keeping redundant variables in the model, and leaving out some important ones.

Choice of Measure of Fit

The deviance and Pearson chi-square tests provide large sample tests of the model fit. These tests are useful depending on the kind of data that is being analyzed. Deviance has an advantage over Pearson chi-square test since it is a likelihood-based test that is useful for comparing nested models. AIC is normally used for comparing competing models without making any inference. The combination of AIC and SC can be used to select the model.

3.3.2 Model Checking

Goodness-of-fit Test

The deviance and the Pearson Chi-square tests are the statistics that could be used for assessing the goodness of fit of the model.

Deviance

In the Generalized Linear Models the fit of the model can be assessed through the deviance. The deviance can also be used to compare the models that are nested. In order to define the deviance we let $l(\hat{\mu}, \phi, y)$ be the log-likelihood of the reduced model at the maximum likelihood estimate and also let $l(y, \phi, y)$ be the log-likelihood estimate of the full or saturated model. The deviance is then given by

$$\text{Deviance} = 2(l(y, \phi, y) - l(\hat{\mu}, \phi, y)). \tag{3.9}$$

For any distribution that has a scale parameter ϕ the scaled deviance is given by

$$D^* = \frac{\text{Deviance}}{\phi}.$$

The Binomial and Poisson distribution has deviance and scaled deviance that are identical because $\phi = 1$ in both distributions. Given that the model is true, as the sample size increase deviance will asymptotically tend towards the chi-square distribution. Suppose that one model provides a deviance D_1 with degree of freedom (df_1) and another model provides a deviance D_2 with degree of freedom (df_2). In order to compare two models, we need to compute the differences between deviances $D_1 - D_2$ and also the degrees of freedom $df_1 - df_2$. This will results in a chi-square distribution. This kind of test works in comparing two models given those parameters of the first model corresponding to D_1 are a subset of the parameters in the second model corresponding to D_2. We now look at the other statistic that can be used to assess the fit of the model.

The Generalized Pearson Chi-square Statistics

The alternative to deviance for testing and comparing models is the generalized Pearson chi-square statistic. The Pearson chi-square the statistic is defined as

$$\chi^2_{\text{Pearson}} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\widehat{\text{var}(\hat{\mu}_i)}} \tag{3.10}$$

where $\widehat{\text{var}(\hat{\mu}_i)}$ is the estimated variance function. The deviance is often preferred over the Pearson chi-square statistic since maximum likelihood estimation in the Generalized linear Models minimizes the deviance while the Pearson does not have the necessary additive properties like the deviance for comparing models.

3.4 Logistic Regression Model

The logistic regression is the most commonly used statistical modeling technique that describes the relationship of several covariates (X's) to a dichotomous response variable. The goal of the logistic regression model with multiple predictors is the same as that of the ordinary multiple linear regression models; in a way that we attempt to construct a model to describe the relationship between a response and one or more predictor variables (David and Mitchel, 1994). In this we study focus on the logistic regression model with more than one predictor variable known as multiple logistic regressions.

3.4.1 Model

Consider the p explanatory/predictor variables of interest denoted by the vector $\boldsymbol{x} = (x_1, x_2, x_3, \dots, x_p)$ for the i^{th} individual. Let the probability that the event, is present, be denoted by $P(Y_i = 1) = \pi_i$ for the i^{th} individual and let the event being not present be denoted by $P(Y_i = 0) = 1 - \pi_i$. The logistic analysis does not require assumptions such as linearity and normality of the dependent variable and residuals. This method is based on the log transformation of the odds and is given by the

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}. \tag{3.11}$$

Thus alternative formula that refers directly to the probability of the outcome of interest is as follow

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})}{1 + \exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})}. \tag{3.12}$$

which is the probability of the event occurring, and the probability of the event is given by $1 - \pi_i$. The ratio of the odds of the event occurring in one group to the odds of it occurring in the other group is known as an odds ratio. The purpose of logistic regression in this study is to find the parameters $\beta_0, \beta_1, \dots, \beta_p$ that best fit the data relating child survival to a number of covariates using 2011-2012 Tanzania HIV/AIDS and Malaria Indicator Survey data. The logistic regression enables researchers to overcome many of the linear regression assumptions that are too restrictive. The following assumptions are relaxed under logistic regression.

- The linear relationship between dependent and independent variables is not assumed.
- The dependent variable do not need to be normally distributed.
- The dependent variables must not have homoscedastic variance (variance do not have to be the same within categories).
- Also normally distributed error terms are not assumed.
- Response variable is required to be binary.

We now look at how the parameters can be estimated using the maximum likelihood.

3.4.2 Parameter Estimation

In this study child survival status which is the dependent (response) variable $Y_i(i = 1, 2, \ldots, n)$ is dichotomous and the underlying probability distribution is Bernoulli. This can be expressed in the form $Y_i \sim Benoulli(\pi_i)$ and the p predictor variables (Czepiel, 2002; Lemeshow and Hosmer, 2000; Wood, 2006). In order to obtain the maximum likelihood estimates. Let

$$Y_i = y_i \mid \ x_{1i}, x_{2i}, \ldots, x_{pi} \sim \text{Benoulli}(\pi_i), i = 1, 2, \ldots, n,$$

hence probability mass function (PMF) for a Bernoulli distribution is:

$$P(Y_i = y_i \mid \ x_{1i}, x_{2i}, \ldots, x_{pi}) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}, i = 1, 2, \ldots, n. \tag{3.13}$$

The likelihood of observing values of the response variable for all the observations is given by

$$\text{L} = Pr(y_1, y_2, y_3, y_4, \ldots, y_n).$$

Assuming that observations are independent, the likelihood function is given by the product of the individual probabilities

$$\mathrm{L} = Pr(y_1)Pr(y_2)\dots Pr(y_n) = \prod_{i=1}^{n} Pr(y_i).$$

Since we know that the responses $y_i^{'}s$ are from Bernoulli so the likelihood function is given by

$$\mathrm{L}(\boldsymbol{\beta} \mid \mathbf{Y}) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}. \tag{3.14}$$

If we Substitute for $\pi_i^{'}s$ in terms of covariates then the Likelihood function becomes

$$\mathrm{L}(\boldsymbol{\beta} \mid \mathrm{Y}) = \prod_{i=1}^{n} \left(\frac{\exp(\boldsymbol{X_i\beta'})}{1+(\exp(\boldsymbol{X_i\beta'}))} \right)^{y_i} \left(\frac{1}{1+(\exp(\boldsymbol{X_i\beta'}))} \right)^{1-y_i} \tag{3.15}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2 \dots, \beta_p)^{'}$ and $\mathbf{X_i}$ is the matrix of covariates with first column containing ones. It is not easy to differentiate the likelihood function thus one needs to simplify the likelihood function further by taking its log. Since the logarithm is a monotonic function, any maximum of the likelihood function will be the maximum of the log likelihood function (Czepiel, 2002). Thus, taking the natural log we obtain the log likelihood expressed as

$$l(\boldsymbol{\beta} \mid \boldsymbol{Y}) = \sum_{i=1}^{n} y_i \log(\pi_i) + \sum_{i=1}^{n}(1-y_i)\log(1-\pi_i). \tag{3.16}$$

In order to obtain the parameter estimates, set the first derivative of log-likelihood with respect to each β equal to zero, so the maximum likelihood estimate for β can be found by setting each of the $K+1$ equation obtained to zero, and solving for each β_k (Czepiel, 2002). Each of such solutions, if any exists, specifies a critical point either a maximum or minimum. The critical point will be the maximum if the matrix of second derivatives is negative definite, which means that every element on the diagonal of the matrix is less than zero. The other useful property of this matrix is that it forms variance-covariance matrix of the parameter estimates. Differentiating each of the $K+1$ equation for the second time with respect to each elements of $\boldsymbol{\beta}$, denoted by β_k leads to the variance-covariance matrix (Czepiel, 2002).

3.4.3 Newton-Raphson Method

After setting the $K + 1$ equations from the first derivative of log-likelihood equate these equations to zero, it results into a system of non-linear equations with each $K+1$ unknown variables. The solution to these equations is the vector with elements $\hat{\beta}_k$. After verifying that the matrix of second partial derivatives is negative definite and that the solution is global maximum instead of a local maximum, then it can be concluded that this vector contains the parameter estimates for which observed data would have the highest probability of occurrence (Czepiel, 2002). However, solving a system of non-linear equation is not easy compared to a system of linear equation. The alternative is to numerically estimate the parameters using iterative methods. The popular method for solving non-linear equation is Newton-Raphson method (Newton's method). Newton-Raphson begins with the initial guess of the solution and uses first two terms of Taylor polynomial evaluated at an initial guess to generate other estimates that are close to the solution. This iterative method process continues until converges to the actual solution (Czepiel, 2002; Moeti, 2010).

3.5 Logistic Model Selection and Checking

3.5.1 Model Selection

Variable Selection

Before fitting the model one has to check the multicollinearity among variables which occur when there is a strong relationship among covariates (Allison, 2012). Multicollinearity does not bias the coefficients but results to unstable coefficients. Good estimates are not guaranteed if two or more variables are highly correlated and may result in large standard errors which lead to invalid statistical inference. Since multicollinearity is the property of predictor variables one can examine it by diagnostic procedure PROC REG with option such as TOL and VIF (Variance Inflation Factor) for logistic regression. The same selection criteria stated in the above subsection 3.3.1 still applies.

Testing Hypothesis About β

The method for testing the significance of the parameter estimates in logistic regression is similar to the approach used for linear regression, but logistic regression uses likelihood function for a binary outcome variable. Once the model is fitted one can test for the significance of each parameter. The distribution of $\hat{\beta}$ in Appendix B is $\boldsymbol{\beta} \sim \mathbf{MVN}(\boldsymbol{\beta}, \boldsymbol{I}^{-1})$ and can be used to test for the significance of $\hat{\beta}_j (j = 1, 2, \ldots, p)$ in the model. The Wald Chi-square is given by

$$\chi^2_{\text{wald}} = \left(\frac{\hat{\beta}_j}{\sqrt{\widehat{V}_j}} \right)^2 \tag{3.17}$$

where V_j's are the diagonal elements of $\boldsymbol{I^{-1}}$. One can use the chi-square distribution with 1 degree of freedom and compare it with Wald Chi-square statistic. The hypothesis being tested here is $H_0 : \beta = 0$ against the alternative $H_a : \beta \neq 0$. If the Wald Chi-square statistic is greater than the table value of chi-square, H_0 is rejected, that means the parameter is significantly different from zero.

Odds Ratio

Let us consider a dichotomous response variable which denotes the occurrence or non-occurrence of an event. Suppose there is one covariate with two categories. The odds ratio is then defined as the ratio of the odds for those with risk factor $(X = 1)$ to the odds for those without the risk factor $(X = 0)$ (Czepiel, 2002). The log of the odds ratio is given by

$$\begin{aligned} \log(\hat{\text{OR}}) &= \log(\text{OR}(x = 1, x = 0)) = \text{logit}(x = 1) - \text{logit}(x = 0), \\ &= (\hat{\beta}_0 + 1 \times \hat{\beta}_1) - (\hat{\beta}_0 + 0 \times \hat{\beta}_1), \\ &= \hat{\beta}_1. \end{aligned} \tag{3.18}$$

The odds ratio can then be computed by exponentiating the difference of the logit between any two population profile and odds ratio is given by

$$\text{OR} = \exp(\hat{\beta}_1). \tag{3.19}$$

The parameter β_1 associated with X represents the change in the log odds from $X = 0$ to $X = 1$. The odds ratio indicates how the odds of the event changes as X change from 0 to 1. suppose we have a continuous variable called X then we can say that as X increases by one unit, the odds of risk factor increase by $\exp(\hat{\beta}_1)$. The confidence interval is discussed in the next subsection.

Confidence Interval for the Odds Ratio

Most of the social science journals often report the point estimates and hypothesis test for coefficients. However, confidence intervals provide a better picture of the sampling variability of the estimates (Allison, 2012). Again the confidence interval for slope and intercept are based on Wald tests. The $100(1 - \frac{\alpha}{2})\%$ confidence interval for the intercept is given by

$$\hat{\beta}_0 \pm Z_{1-\frac{\alpha}{2}} \sqrt{V_0} \tag{3.20}$$

where $\sqrt{V_0}$ is the standard error of β_0. The $100(1 - \frac{\alpha}{2})\%$ confidence interval for intercept is given by

$$\hat{\beta}_j \pm Z_{1-\frac{\alpha}{2}} \sqrt{V_j} \tag{3.21}$$

where $\sqrt{V_j}$ is the standard error of β_j. Here, $Z_{1-\frac{\alpha}{2}}$ is the upper $100(1 - \frac{\alpha}{2})\%$ value from the standard normal distribution. Since these confidence intervals are on the logit scale they have to be transformed by exponentiation in order to get corresponding $100(1-\frac{\alpha}{2})\%$

$$\exp(\hat{\beta}_j \pm Z_{1-\frac{\alpha}{2}} \sqrt{V_j}). \tag{3.22}$$

This is the confidence interval for odds ratio associated with β_j where $j = 1, 2, 3, \ldots, p$. In the next subsection model checking is discussed.

3.5.2 Model Checking

The Hosmer-Lemeshow goodness-of-fit(GOF) statistic χ^2_{HL} is obtained by computing the Pearson Chi-square statistic from the $g \times 2$ table of observed and estimated expected frequencies. Where g is the number of groups. The Hosmer-Lemeshow statistics is given by

$$\chi^2_{HL} = \sum_{k=1}^{g} \frac{(O_k - n_k^{'}\overline{\pi}_k)^2}{n_k^{'}\overline{\pi}_k(1-\overline{\pi}_k)} \tag{3.23}$$

where,$n_k^{'}$ is the total number frequency of subjects in the k^{th} group, O_k is the total frequency of the event outcomes in the k^{th} group and $\overline{\pi}_k$ is the average estimated predicted probability of an event outcome for k^{th} group. The Hosmer-Lemeshow statistics is compared to the Chi-square distribution with $(n-g)$ degrees of freedom, where in SAS the value n can be specified using lack of fit option in the model statement. The default value is $n = 2$ in SAS. The null hypothesis being tested here is H_0 : a model is a good fit against the alternative H_a : the model is not a good fit. The large value of Hosmer-Lemeshow statistic (p-value less than 0.05) suggests a lack of fit of the model. Below the statistics for measuring the predictive power is discussed.

3.5.3 Logistic Regression Diagnostics

Influential Observations

We now focus on detecting potential observations which have a significant impact on the model. Under the ordinary least square regression, we have different types of residuals and influence measure which help us understand the behavior of each observation in the model, such observations turn to be far away from the rest observations. If the observation has too much leverage on the regression line we can view it as an observation that has a significant impact on the model (Hosmer and Lemeshow, 2004). The same methods have been developed for logistic regression.

Leverage of an Observation

This is another measure where the observation with an extreme value on the predictor variable is known as a point with high leverage (Hosmer and Lemeshow, 2004). The leverage is defined as a measure of how far an independent variable deviates from its corresponding mean. The large values suggest covariate patterns far from the average covariate pattern which can have a larger effect on the fitted model even if the corresponding residuals are small (Hosmer and Lemeshow, 2004).

Standardized Pearson Residual

The Standardized Pearson residual is defined to be the standardized difference between the observed frequency and predicted frequency. This residuals measures the relative deviations between observed and fitted values (this applies for logistic regression only) (Hosmer and Lemeshow, 2004). The standardized Pearson residual is given by

$$r_i^{student} = \frac{(y_i - \hat{\pi}_i)}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)(1-h_i)}} \tag{3.24}$$

where h_i is the i^{th} for subject leverage, $\hat{\pi}_i$ is the estimated probability that $y_i = 1$ subject i.

Deviance Residual

This is another type of residual that measures the disagreement between the maxima of the observed and fitted log-likelihood functions. The logistic regression uses maximum likelihood principle, where its objective is to minimize the sum of deviance residuals. This residual is similar to the raw residual in ordinary least square regression (Lemeshow and Hosmer, 2000; Hosmer and Lemeshow, 2004). The objective of the ordinary least square regression is to minimize the sum of square residual. The deviance residual is given by

$$d_i = \sqrt{2 \mid ln(\hat{\pi}_i) \mid}, \text{if } y_i = 1,$$

$$d_i = \sqrt{2 \mid ln(1-\hat{\pi}_i) \mid}, \text{if } y_i = 0$$

where d_i is the individual component known as the deviance residual.

Predictive Accuracy/Ability of the Model

In order to check for the predictive accuracy SAS procedure PROC LOGISTIC produces other model statistics namely, Somer's D, Gamma, C, and Tau-a. All these statistic range between 0 and 1. In all, larger values correspond to a strong association between predicted and observed values. These measures of association are given by

$$\text{Tau-a} = \frac{C - D}{N},$$

$$\text{Gamma} = \frac{C - D}{C + D},$$

$$\text{Somer's D} = \frac{C - D}{C + D + T},$$

$$\text{C} = 0.5(1 + \text{Somer's D}).$$

The C statistic is the proportion of observation pairs with different observed outcomes for which the model correctly predicts a higher probability for observations with the event outcome than the probability for non-event observation. A value of one means that the model assigns the higher probability to all observations with the event outcome compared to non-event observations. We use concordant and discordant pairs to describe the relationship between pairs of observations. The pair said to be Concordant (C) if the subject ranked higher on predictor variable X also ranked higher on response variable Y. The pair is said to be Discordant if the subject ranking higher on predictor variable X ranks lower on the response variable Y. The pair is said to be Tied (T) if subjects have the same classification on predictor and response variable. The total number of pairs is given by N. The value of C correspond to the receiver operating characteristics (ROC) curve in the case of the binary response which is defined below (Šimundić, 2008).

Area Under the Receiver Operating Characteristics

The specificity and sensitivity rely on the cutoff point to classify the result as positive (Lemeshow and Hosmer, 2000). To plot the ROC curve, one needs to plot sensitivity versus 1-specificity. Sensitivity measures the proportion of correctly classified positive outcome or event of interest (death) and specificity measures the proportion of correctly classified event free outcome (no death). ROC provides a complete description of classification accuracy and can be used as a graphical display of the prediction accuracy of the model (Vittinghoff et al., 2011; Šimundić, 2008). The area under the curve (AUC) is between 0 and 1 as shown in figure 3.1. The ROC gives the measure of model ability to classify between subjects which have experienced the outcome versus those who did not.

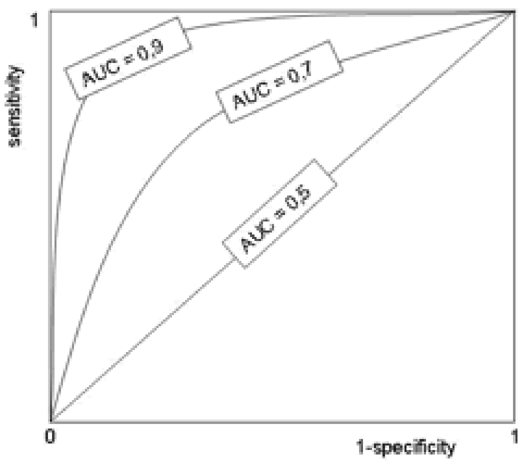

Figure 3.1: Receiver Operating Characteristic (ROC) curve

Source: Šimundić (2008).

The Area under the curve (AUC) is known as the global measure of diagnostic accuracy. This area measures the prediction accuracy of the model. AUC does not tell us anything about individual parameters (Šimundić, 2008). If the area under the curve is large, the better the diagnostic accuracy of the test. Suppose three logistic models were fitted, and model 1 produced AUC of 0.5, model 2 produced AUC of 0.9 also model 3 produced an AUC of 0.7. One can classify model 2 as the better model since it has an excellent diagnostic accuracy thus have a better accuracy. An AUC of 0.5 is not good because the test cannot discriminate between correctly classified positive outcome and those falsely classified as positive. One can classify the relationship between the AUC and diagnostic accuracy as shown in the Table 3.1 below.

Table 3.1: Relationship between area under the curve and diagnostic accuracy

area	diagnostic accuracy
0.9 - 1.0	excellent
0.8 - 0.9	very good
0.7 - 0.8	good
0.6 - 0.7	sufficient
0.5 - 0.6	bad
< 0.5	test not useful

Source:Šimundić (2008).

3.6 Fitting the Logistic Regression Model

The model was fitted using PROC LOGISTIC in SAS; first, univariate models were fitted to identify potential candidate variables associated with the outcome without considering the combined effects of covariates on the response. The multiple logistic models were then fitted with all variables that were identified as significant in the univariate analysis. The goodness-of-fit was tested using the Hosmer-Lemeshow test and the predictive accuracy of the model was assessed through the ROC. The coefficient and odds ratios were interpreted and the limitations of the logistic regression outlined in this section.

3.6.1 Univariate Logistic Regression Model

Table 3.2 displays parameter estimates, Standard errors, p-values and odds ratios for the univariate models. The results are shown in this table confirm some of the bivariate results in section 2.4 in Table **??**. The variables that were found to be significant had p-values which were less than 0.05.

The effect of not breastfeeding was found to be positively associated with under-five child mortality (p-value=0.0001). The corresponding odds ratio was 1.686 (with 95% CI: 1.5189 ; 1.8723). The odds of death for a child from a mother who does not breastfeed were 1.686 times the odds of death for a child from a mother who breastfeed. The effect of a mother being married was found to be negatively associated with under-five child mortality (p-value=0.0002). The corresponding odds ratio was 0.817 (with 95% CI: 0.7359 ; 0.9074). The odds of death for a child from a mother that is married were

Table 3.2: Univariate coefficients, standard errors, p-values and odds ratios.

Effects	**Estimate**	**Standard Errors**	**P-value**	**Odds Ratio**	**95% confidence limits**	
					lower	Upper
Sex of a child (ref. Male)						
Intercept	-3.1768	0.0499	0.0001			
Female	-0.0684	0.0499	0.1709	0.934	-0.1662	1.0298
Age of household head (ref. over 34 years						
Intercept	-2.9053	0.1462	0.0001			
21 to 34 years	-0.1297	0.1542	0.4	0.878	0.6503	1.1883
less than 20 years	0.4781	0.286	0.0946	1.613	0.9235	2.8254
Currently breastfeeding (ref. Yes)						
Intercept	-3.2398	0.0535	0.0001			
No	0.5223	0.0535	0.0001	1.686	1.5189	1.8723
Marital status (ref. Unmarried)						
Intercept	-3.0843	0.0536	0.0001			
Married	-0.2022	0.0536	0.0002	0.817	0.7359	0.9074
Mother's HIV status (ref. HIV-Negative)						
Intercept	-2.8119	0.0876	0.0001			
HIV-Positive	0.4163	0.0876	0.0001	1.516	1.2782	1.8004
Birth order number (ref. Less than 2 births)						
Intercept	-3.2046	0.0517	0.0001			
2 to 4 births	-0.0169	0.0767	0.8256	0.983	0.8467	1.1427
above 4 births	-0.2345	0.0756	0.0019	0.791	0.6825	0.9173
Number of children ever born (ref. less than 2 children)						
Intercept	-3.1601	0.05	0.0001			
2 to 4 children	0.1011	0.0706	0.1523	1.106	0.9641	1.2706
more than 4 children	-0.1666	0.069	0.0158	0.847	0.7400	0.9691
Mother's age (ref. over 34 years)						
Intercept	-3.1506	0.0797	0.0001			
Between 20-34	0.0572	0.0866	0.5088	1.059	0.8943	1.2547
less than 20 years	0.2558	0.1429	0.0735	1.292	0.9774	1.7090
Wealth index (ref. rich						
Intercept	-3.2302	0.0553	0.0001			
middle	-0.2571	0.0909	0.0047	0.773	0.6477	0.9241
poor	0.1114	0.0698	0.1104	1.118	0.9756	1.2817
Education Level (ref. higher education)						
Intercept	-3.1455	0.0612	0.0001			
No education	-0.1717	0.0868	0.048	0.842	0.7111	0.9984
Primary education	-0.0186	0.0708	0.7933	0.982	0.8550	1.1277
Number of children 5 or under (ref. less than 2 children)						
Intercept	-3.5683	0.11	0.0001			
2 to 4 children	-0.3309	0.138	0.0165	0.718	0.5488	0.9414
more than 4 children	-0.2567	0.2012	0.2021	0.774	0.5225	1.1476
Working status of a mother (ref. Yes)						
Intercept	-3.0793	0.0705	0.0001			
No	0.1292	0.0705	0.0669	1.138	0.9918	1.3065
Number of children living (ref. Less than 2 children)						
Intercept	-3.3052	0.0562	0.0001			
More than 4 children	-0.631	0.0888	0.0001	0.532	0.4475	0.6332
2 to 4 children	-0.1241	0.0802	0.1218	0.883	0.7554	1.0336
Type of place of residence (ref. Urban)						
Intercept	-3.0587	0.0623	0.0001			
Rural	-0.1782	0.0623	0.0042	0.837	0.7411	0.9455
Main floor material (ref. Unfinished)						
Intercept	-3.1505	0.0556	0.0001			
Finished	0.0513	0.0556	0.3559	1.053	0.9445	1.1738
Source of drinking water (ref. unsafe water)						
Intercept	-3.1727	0.0498	0.0001			
Safe water	0.0374	0.0498	0.4528	1.038	0.9420	1.1445

0.817 times the odds of death for a child from a mother who is not married. The effect of mothers HIV status (positive) was found to be positively associated with under-five child mortality (p-value=0.0001). The corresponding odds ratio was 1.516 (with 95% CI: 1.2782 ; 1.8004). The odds of death for a child from a mother who is HIV-positive were 1.516 times the odds of death for a child from a mother who is HIV-negative. The effect of childbirth order number above four was found to be negatively associated with under-five child mortality (p-value=0.0019). The corresponding odds ratio was 0.791 (with 95% CI: 0.6825 ; 0.9173). The odds of death for a child whose birth order number is above four were 0.791 times the odds of death for a child whose birth order number is less than two. The effect of a number of children ever born that is above four was found to be negatively associated with under-five child mortality (p-value=0.0158). The corresponding odds ratio was 0.847 (with 95% CI: 0.7400 ; 0.9691). The odds of death for a child from a mother that gave birth to more the four children were 0.847 times the odds of death for a child from a mother who gave birth to less than two children. The effect of a mother with no education was found to be negatively associated with under-five child mortality (p-value=0.048). The corresponding odds ratio was 0.842 (with 95% CI: 0.7111 ; 0.9984). The odds of death for a child from a mother with no education were 0.842 times the odds of death for a child from a mother with higher education level. The effect of a number of living children that is above four was found to be negatively associated with under-five child mortality (p-value=0.0001). The corresponding odds ratio was 0.532 (with 95% CI: 0.4475 ;0.6332). The odds of death for a child from a mother with more than four children alive were 0.0.532 times the odds of death for a child from a mother who has less than two children alive. The effect of type of place of residence (rural area) was found to be negatively associated with under-five child mortality (p-value=0.0042). The corresponding odds ratio was 0.837 (with 95% CI: 0.7411 ; 0.9455).

3.6.2 Multiple Logistic Regression Model

Model Selection

Stepwise, forward and backward selection procedures were used to select important variables associated with the outcome variable (survival status) in Tanzania. All three procedures provided similar variables/factors that were identified to be important. In the model, two-way interaction effects found to be significant was included. Table 3.3 shows the model fit statistics that is used in comparing two models.

Table 3.3: Model fit statistics for logit model.

Model fit statistics	
AIC	3181.287
BIC	3304.65
DF	17

Model Checking

Multicollinearity was checked for the variables in the model and two variables that are, number children alive and a number of children ever born were found to have the Tolerance less than 20% or variance inflation factor above five (see Table presented in Appendix C). The variables were the number of children ever born and a number of children alive. This suggests that these variables contain similar information hence one can be dropped from the model. To test for the goodness of fit of the model one can use the Hosmer-Lemeshow test using 10 as the number of groups. The goodness-of-fit Chi-square statistics for Hosmer and Lemeshow is 2.34 with 8 degrees of freedom and the corresponding p-value is 0.9686 as shown in Table 3.4. This indicates that there is insufficient evidence to claim that the model does not fit the data adequately thus one can conclude that the model fitted the data adequately , that is, predicted probabilities are approximately the same as the observed values.

Table 3.4: Hosmer and Lemeshow Goodness-of-fit test.

Goodness-of-fit test	
Number of observations	11013
Number of groups	10
Hosmer-Lemeshow Chi-Square	2.34
P-value	0.9860

3.6.3 Prediction Accuracy of the Model

It is important to check how much the predicted probability agrees with outcomes. The main objective is to have a model which maximizes the chance and sensitivity of identifying individuals that need justified intervention (Moeti, 2010). This means that one is interested in reducing the proportion of individuals that are classified incorrectly as having outcome or failure. One can validate the model by checking the prediction accuracy, in which this could be done by checking how often the model predicts correctly predicts the outcome. Table 3.5 shows the association of predicted probabilities and observed outcomes with the area under the curve being c=0.747 and a concordant rate of 72.5 which tells us how good the model is for separating the 0's and 1's with a chosen model. Figure 3.2 shows the ROC curve of the fitted model and the area under the curve C=0.747 which implies that 75% of the probabilities are predicted correctly, which is a good predictive accuracy. The model correctly assigned higher probability to child status (not alive). The measures Somer's D, Gamma, and Tau-a are the summaries of the table of concordant and discordant pairs. These measures are most likely to lie between 0 and 1 where the large values indicate better predictive ability of the model. These can be viewed as the measures of strength and direction of the relationship between pairs. The value for Gamma is 0.520 which suggest that there is no perfect association. It is interpreted as 52% fewer errors are made in prediction by utilizing the estimated probabilities than by a chance alone. One of the problems with this statistic is the tendency to overstate the strength of association between probabilities and outcome. The value for Somer's D is 0.496. This shows that not all pairs are concordant and one may use it to compare model.

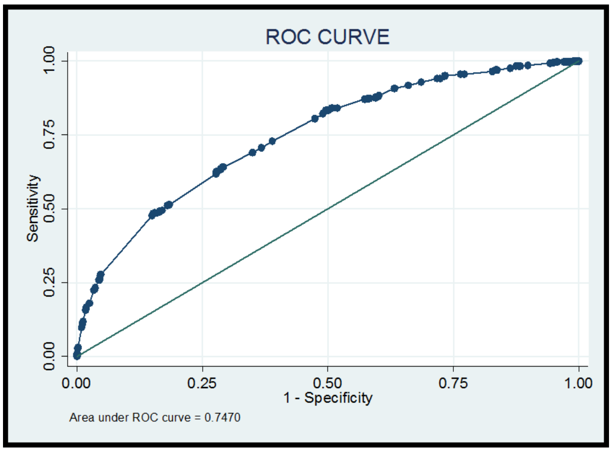

Figure 3.2: Receiver Operating Characteristic (ROC) curve for logit model.

Table 3.5: Association of predicted probabilities and observed outcome.

Association of predicted probabilities and observed responses			
Percent Concordant	72.5	**Somers' D**	0.496
Percent Discordant	22.9	**Gamma**	0.52
Percent Tied	4.7	**Tau-a**	0.038
Pairs	4217640	**c**	0.748

Interpretation of the Coefficient of the Model and the Odds Ratio

Table 3.6 shows the estimated coefficients, standard errors, and p-value for the logistic regression model. The calculated odds ratios and corresponding 95% confidence intervals are also shown. The effect of not breastfeeding was found to be positively associated with the under-five mortality (p-value=0.0001). The corresponding odds ratio was 1.973 (with 95% CI:1.6122;2.4142). The odds of death for a child from a mother who does not breastfeed were 1.973 times the odds of death for a child from a mother who does breastfeed. The effect of HIV status of a mother who is HIV-positive was found to be positively associated with the under-five mortality (p-value=0.007). The corresponding odds ratio was 1.282 (with 95% CI:1.0702;1.5362). The odds of death for a child from a mother who is HIV positive were 1.282 times the odds of death for a child from a mother who is HIV negative. The effect childbirth order number above four was found to be positively associated with under-five mortality (p-value=0.0001). The corresponding odds ratio was 2.842 (with 95% CI:2.1233;3.8047). The odds of death for a child whose birth order number is above four were 2.842 times the odds of death for a child whose birth order number is less than two. The effect of the number of children alive who are more than four was found to negatively associated with under-five mortality (p-value=0.0001). The corresponding odds ratio was 0.251 (with 95% CI:0.1862;0.3396). The odds of death for a child from a mother with more than four children alive were 0.251 times the odds of death for a child from a mother with less than two children alive. The effect of childbirth order number above four depends on not breastfeeding and was found to be positively associated with under-five mortality (p-value=0.0008). The corresponding odds ratio was 1.647 (with 95% CI:1.2307;2.2053). The odds of death for a child whose birth order number is above four and from a mother who does not breastfeed were 1.647 times the odds of death for a child whose birth order number is less than two and from a mother who breastfeeds. The effect of mother's age from 20 to 34 years depends on not breastfeeding and is found to be negatively associated with under-five (p-value=0.0001). The corresponding odds ratio was 0.647 (with 95% CI:0.5254;0.7974). The odds of death for a child from a mother with age between 20 and 34 years who does not breastfeed were 0.647 times the odds of death for a child from a mother with age over 34 years and breastfeed. The effect of mother's age less than 20 years depends on not breastfeeding and was found to be positively associated with the under-five mortality (p-value=0.0004). The corresponding odds ratio was 1.888 (with 95% CI:1.3246;2.6919). The odds of death for a child from a mother with age less than 20 years and does not breastfeed were 1.888 times the odds of death for a child from a mother with age over 34 years and does breastfeed. The effect of the number of children alive greater than five depends on not breastfeeding and was found to be negatively associated with under-five mortality (p-value=0.002). The corresponding odds ratio was 0.625 (with 95% CI:0.4639;0.8427). The odds of death for a child from a mother with more than four children alive and not breastfeeding was 0.625 times the odds of death for a child from a mother with less than two children alive and does breastfeed.

Table 3.6: Logistic regression model coefficients, standard errors and odds ratios.

Analysis of Maximum Likelihood Estimates				**95% confidence interval**		
Effects	**Estimate**	**Standard error**	**P-value**	**Odds ratio**	**Lower**	**Upper**
Socio-demographic characteristics						
Intercept	-3.4154	0.1608	0.0001			
Breastfeeding(BF)						
Yes(reference)						
No	0.6795	0.103	0.0001	1.973	1.6122	2.4142
HIV Status(HS)						
Negative(reference)						
Positive	0.2486	0.0922	0.007	1.282	1.0702	1.5362
Respondent Age(MA)						
Over 34 years(reference)						
20-34 years	0.0749	0.1064	0.4813	1.078	0.8749	1.3277
Birth Order Number(BON)						
Less than 2 births(reference)						
2-4 births	-0.058	0.1021	0.5702	0.944	0.7725	1.1527
above 4 births	1.0446	0.1488	0.0001	2.842	2.1233	3.8047
Socio-economic characteristics						
Children five and under(C5)						
Less than 2 children(reference)						
2-4 children	-0.1155	0.1418	0.4153	0.891	0.6747	1.1764
over 4 children	-0.178	0.2049	0.3849	0.837	0.5601	1.2506
Number of children alive(CL)						
Less than 2 children(reference)						
5 or more children	-1.3805	0.1533	0.0001	0.251	0.1862	0.3396
2-4 children	-0.1658	0.0991	0.0943	0.847	0.6977	1.0288
Interaction between Socio-demographic and Socio-economic characteristics						
Breastfeeding by birth order number						
Yes by less than 2(reference)						
No by 2-4	0.0318	0.102	0.7553	1.032	0.8453	1.2608
No by above 4	0.4992	0.1488	0.0008	1.647	1.2307	2.2053
Less than 20 years	0.1088	0.1815	0.5489	1.115	0.7812	1.5913
Breastfeeding by Respondent age						
Yes by over 34 years(reference)						
No by 20-34 years	-0.435	0.1064	0.0001	0.647	0.5254	0.7974
No by less than 20 years	0.6357	0.1809	0.0004	1.888	1.3246	2.6919
Breastfeeding by Number of children alive						
Yes by less than 2(reference)						
No by 5 or more children	-0.4696	0.1523	0.002	0.625	0.4639	0.8427
No by 2-4 children	-0.068	0.0991	0.4925	0.934	0.7693	1.1345

3.6.4 Logistic Regression Diagnostics Plots

Different techniques of diagnostics have been discussed in section 3.5.3. We will now focus on detecting potential observation that has significant impact on the model. The importance of focusing on this help us to detect if there was any error in data entry and influential data may badly influence or skew the regression estimation. The residual and influence measures that help us understand how observations behave in the model discussed includes Standardized Pearson residuals, Standardized residuals, and Deviance residuals. Figure 3.3 displays influence diagnostics which were produced by using INFLUENCE option in procedure PROC LOGISTIC to fit a logistic regression model to the data. The vertical axis on each plot represents the value of the diagnostic, and the horizontal axis represents case number of the observation. These plots are useful for identification of extreme values. The observations that are further away from zero are influential observation. The plots of the Pearson residual and standardized Pearson residuals indicates that case such as 1214, 2993, 4911, 8622 and many other cases are poorly accounted for by the model.

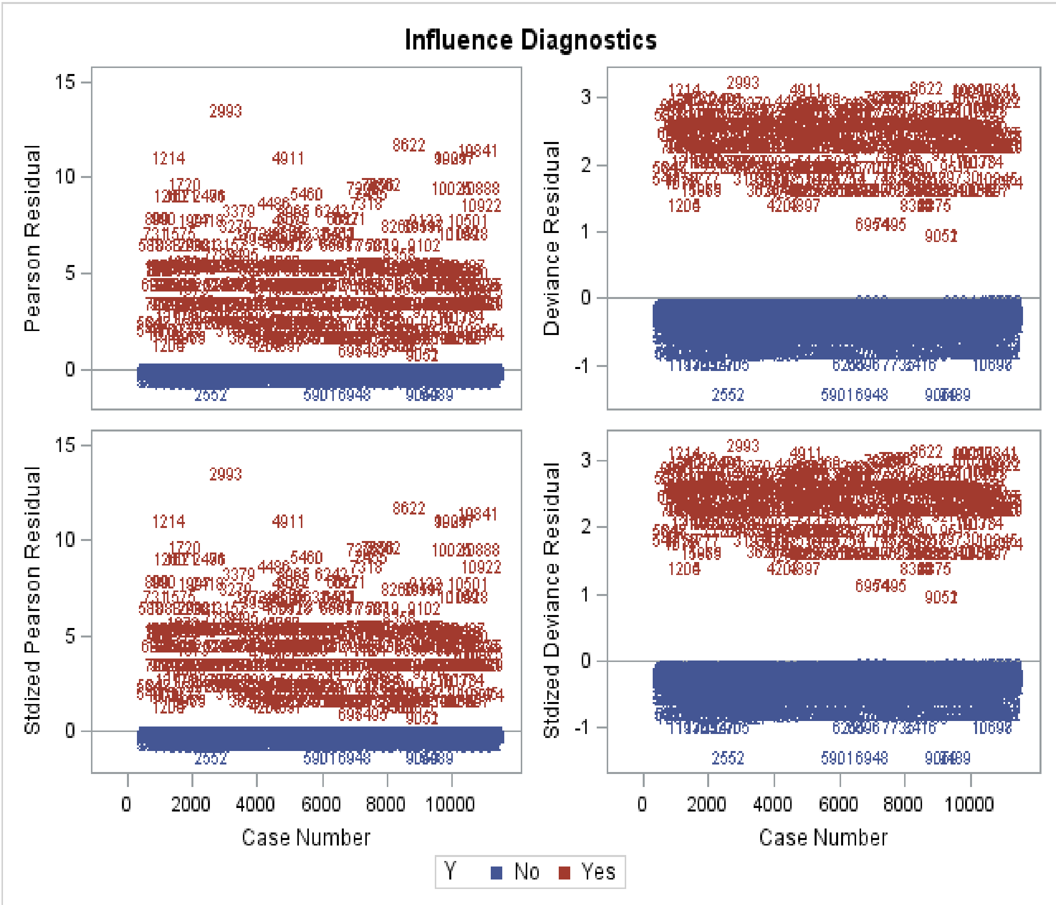

Figure 3.3: Logistic regression diagnostics plots from influence option.

3.7 Limitations of Logistic Regression

In logistic regression, no assumptions are made about distributions of the covariates, but covariates should not be highly correlated to one another since it may lead to problems with estimation. A large sample size is required to obtain sufficient numbers in both categories and response. More covariates require larger sample sizes. The smaller the sample sizes, the less powerful is the Hosmer-Lemeshow. The other limitation that when there is non-linear relationship between log odds and covariates one may obtain invalid results and furthermore ordinary logistic regression does not account for the complex nature of the survey design which can lead to invalid statistical inference. In the next Chapter we consider the method which takes into account the survey design features.

3.8 Survey Logistic Regression Model

Many statistical analyses assume that the data being analyzed is drawn from a finite population by a simple random sampling, where every unit in the population has an equal chance of being chosen during sample selection. However, in real-life survey data are collected from finite population, where the population is stratified by variable of interest (e.g. region, type of place of residence). This ensures the balance in the number of respondent for each category of the variable (An, 2002). Survey logistic regression model has a similar theory as ordinary logistic regression. However, survey logistic regression accounts for the complexity of the survey design (Moeti, 2010). We can make a valid statistical inference by using survey logistic regression which to account for stratification, clustering, and unequal weighting. In the ordinary logistic regression, a model is fitted and selected based on the assumption that the data are collected using simple random sampling. If the complexity of the design is ignored when modeling, the standard errors would be underestimated or overestimated that hence leading to wider or narrow confidence intervals. Survey logistic regression and ordinary logistic regression would be identical if the data are collected using simple random sampling. The main advantage of stratification is that the survey is easier to administer, and parameters could be estimated for each stratum in which themselves can be important. Dividing the population into strata could reduce the variance of the estimator of a population total (An, 2002; Lemeshow and Hosmer, 2000). The methods of parameter estimation for survey logistic are presented in the section that follows.

3.8.1 Parameter Estimation

In complex survey design, the independent assumption does not hold, when cluster are drawn they might introduce correlation among observations. We need to appropriately estimate the standard errors associated with the model coefficients. In order to do such, we need to account for the complexity of the sample design. The standard error produced while assuming a simple random sample will probably underestimate the true population value (Siller and Tompkins, 2006). In the data considered the primary sample units were sampled in the first stage in each stratum (e.g. Location or region). In the second stage, the household was sampled. Thus we specify the response variable as y_{hijk} ($h = 1, 2, \ldots, H_{kji}; i = 1, 2, \ldots, n_{kj}; j = 1, 2, \ldots, m_k; k = 1, 2, \ldots, K$) which is 1 if the event occurred in h$^{\text{th}}$ individual within i$^{\text{th}}$ household within j$^{\text{th}}$ primary sample units nested within k$^{\text{th}}$ stratum, and 0 otherwise. The total number of observations is given by $n = \sum_{k=1}^{K} \sum_{j=1}^{m_k} n_{kj}$ and sampling design weight for the kjih$^{\text{th}}$ are given in the dataset which are denoted by $\boldsymbol{w}_{kjih}$. The weights are based on sampling probability calculated at each stage. These design weights were obtained by multiplying household design weights by the inverse of the household response rate, by stratum. Let the probability that the event occurred in h$^{\text{th}}$ individual within i$^{\text{th}}$ household within j$^{\text{th}}$ primary sample units nested within k$^{\text{th}}$ stratum be $\pi_{kjih} = P(y_{hijk} = 1)$ and the probability that the event did not occur in h$^{\text{th}}$ individual within i$^{\text{th}}$household within j$^{\text{th}}$ primary sample units nested within k$^{\text{th}}$ stratum be $1 - \pi_{kjih} = P(y_{hijk} = 0)$. The pseudo maximum likelihood is constructed as the product of individual contributions to the likelihood (Lemeshow and Hosmer, 2000). The contribution of a single observation using pseudo maximum likelihood is given by

$$\pi_{kjih}^{w_{kjih} y_{kjih}} (1 - \pi_{kjih})^{(1 - w_{kjih} y_{kjih})}.$$

Thus the pseudo- likelihood function is given by

$$L(\boldsymbol{\beta};\boldsymbol{Y}) = \prod_{k=1}^{K}\prod_{j=1}^{m_k}\prod_{i=1}^{n_{kj}}\prod_{h=1}^{H_{kji}} \pi_{kjih}^{w_{kjih}y_{kjih}}(1-\pi_{kjih})^{(1-w_{kjih}y_{kjih})}. \tag{3.25}$$

The pseudo log-likelihood function is given by

$$l(\boldsymbol{\beta};\boldsymbol{Y}) = \sum_{k=1}^{K}\sum_{j=1}^{m_k}\sum_{i=1}^{n_{kj}}\sum_{h=1}^{H_{kji}} \left\{ w_{kjih}y_{kjih}log\left(\frac{\pi_{kjih}}{1-\pi_{kjih}}\right) - log\left(\frac{1}{1-\pi_{kjih}}\right)\right\}. \tag{3.26}$$

Differentiating the log-likelihood with respect to unknown regression coefficients we obtain the vector of $p+1$ score equations which are compactly written as

$$\boldsymbol{X}'\boldsymbol{W}(\boldsymbol{y}-\boldsymbol{\pi}) = \boldsymbol{0} \tag{3.27}$$

where $\boldsymbol{X}$ is the $n \times (p+1)$ matrix of covariate values, $\boldsymbol{W}$ is a $n \times n$ diagonal matrix containing weights, $\boldsymbol{y}$ is the $n \times 1$ vector of observed outcome values and $\boldsymbol{\pi} = [\pi_{1111}, \ldots, \pi_{km_kn_{kj}H_{kji}}]'$ is the $n \times 1$ vector of logistic probabilities. The survey logistic regression model is given by

$$\text{logit}(\pi_{kjih}) = log\left\{\frac{\pi_{hjih}}{1-\pi_{kjih}}\right\} = \boldsymbol{X}'_{\boldsymbol{kjih}}\boldsymbol{\beta} \tag{3.28}$$

where,$\boldsymbol{X_{kjih}}$ is the vector that correspond to the characteristics of the h^{th} individual within i^{th} household within j^{th} primary sample unit nested within k^{th} stratum, and also $\boldsymbol{\beta}$ is the vector of unknown model coefficients. In the following section model selection and checking procedure are discussed.

3.9 Survey Logistic Model Selection and Checking

3.9.1 Model Selection

Variable Selection Procedures and Model Selection

In the survey logistic procedure in SAS, the variable selection procedures such as backward selection, forward selection, score and stepwise selection are not implemented. However, one can manually add or remove one variable in the model at a time by using the type 3 analysis of effects and observe the effect of the remaining variables. Type 3 analysis of effects are often used when the effect of one explanatory variable is influenced by the effect of another explanatory variable. One can remove variable the is not significant at a time and refit the model without that variable. This manual approach can be done until all remaining variables in the model are significant. The same model fitted in Chapter 3 is fitted using procedure PROC SURVEYLOGISTIC.

Testing Hypothesis about β

The computation of the standard errors of the parameter estimates used to construct confidence intervals and perform statistical tests is much complicated if data are from a complex design (Moeti, 2010). The estimate of the covariance matrix of the estimator of coefficients is given by

$$\widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}})} = (\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X})^{-1}\boldsymbol{S}(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X})^{-1} \tag{3.29}$$

where, $\boldsymbol{D} = \boldsymbol{W}\boldsymbol{V}$ is the $n \times n$ diagonal matrix with general elements

$$w_{kjih}\pi_{kjih}(1 - \pi_{kjih}).$$

The matrix $\boldsymbol{S}$ is a pooled within-stratum estimator of the covariance matrix in the left side of equation (3.27). Let us denote the general element of the vector of the score equation as $Z'_{kjih} = w_{kjih}\pi_{kjih}(1-\pi_{kjih})$. Thus

$$z_{kj} = \sum_{i=1}^{n_{kj}} z_{kjih}. \tag{3.30}$$

The stratum-specific mean is given by

$$\bar{z}_k = \frac{1}{m_k}\sum_{j=1}^{m_k} z_{kj}.$$

The within stratum estimator for the k^{th} stratum variance is given by

$$S_k = \frac{m_k}{M_k}\sum_{j=1}^{m_k}(z_{kj}-\bar{z}_k)(z_{kj}-\bar{z}_k)'. \tag{3.31}$$

The pooled estimator $S = \sum_{k=1}^{K}(1-f_k)S_k$. $(1-f_k)$, is the finite population correction factor and $f_k = \frac{m_k}{M_k}$ is the ratio of the number of sampling unit to the total number of primary sampling unit in the stratum k. Generally if M_k is unknown, one can assume that M_k is large enough so that f_k approaches zero, thus finite population correction factor will be 1 (Lemeshow and Hosmer, 2000). The Wald statistic for testing all coefficients in the fitted model are equal to zero is given by

$$\text{Wald} = \hat{\boldsymbol{\beta}}'[\widehat{var(\hat{\boldsymbol{\beta}})}_{p\times p}]^{-1}\hat{\boldsymbol{\beta}} \tag{3.32}$$

where,$\hat{\boldsymbol{\beta}}$ is the vector of p slope coefficients and $\widehat{var(\hat{\boldsymbol{\beta}})}_{p\times p}$ is the sub-matrix from a $(p+1)(p+1)$ matrix of $\widehat{var(\hat{\boldsymbol{\beta}})}$, and the p-value can be computed using χ^2 distribution with p degrees of freedom, that is to say,

$$\text{p-value} = P(\chi^2(p) \geqslant wald).$$

The SAS procedure PROC SURVEYLOGISTIC produces the covariance matrix of parameters through the Taylor expansion approximation procedure (Vittinghoff et al., 2011; **?**). This procedure estimates the variance in the variation between clusters and calculates the overall variance through pooling the stratum variance together. In this case, t-test statistics could be used for testing significance of the parameter estimates and constructs the confidence interval if the sample size is small. However, if the sample size is large, the sampling distribution of the parameter estimators are almost normally distributed (Lemeshow and Hosmer, 2000; **?**). The Wald statistics will be used to test and construct the confidence intervals given by

$$\hat{\beta}_j \pm Z_{1-\frac{\alpha}{2}}\sqrt{V_j} \tag{3.33}$$

where α is the level of significant, $z_{1-\frac{\alpha}{2}}$ is $100(1-\frac{\alpha}{2})$ percentile of the standard normal distribution and V_j is the variance obtained from the diagonal of the variance-covariance matrix. One can take the exponent of the confidence interval since it's on the logit scale. A similar hypothesis as that discussed in section 3.5 is tested here. In SAS the procedure PROC SURVEYLOGISTIC uses both Taylor expansion (linearization method) and maximum likelihood. There are other procedures such as Jackknife Repeated Replication(JRR) and Balanced Repeated Replication(BRR) can be used to estimate variance for each parameter. This procedure is used in this study to construct logistic regression model that account for the complex nature of the survey design. The odds ratio is still obtained as described earlier.

3.9.2 Model Checking

Model fit Test

The Hosmer-Lemeshow statistic is not produced in PROC SURVEYLOGISTIC. Since this statistic is not yet available, however can also, use Akaike's Information Criterion (AIC) and Schwarz Criterion (SC) to compare the goodness of fit(GOF) of two nested models (Moeti, 2010). The GOF test for logistic regression that is applied to complex survey data is obtained in the following manner: once the usual logistic regression model is fitted, the residuals $\hat{r_{ji}} = y_{ji} - \hat{\pi}(x_{ji})$ can be obtained. The GOF test is based on the residuals because of the large departures between observed and estimated value that indicates lack of fit (Hosmer and Lemeshow, 2004; Shackman, 2001; Archer et al., 2007). If we use grouping strategy, the observations are sorted into deciles based on their weight and estimates residuals. The survey estimates of sum of the residual by decile of risk $\hat{\boldsymbol{T}}' = (\hat{T_1}, \hat{T_2}, \ldots, \hat{T_{10}})$ are obtained such that the $\hat{T_g} = \sum_j \sum_i \bar{w}\hat{r}_{ji}$ $(g = 1, \ldots, 10)$. The associated estimated variance-covariance matrix $\hat{\boldsymbol{V}}(\hat{\boldsymbol{T}})$ is obtained using linearization. The linearization method can be used to construct an approximation to the functional form of the estimated population characteristics (two plus). In the first step, the functional form of the estimated population characteristics is approximated by a first order Taylor series, and the result is an approximation that is linear in the sample observation. The design based methods are used to estimate its variance. Using this method, the F-adjusted can be estimated as

$$F_{adjusted} = \frac{f - g + 2}{f_g} \hat{\boldsymbol{T}}' \boldsymbol{V}(\hat{\boldsymbol{T}})^{-1} \hat{\boldsymbol{T}} \tag{3.34}$$

where f is the number of sampled cluster minus the number of strata and g is the number of groups. We assume that the covariances are zero. The hypothesis being tested here is as follow H_0 : model is a good fit versus H_a : model, not a good fit. We compare the calculated $F_{adjusted}$ value with $F_{critical}$. We reject the null hypothesis if the $F_{adjusted}$ is greater than the $F_{critical}$ and we say a model is not a good fit.

Predictive Accuracy/Ability of the Model

In order to check for the predictive accuracy SAS procedure PROC SUVEYLOGISTIC produces for other statistics namely, Somer's D, Gamma, c and Tau-a. All these statistics ranges between 0 and 1. The larger value corresponds to a strong association between predicted and observed values. These measure of association are as discussed before.

3.10 Design Effects

3.10.1 Background

The sample size and sampling design determine the precision of the parameter estimates. Due to the practical constraints such as cost and manpower, the national survey would not adopt the simple random sampling (Shackman, 2001).The complex design would be adopted instead. The problem we face in complex sample design is that sampling errors for survey estimates can not be easily computed using the formulae found in statistical texts (Shackman, 2001). The loss of effectiveness in using complex instead of simple random sampling is known as design effects. The design effect is basically defined as the ratio of actual variance, under the sampling method actually used, to the variance computed under the assumption of simple random sampling (Shackman, 2001).The design effect is a technique that is widely used in survey sampling for planning a sample design in estimation and analysis (Park and Lee, 2001). One may use DEFF option in the model statement. PROC SURVEYLOGISTIC calculates the design effect for the regression coefficients.The design effect is given by

$$\text{DEFF} = \frac{\text{variance under the complex design}}{\text{variance under simple random sampling}}. \tag{3.35}$$

The denominator of equation(3.35) is computed under the assumption that the design is simple random sampling where we do not account for clustering, stratification, and weighting. One may compute the variance under the assumption of simple random sampling. If we consider both sampling weights and population total for the analysis, the sampling rates (or population total) under the assumption of simple random sampling: it is given by

$$f_{srs} = \frac{n}{w}$$

where n is the sample size and w estimates the population size. When the estimated population size is less than the sample size then f_{srs} is set to zero.

3.10.2 Design Effect Interpretation

The design effect (DEFF) may be used to compare variance under the assumption that data was obtained using simple random sampling and complex design. One can also use DEFT which is simply the square root of DEFF. The DEFT may be used to reduce variability since DEFT is less variable than DEFF. The DEFT can be used to estimate confidence interval directly (Shackman, 2001). The DEFT shows how much the standard error and confidence intervals increase. Suppose we have a value of DEFT equal to k, then we say confidence interval has to be k times as large as they would for a simple random sample.

3.11 Fitting the Survey Logistic Regression Model

Multiple logistic regression was fitted for the 2011-2012 Tanzania HIV/AIDS and Malaria Indicator Survey (THMIS) data using SAS. PROC SURVEYLOGISTIC was considered for this study to estimate parameter estimates, standard errors and odds ratios. A similar model as the one in subsection 3.6.2 was fitted and interpreted.

3.11.1 Model Selection

In PROC SURVEYLOGISTIC the option for variable selection that are associated with the outcome is not available. Since this option is not available, one has to manually add or remove one variable at a time in the model based on the results presented in Table 3.7 for type three analysis of effect, and fit the model again without the insignificant variable. The model which was fitted also involves two-way interaction effects which were found to be significant at 5% significant level as shown in Table 3.7.

Model Checking

The PROC SURVEYLOGISTIC in SAS does not produce plots and Hosmer-Lemeshow statistics, so one may use the AIC and SC to check if the model is a good fit or not. The AIC of the full model (contains intercept and other variables) is smaller compared

Table 3.7: Type 3 analysis of effects.

Type 3 analysis of effects			
Effect	**Degrees of Freedom**	**Wald Chi-square**	**p-value**
Breast feeding	1	32.0664	0.0001
HIV status	1	5.6423	0.0175
Birth order number	2	61.3484	0.0001
Breast feeding by Birth order number	2	14.0087	0.0009
Respondent age	2	0.5012	0.7783
Breast feeding by Respondent age	2	17.6328	0.0001
Children under five years	2	3.9799	0.1367
Children alive in a household	2	112.455	0.0001
Breast feeding by Children alive	2	15.3994	0.0005

to the AIC of the reduced model (contains only the intercept); this indicates that the fitted model better explains the data.

Table 3.8: Model fit statistics for survey logistic model.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
Akaike Information Criterion(AIC)	3.34E+09	2.73E+09
Schwarz Criterion(SC)	3.34E+09	2.73E+09
-2logLikelihood	3.34E+09	2.73E+09

Prediction Accuracy of the Model

In order to check how much the predicted probability agrees with outcome. One can use the receiver operating characteristics (ROC) curve. However, in PROC SURVEYLOGISTIC, the curve is not produced by the procedure but an association of predicted probabilities and the observed outcome can be produced. The concordance rate was 72.0% as shown in Table 3.9; this value tells us how good the model was in separating 0's and 1's. The value c=0.746 is the area under the ROC curve. The meaning of the area under curve ROC of 0.746 implies that 74.6 % of probabilities were predicted correctly by the model and shows that this model has good prediction accuracy. The Gamma statistic has a value of 0.521 indicates a moderate positive association between variables. The Somer's D statistic is 0.493 suggesting that not all pairs are Concordant.

Interpretation of the Coefficient of the Model and the Odds Ratio

Table 3.10 shows the estimated coefficients, standard errors, and p-value for the logistic regression model. The logit link is used all the time and calculated odds ratios and corre-

Table 3.9: Association of predicted probabilities and observed responses.

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	72.0	Somers' D	0.493
Percent Discordant	22.7	Gamma	0.521
Percent Tied	5.4	Tau-a	0.038
Pairs	4217640	c	0.746

sponding 95% confidence intervals are also shown. The effect of mother not breastfeeding was found to be positively associated with the under-five mortality (p-value=0.0001). The corresponding odds ratio was 2.067 (with 95% CI:1.6166-2.6419). The odds of death for a child from a mother who does not breastfeed were 2.067 times the odds of death for a child from a mother who does breastfeed. The effect of HIV status of a mother which was HIV-positive was found to be positively associated with the under-five mortality (p-value=0.0083). The corresponding odds ratio was 1.313 (with 95% CI:1.0724-1.6065). The odds of death for a child from a mother who is HIV-positive were 1.313 times the odds of death for a child from a mother who is HIV-negative. The effect child birth order number that is above four was found to be positively associated with under-five mortality (p-value=0.0001). The corresponding odds ratio was 3.523 (with 95% CI:2.4609-5.0444). The odds of death for a child whose birth order number is above four were 3.523 times the odds of death for a child whose birth order number is not more than one. The effect of the mother with number of children alive which is is more than four was found to be negatively associated with under-five mortality (p-value=0.0001). The corresponding odds ratio was 0.191 (with 95% CI:0.1222-0.2991). The odds of death for a child from a mother with more than four children alive were 0.191 times the odds of death for a child from a mother with less than two children alive. The effect of childbirth order number above four depends on not breastfeeding and was found to be positively associated with under-five mortality (p-value=0.0088). The corresponding odds ratio was 1.613 (with 95% CI:1.1277-2.3080). The odds of death for a child whose birth order number is above four and from a mother who does not breastfeed were 1.613 times the odds of death for a child whose birth order number is less than two and from a mother who breastfeed. The effect of mother's age from 20 to 34 years depends on not breastfeeding and was found to be negatively associated with under-five (p-value=0.0007). The corresponding odds ratio was 0.624 (with 95% CI:0.4749-0.8208). The odds of death for a child from a mother with age between 20 and 34 years who does not breastfeed were 0.624 times the odds of death for a child from a mother with age over 34 years and breastfeed. The effect of mother's age less than 20 years depends on not breastfeeding and was found to be positively associated with the under-five mortality (p-value=0.0001). The corresponding odds ratio was 2.39 (with 95% CI:1.5337-3.7253). The odds of death for a child from a mother with age less than 20 years and does not breastfeed were 2.39 times the odds of death for a child from a mother with age over 34 years and does breastfeed.

Table 3.10: Survey logistic regression model coefficients, standard errors and odds ratios.

Analysis of Maximum Likelihood Estimates				**95% confidence interval**		
Effects	**Estimate**	**Standard error**	**P-value**	**Odds ratio**	**lower**	**upper**
		Socio-demographic characteristics				
Intercept	-3.5209	0.1873	0.0001			
Breastfeeding(BF)						
Yes(reference)						
No	0.7259	0.1253	0.0001	2.067	1.6166	2.6419
HIV Status(HS)						
Negative(reference)						
Positive	0.272	0.1031	0.0083	1.313	1.0724	1.6065
Birth Order Number(BON)						
Less than 2 births(reference)						
2-4 births	-0.1911	0.1278	0.1347	0.826	0.6430	1.0612
above 4 births	1.2594	0.1831	0.0001	3.523	2.4609	5.0444
Respondent Age(MA)						
Over 34 years(reference)						
20-34 years	0.0979	0.1372	0.4756	1.103	0.8428	1.4431
Less than 20 years	-0.0729	0.2308	0.7521	0.93	0.5914	1.4615
		Socio-economic characteristics				
Children five and under(C5)						
Less than 2 children(reference)						
2-4 children	0.043	0.1772	0.8084	1.044	0.7376	1.4774
over 4 children	-0.3271	0.2727	0.2303	0.721	0.4225	1.2305
Number of children alive(CL)						
Less than 2 children(reference)						
5 or more children	-1.6547	0.2284	0.0001	0.191	0.1222	0.2991
2-4 children	-0.067	0.1341	0.6173	0.935	0.7190	1.2163
		Interaction between Socio-demographic and Socio-economic characteristics				
Breastfeeding by birth order number						
Yes by less than 2(reference)						
No by 2-4	0.0744	0.1265	0.5562	1.077	0.8407	1.3804
No by above 4	0.4783	0.1827	0.0088	1.613	1.1277	2.3080
Breastfeeding by Number of children alive						
Yes by less than 2(reference)						
No by 5 or more children	-0.3613	0.2222	0.104	0.697	0.4508	1.0770
No by 2-4 children	-0.2137	0.1335	0.1093	0.808	0.6217	1.0491
Breastfeeding by Respondent age						
Yes by over 34 years(reference)						
No by 20-34 years	-0.4711	0.1396	0.0007	0.624	0.4749	0.8208
No by less than 20 years	0.8714	0.2264	0.0001	2.39	1.5337	3.7253

3.12 Comparison of Logistic and Survey Logistic Regression

Tables 3.6 and 3.10 contains odds ratios and confidence interval for logistic and survey logistic regression respectively. Since the sample was not from the simple random sample, the parameter estimates for both models are not the same. However, they are closer to one another. One of the assumptions for logistic regression is that the observation are independent, but for complex design this assumption is violated thus a better model may be the one fitted using PROC SURVEYLOGISTIC since it accounts for the complexity of the design. The models fitted by both methods produce the areas under the curve which are between 0.7 and 0.8. This suggests that both models had good prediction accuracy. Table 3.11 shows the DEFF and DEFT which is the square root of DEFF for each estimated coefficient. The effect of breastfeeding has the DEFF value of 1.1758 and DEFT value of 1.0843. The standard error and confidence interval are 1.0843 times as larger as they would be for simple random sampling. The effect of mothers HIV status which is positively associated with the under-five mortality has DEFF=1.4929 and DEFT=1.2219. The standard error and confidence interval are 1.2219 times as large as they would be for simple random sampling. The effect of childbirth order number above four which is positively associated with the under-five mortality has DEFF=1.3977 and DEFT=1.1822. The standard errors and confidence interval are 1.1822 times large as they would be for simple random sampling. The effect the number of children alive that is more than four is negatively associated with under-five mortality has the DEFF=1.342 and DEFT=1.1584. The standard error and confidence interval have to be 1.1584 times as large as they would be for simple random sampling. The effects for the number of children alive which is less than two is positively associated with the under-five mortality has the DEFF=1.4458 and DEFT=1.2021. The standard error and confidence interval have to be 1.2021 times as large as they would be for simple random sampling. The effect of birth order number above four depends on whether the mother does breastfeed with DEFF=1.7086 and DEFT=1.3071. The standard error and confidence interval have to be 1.3071 times as large as they would be for simple random sampling. The effect of mother's age between 20 to 34 years depends on whether the mother does breastfeed with DEFF=1.1989 and DEFT=1.0949. The standard errors and confidence interval have to be 1.0949 times as large as they would be for simple random sampling. The effect of mother's age over 34 years depends on whether the mother does breastfeed has DEFF=1.344 and DEFT=1.1593. The standard error and confidence interval have to be 1.1593 times as large as they would be for simple random sampling. The effect of the number of children alive less than two depends on whether the mother does breastfeed with DEFF=1.8466 and DEFT=1.3689. The standard error and confidence interval have to be 1.3689 times as large as they would be for simple random sampling. We observe that the design effects values are above one this tell us that variance was under-estimated while using logistic regression model were smaller compared to those computed while using complex design. This confirm that standard errors are larger under survey logistic. This shows that there was an under-estimation of variance while using logistic regression assuming that data was sampled using SRS. Hence, using the model like survey logistic regression is good since it takes into account of survey design features.

Table 3.11: Survey logistic, coefficients, standard errors, p-values, odds ratios, confidence interval and design effects.

Analysis of Maximum Likelihood Estimates				95% confidence interval			Design Effects	
Effects	**Estimate**	**Standard error**	**P-value**	**Odds ratio**	**lower**	**upper**	**DEFF**	**DEFT**
Socio-demographic characteristics								
Intercept	**-3.5209**	0.1873	0.0001				1.9613	1.4005
Breastfeeding(BF)								
No(reference)								
Yes	**-0.7259**	0.1253	0.0001	0.484	0.3785	0.6186	1.1758	1.0843
HIV Status(HS)								
Negative(reference)								
Positive	**0.272**	0.1031	0.0083	1.313	1.0724	1.6065	1.4929	1.2219
Birth Order Number(BON)								
Less than 2 births(reference)								
2-4 births	-0.1911	0.1278	0.1347	0.826	0.6430	1.0612	1.7569	1.3255
above 4 births	**1.2594**	0.1831	0.0001	3.523	2.4609	5.0444	1.3977	1.1822
Respondent Age(MA)								
less than 20 years(reference)								
20-34 years	0.0979	0.1372	0.4756	1.103	0.8428	1.4431	2.0312	1.4252
over 34 years	-0.0729	0.2308	0.7521	0.93	0.5914	1.4615	1.6249	1.2747
Socio-economic characteristics								
Children five and under(C5)								
Less than 2 children(reference)								
2-4 children	0.043	0.1772	0.8084	1.044	0.7376	1.4774	1.55	1.245
over 4 children	-0.3271	0.2727	0.2303	0.721	0.4225	1.2305	1.5883	1.2603
Number of children alive(CL)								
2-4 children(reference)								
5 or more children	**-1.6547**	0.2284	0.0001	0.191	0.1222	0.2991	1.342	1.1584
Less than 2 children children	**1.7217**	0.1709	0.0001	5.594	4.0018	7.8198	1.4458	1.2024
Interaction between Socio-demographic and Socio-economic characteristics								
Breastfeeding by Number of children alive								
2-4 children(reference)								
Yes by 5 or more children	0.3613	0.2222	0.104	1.435	0.9285	2.2185	1.344	1.1593
Yes by less 2 children children	**-0.575**	0.1672	0.0006	0.563	0.4055	0.7809	1.8466	1.3689
Breastfeeding by Respondent age								
No by less than 20 years(reference)								
Yes by 20-34 years	**0.4711**	0.1396	0.0007	1.602	1.2183	2.1058	1.1989	1.0949
Yes by over 34 years	**0.4003**	0.199	0.0442	1.492	1.0103	2.2041	1.344	1.1593
Breastfeeding by birth order number								
No by less than 2(reference)								
Yes by 2-4	-0.0744	0.1265	0.5562	0.928	0.7245	1.1895	1.4312	1.1964
Yes by above 4	**-0.4783**	0.1827	0.0088	0.62	0.4333	0.8867	1.7086	1.3071

3.13 Limitations of Survey Logistic Regression

Despite the fact that the survey logistic account for the complexity of the survey design. It may present some limitations due to unavailability of Hosmer-Lemeshow test. We may not be able to test if the model is a good fit or not a good fit. The variable selection procedures are not available thus one is required to select variable manually which can be time-consuming when many variables are involved, and possible errors may occur while choosing variables. The model has to be chosen based on the AIC and SC both of which introduce a penalty to the -2log-likelihood of having many parameters. Since they both have -2logL term in their formulation, they are used only in the case of ungrouped data (Lemeshow and Hosmer, 2000).

Chapter 4

Generalized Linear Mixed Models

4.1 Introduction

The Generalized Linear Models (GLMs) discussed in Chapter 3 might not be appropriate for the data of interest. In GLMs under which the logistic regression falls, the complexity of the survey design is ignored in the sense that the random effect on child survival status is ignored. The inclusion of random effects in the analysis results into generalized linear mixed models (GLMMs). These models are powerful since they combine features of both linear mixed models (including both fixed effects and random effects) and generalized linear models, such that they handle a wide range of response distributions and data with observations sampled in some group structure instead of completely independent (Molenberghs and Verbeke, 2006; Waagepetersen, 2007). GLM allows modeling of different kind of responses such a binary (McCullagh and Nelder, 1989). The models that incorporate random effects are known as linear mixed models (LMMs). In order to make a valid statistical inference, one has to account for subject-specific effects. The subject-specific effects in the studies with natural occurring groups (i.e. responses collected from members same group/family tends to be more similar). In this section, the theory of linear mixed models is reviewed. The theory of generalized linear mixed model is outlined and is utilized in modeling the data of interest.

4.2 Review of Linear Mixed Models

The generalized linear model discussed in Chapter 3 do not account for the random effect. Instead, it is necessary to expand the model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4.1}$$

to become.

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{U} + \boldsymbol{\epsilon} \tag{4.2}$$

Y is the $n \times 1$ vector of responses,

where $\boldsymbol{X}$ is a $n \times (p+1)$ design matrix for fixed effects,

$\boldsymbol{\beta}$ is a $(p+1) \times 1$ vector of unknown fixed effects parameters,

$\boldsymbol{Z}$ is a $n \times q$ design matrix for random effects,

$\boldsymbol{U}$ is a $q \times 1$ vector of unknown random effects parameters, and

$\boldsymbol{\epsilon}$ is a $n \times 1$ vector of error term which have multivariate normal distribution with mean vector $\mathbf{0}$ and variance-covariance matrix $\mathbf{R}$ i.e. $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \boldsymbol{R})$. Given the nature random effect hypothesis, $\boldsymbol{U}$ is treated differently from $\boldsymbol{\beta}$. Statistical linear mixed models state that observed data consist of two parts, that is, random and fixed effects (Littell et al., 2000). We define fixed effects as the expected value of the observation and random effects is defined as variance and covariance of the observation. we may assume that observations on the same unit are correlated. Hence, Linear mixed models address the issue of covariation between measures on the same unit (Kincaid, 2005; Littell et al., 2000). Representing variance of the model as $V(\boldsymbol{y})$ shown in equation (4.3) is known as Modelling covariance structure. It is modelled as a function of relatively small number of parameters (Littell et al., 2000). The specification of the covariance structure for mixed model is done through $\boldsymbol{G}$ and $\boldsymbol{R}$ as.

$$V(\boldsymbol{Y}) = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}' + \boldsymbol{R} \tag{4.3}$$

where $\boldsymbol{ZGZ'}$ represents the between subject portion of the covariance structure and $\boldsymbol{R}$ represents within subject portion. In linear mixed models with more than one random effects, the random effects are assumed to come from a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{G}$. The random effect can be predicted and not estimated. The variance components are estimated instead. The diagonal elements of matrix $\boldsymbol{G}$ is the variance component for each random effect while off-diagonal elements are covariances that exist between different dimensions. Suppose that there is one random effect in the model, then G will have only one element that is the variance component of random effects. If they are more than one random effects, $\boldsymbol{G}$ will be a $k \times k$ for k random effect. Suppose $k = 3$ random, we present five different covariance structures in the Table 4.1 and discuss them. Table 4.1 shows the list of covariance structures which

Table 4.1: List of simpler covariance structures.

Structure	Description	Number of parameters	i,jth element
AR(1)	Autoregressive lag 1	2	$\sigma_{ij} = \sigma^2 \rho^{\|i-j\|}$
CS	Compound Symmetry	2	$\sigma_{ij} = \sigma_1 + \sigma^2 1(i = j)$
UN	Unstructured	t(t+1)/2	$\sigma_{ij} = \sigma_{ij}$
TOEP	Toeplitz	t	$\sigma_{ij} = \sigma_{\|i-j\|+1}$
VC	Variance Component	q	$\sigma_{ij} = \sigma_k^2 1(i = j)$

can be modeled in SAS using PROC MIXED procedure. We firstly look at covariance structure known to be simple.

Simple or Variance Component(VC)

The variance component structure is the standard variance components and is the default structure if the random or repeated statement is not used in SAS.

$$VC = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix}.$$

Compound Symmetry(CS)

The Compound Symmetry structure is often required for split-plot design. The variances are homogenous for this covariance structure. There is a correlation between two measurements and we may assume that the correlation is constant regardless of the distance between two measurements (Kincaid, 2005; Littell et al., 2000).

$$CS = \sigma^2 \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}.$$

Autoregressive Lag 1(AR(1)

The covariance structure known as autoregressive has homogenous variances and correlation decline exponentially with distance. This means that two measurements that are right next to each other in time are considered to be correlated. However, as measurements get further apart they are less correlated (Kincaid, 2005; Littell et al., 2000). This structure is applicable for evenly spaced time interval for repeated measures.

$$AR(1) = \sigma^2 \begin{pmatrix} 1 & \rho^1 & \rho^2 \\ \rho^1 & 1 & \rho^1 \\ \rho^2 & \rho^1 & 1 \end{pmatrix}.$$

Toeplitz(TOEP)

The banded structure, also known as Toeplitz, specifies that covariance depends only on lag, but not as a mathematical function with smaller number of parameters. Toeplitz structure is similar to the autoregressive (AR(1)) in that all measurement next to each other have the same correlation measurements which are two apart have same correlation different from the first. However, the correlations do not necessarily have the same pattern. The AR(1) is basically a special case of Toeplitzitep(Kincaid, 2005).

$$TOEP = \begin{pmatrix} \sigma^2 & \sigma_1 & \sigma_2 \\ \sigma_1 & \sigma^2 & \sigma_1 \\ \sigma_2 & \sigma_1 & \sigma^2 \end{pmatrix}.$$

Unstructured(UN)

The Unstructured covariance structure specifies no pattern in the covariance matrix, and completely general. The generality of this structure has drawback for having a large number of parameter to be tested(Kincaid, 2005).

$$UN = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{pmatrix}.$$

The assumptions made for generalized linear models (GLMs) are retained in GLMMs. It is possible to have a variable that appears in both $\boldsymbol{X}$ and $\boldsymbol{Z}$, in this case the fixed effect is an average across all levels of random effects. In the latter case, the estimate is the amount of variance in the effect between levels. If $\boldsymbol{X}$ contains a single column of ones, then this lead to the random intercept model. If $\boldsymbol{X}$ contains an extra column, then this is known as the random slope model. However, the draw back for this model is that it requires the responses to be normally distributed. The models which accommodate normal and non-normal data in which they are a member of exponential family of distributions known as generalized linear mixed models (GLMMs)(McCullagh and Nelder, 1989). The linear mixed model can be viewed as a special case of the generalized linear mixed model (GLMMs).

4.3 Generalized Linear Mixed Models

Generalized linear mixed models are an extension of linear mixed models with a relaxation of some of the assumptions of LMMs. GLMMs provides all advantages of a logistic regression such as information on a sample size, they are able to do one analysis with all random effects on it, and they accommodate the binary response variable. Furthermore, the advantage of GLMMs is its ability to handle unbalanced data due to missing observations and its ability to account for correlated data(Manning, 2007). The observation in the data used might be correlated since clusters were drawn. In Chapter 3 the linear predictor for the generalized linear model is $\boldsymbol{\eta} = \boldsymbol{X\beta}$.

4.3.1 Model Formulation

Suppose we now relax the normality assumption of $f(\boldsymbol{Y} \mid \boldsymbol{\theta})$. It can be assumed that $\boldsymbol{Y}$ and $\boldsymbol{\theta}$ are independent and $f(\boldsymbol{Y} \mid \boldsymbol{\theta})$ is the member of exponential family of distribution (McCullagh and Nelder, 1989).

$$f(\boldsymbol{Y} \mid \boldsymbol{\theta}) = \exp\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} - c(y_i, \phi)\} \tag{4.4}$$

where ϕ is the scale parameter. Based on the model the conditional y related to θ_i is given by

$$E(\boldsymbol{y} \mid \boldsymbol{\theta}) = \frac{\partial b(\theta_i)}{\partial \theta_i}.$$

The model with both random and fixed effects is given by.

$$g(\theta_i) = \boldsymbol{X_i'\beta} + \boldsymbol{Z_i'U_i} \tag{4.5}$$

where, $\eta_i = g(\theta_i)$, g is the link function and $\boldsymbol{U}_i$ is a vector of random effects. In this study survival status is either 0 (child alive) or a 1(child not alive). Thus we use the logistic regression where we consider $g(.)$ as the logit link, with X_i and $Z_i (i = 1, 2, \ldots, n)$ being p-dimension and q-dimension a vectors of known covariates values, while β is a p-dimension vector of unknown fixed effects regression coefficient.

4.4 Maximum Likelihood Estimation

In linear mixed models the marginal distribution of Y could be computed as the multivariate normal, meaning $f(Y)$ is a density function of a multivariate normal distribution. However, for generalized linear mixed models, it is difficult to evaluate the integral because of the presence of N q-dimensional integral over the random effects (Vittinghoff et al., 2011; Bolker et al., 2009). The random effect model could be fitted by maximization of marginal likelihood, and that is obtained by integrating out the random effects. The likelihood is given by

$$\begin{aligned} L(\boldsymbol{\beta}, G, \phi) &= \prod_{i=1}^{N} f_i(\boldsymbol{Y_i} \mid \boldsymbol{\beta}, \boldsymbol{G}, \boldsymbol{\phi}) \\ &= \prod_{i=1}^{N} \int f_i(\boldsymbol{Y_i} \mid \boldsymbol{\beta}, \boldsymbol{G}, \boldsymbol{\phi}).f(\boldsymbol{U_i}, \boldsymbol{G})\boldsymbol{du_i} \end{aligned} \tag{4.6}$$

where,$f_i(\boldsymbol{Y_i} \mid \boldsymbol{\beta}, \boldsymbol{G}, \boldsymbol{\phi}) = \int \prod_{j=1}^{n_i} f_{ij}(\boldsymbol{Y_{ij}} \mid \boldsymbol{\beta}, \boldsymbol{G}, \boldsymbol{\phi}).f(\boldsymbol{U_i}, \boldsymbol{G})\boldsymbol{du_i}$(Molenberghs and Verbeke, 2006). In general, numerical approximations have to be used to evaluate likelihood of GLMMs.

4.4.1 Estimation: Approximation of the Integrand

The Laplace method is one of the approaches of approximating the integrand and is one of the natural alternatives when exact the likelihood function is difficult to compute (Molenberghs and Verbeke, 2006). When the integrands are approximated, the objective is to obtain traceable integrals such that closed form expressions can be obtained which make numerical maximization of the approximated likelihood feasible (Molenberghs and Verbeke, 2006). Suppose we wish to approximate the integral of the form

$$I = \int \exp^{(-q(x))}\, d\boldsymbol{x} \tag{4.7}$$

where $q(.)$ is a well behaved function in a way that its minimum value is at x = x̃ with $q^{'}(\tilde{\mathrm{x}}) = 0$ and $q^{''}(\tilde{x}) > 0$. we can consider the Taylor expansion about $\tilde{x}$ given by

$$q(\mathbf{x}) \approx q(\tilde{\mathrm{x}}) + \frac{1}{2}q^{''}(\tilde{\mathrm{x}})(\mathrm{x} - \tilde{\mathrm{x}}) + \ldots. \tag{4.8}$$

This gives an approximation to (4.7) as

$$\int \exp\left(-q(x)\right)dx \approx \sqrt{\frac{2\pi}{q^{''}(\tilde{\mathrm{x}})}} \exp\left(-q(\tilde{\mathrm{x}})\right). \tag{4.9}$$

We may also have the multivariate extension of (4.9), which is often useful. Let $q(\alpha)$ be a well behaved function with its minimum at $\alpha = \tilde{\alpha}$ with $q^{'}(\tilde{\alpha}) = 0$ and $q^{''}(\tilde{\alpha}) > 0$, where $q^{'}$ and $q^{''}$ are the gradient and Hessian of q respectively. We have

$$\int \exp\left(-q(x)\right)dx \approx c \mid q^{''}(\tilde{x}) \mid^{-\frac{1}{2}} \exp\left(-q(\tilde{x})\right) \tag{4.10}$$

where c is a constant depending on the dimension of the integral and $\mid q^{''}(\tilde{x}) \mid$ is the determinant of matrix $q^{''}(\tilde{x})$. In which $q^{''}(\tilde{x}) > 0$ implies matrix $q^{''}(\tilde{x})$ is positive definite.

4.4.2 Estimation: Approximate of Data

There is another class of estimation approach based on a decomposition of the data into mean and error terms. With the Taylor series expansion of the mean which is a non-linear function of predictors. The method in this class differs in the order of the Taylor approximation. The decomposition that is considered is

$$Y_{ij} = \mu_{ij} + \epsilon_{ij} = h(X_{ij}^{'}\boldsymbol{\beta} + Z_{ij}^{'}\boldsymbol{U}) + \epsilon_{ij} \tag{4.11}$$

where, $h(.)$ is the inverse link function, and error term have an appropriate distribution with variance equal to $\mathrm{var}(Y_{ij} \mid U_i) = \phi V(\mu_{ij})$. Here, $V(.)$ is the usual variance function in the exponential family (Molenberghs and Verbeke, 2006). Consider a binary outcome with logit link function. One then has

$$\mu_{ij} = h(X_{ij}'\boldsymbol{\beta} + Z_{ij}'\boldsymbol{U}) = P_{ij} = \frac{\exp(X_{ij}'\boldsymbol{\beta} + Z_{ij}'\boldsymbol{U})}{1 + \exp(X_{ij}'\boldsymbol{\beta} + Z_{ij}'\boldsymbol{U})} \tag{4.12}$$

where $h(X_{ij}'\boldsymbol{\beta} + Z_{ij}'\boldsymbol{U})$ is the inverse for the logit link function which is the logistic function. x_i and z_i are as in the definition of generalized linear mixed model. This is considered as the special case of GLMM where the exponential the family is Bernoulli and corresponding link function is $g(\mu) = logit(\mu)$.

4.4.3 Penalized Quasi-Likelihood

The Penalized Quasi-Likelihood (PQL) is one of the methods that approximates data by mean plus error term with variance equals to $\mathrm{Var}(Y_{ij} \mid U_i\ )$. This method uses Taylor expansion around estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{U}}$ of fixed effects and random effects respectively (Bolker et al., 2009; Moeti, 2010). One then has

$$\begin{aligned} Y_{ij} &= \mu_{ij} + \epsilon_{ij} = h(X_{ij}'\boldsymbol{\beta} + Z_{ij}'\boldsymbol{U}) + \epsilon_{ij} \\ &\approx h(X_{ij}'\hat{\boldsymbol{\beta}} + Z_{ij}'\hat{\boldsymbol{U}}) + h(X_{ij}'\hat{\boldsymbol{\beta}} + Z_{ij}'\hat{\boldsymbol{U}})X_{ij}'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + h(X_{ij}'\hat{\boldsymbol{\beta}} + Z_{ij}'\hat{\boldsymbol{U}})Z_{ij}'(\boldsymbol{U} - \hat{\boldsymbol{U}}) + \epsilon_{ij} \\ &= \hat{\mu}_{ij}V(\hat{\mu}_{ij})X_{ij}'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + V(\hat{\mu}_{ij})Z_{ij}'(\boldsymbol{U} - \hat{\boldsymbol{U}}) + \epsilon_{ij}, \end{aligned} \tag{4.13}$$

and

$$\boldsymbol{Y_i} = \boldsymbol{\hat{\mu}_i} + \boldsymbol{\hat{V}_i X_i}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \boldsymbol{\hat{V}_i Z_i}((U) - \hat{\boldsymbol{U}}) + \boldsymbol{\epsilon_i}$$

where $\boldsymbol{\hat{\mu}_i}$ contains values of $\boldsymbol{\hat{\mu}_{ij}} = h(X_{ij}'\hat{\boldsymbol{\beta}} + Z_{ij}'\hat{\boldsymbol{U}})$, $\boldsymbol{V_i}$ is the diagonal matrix with elements $V(\hat{\mu}_{ij}) = h(X_{ij}'\hat{\boldsymbol{\beta}} + Z_{ij}'\hat{\boldsymbol{U}})$ and $\boldsymbol{X_i}$ and $\boldsymbol{Z_i}$ contain the X_{ij}' and Z_{ij}' respectively. Re-ordering the above expression and pre-multiply with $\widehat{\boldsymbol{V_i^{-1}}}$ we obtain

$$\begin{aligned} Y_i^* &= \widehat{\boldsymbol{V_i^{-1}}}(\boldsymbol{Y_i} - \boldsymbol{\hat{\mu}_i}) + \boldsymbol{X_i}\hat{\boldsymbol{\beta}} + \boldsymbol{Z_i}\hat{\boldsymbol{U}} \\ &\approx \boldsymbol{X_i}\hat{\boldsymbol{\beta}} + \boldsymbol{Z_i}\hat{\boldsymbol{U}} + \boldsymbol{\epsilon_i^*}. \end{aligned} \tag{4.14}$$

For $\boldsymbol{\epsilon_i^*}$ equal to $\hat{\boldsymbol{V_i^{-1}}}\boldsymbol{\epsilon_i}$ and has a zero mean. This can be viewed as a linear mixed model for a pseudo data $\boldsymbol{Y_i^*}$ with error term $\boldsymbol{\epsilon_i^*}$. This gives the algorithm for fitting original generalized linear mixed models.

Algorithm

Step 1: Given starting value for parameter β, ϕ and G. In the marginal likelihood empirical Bayes estimates are calculated for $\boldsymbol{U_i}$ and pseudo data $\boldsymbol{Y_i^*}$ are computed.
Step 2: Approximate linear mixed model is fitted, which gives updated estimates for β, ϕ and G. then updated estimates are used to update the pseudo data. This whole scheme is iterated until convergence is reached, and resulting estimates are called penalized quasi-likelihood estimate. They are obtained from optimizing a quasi-likelihood function that involves first and second order conditional moments, augmented with a penalty term on the random effects (Molenberghs and Verbeke, 2006).

4.4.4 Marginal Quasi-Likelihood

Marginal Quasi-Likelihood (MQL) is an approximation method which is similar to PQL method. However, it is based on a linear Taylor expansion of the mean around current estimate $\hat{\boldsymbol{\beta}}$ for fixed effects, and around $\boldsymbol{U} = \boldsymbol{0}$ for random effects (Bolker et al., 2009; Moeti, 2010). This gives same expansion as shown for PQL, but now the current predictor is of the form $h(X_{ij}'\hat{\boldsymbol{\beta}})$. The pseudo-data are now of the form

$$\boldsymbol{Y_i^*} = \hat{\boldsymbol{V_i^{-1}}}(Y_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{i}}) + \boldsymbol{X_i}\hat{\boldsymbol{\beta}} \tag{4.15}$$

and satisfy the approximate linear mixed model

$$\boldsymbol{Y_i^*} \approx \boldsymbol{X_i\beta} + \boldsymbol{Z_i U} + \boldsymbol{\epsilon_i^*}. \tag{4.16}$$

The model fitting is also done by iteration between the calculation of the pseudo data and fitting of approximate linear mixed model for these pseudo data (Molenberghs and Verbeke, 2006).. The resulting estimates are known as quasi-likelihood estimates (MQL).

4.4.5 Discussion of MQL and PQL

There is no main difference between penalized quasi-likelihood (PQL) and marginal quasi-likelihood (MQL); they both do not incorporate the random $\boldsymbol{U_i}$ in the linear predictor (Bolker et al., 2009). Both of these methods are based on similar ideas and will have almost similar properties. However, the accuracy of both models depends on the accuracy of the linear mixed model for pseudo data $\boldsymbol{Y_i^*}$. The Laplace method, PQL and MQL perform poorly in the cases of binary with repeated observations small number of repeated observations available (Molenberghs and Verbeke, 2006). The MQL completely ignores the random effects variability in linearization of the mean. The Laplace method is more accurate than penalized quasi-likelihood. However, Laplace is slower and less flexible compared to penalized quasi-likelihood (Bolker et al., 2009). The MQL remains biased while PQL will be consistent with an increased number of measurements.

4.5 Generalized Linear Mixed Models (GLMMs) in SAS

The Statistical Analysis Software (SAS) procedure PROC GLIMMIX accommodates features of GLMMs. This procedure combines both procedures namely PROC GENMOD and PROC MIXED. The estimation of the parameter estimates utilizing this procedure follows likelihood based techniques; the default is pseudo-likelihood procedure (Moeti, 2010). The procedure allows one to change estimation method and specify covariance structures. The construction of the Wald test statistics and confidence intervals for the estimates depends on Taylor series expansion method. The Wald-type tests together with the estimate variance-covariance matrix are used for hypothesis test for fixed effects.

4.6 Interpretation of Generalized Linear Mixed Model Parameter

Just like in GLMs the model parameter can be interpreted once they are obtained. The example is given below on how to interpret the result (estimates). Consider the instance where the binary random variables and logit link is as follows

$$\text{logit}(\pi_{ij}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{2i} X_{3i}.$$

Let X_{1i} be a continuous predictor variable, X_{2i} be a categorical variable with 2 categories (i.e. gender) and X_{3i} be categorical variable with 2 categories for place of residence (for individual i and cluster j). Parameter β_1 is the increase in the log odds of event given a 1 unit increase in X_{1i}. Parameter β_2 is the increase in the log odds of the event comparing two individuals with different genders but with the same value of u_i. The parameter β_3 is the increase in the log odds of the event comparing two individuals with different categories (i.e. type of place of residence). The parameter β_4 is the increase in log odds of the event comparing two individuals with different gender by different categories (a type of place of residence i.e. rural or urban).

4.7 Application of Generalized Linear Mixed Model

The GLIMMIX procedure fits statistical models to data with correlations or non-constant variability. In this procedure, the response does not need to be normally distributed and allows different estimation methods to be specified (i.e. Laplace). The model was fitted three times, each time specifying different estimation methods discussed earlier. The random effect were the clusters. Another approach for interpreting the model parameters used in this section is known as pairwise comparisons of least-square means. The pairwise comparison of the least-square means for interaction effects is performed. The Diffogram which displays a line for each comparison and axes of the plot represents the scale of the least-square means (Moeti, 2010). The 45-degree line is the reference line of the plot. In the analysis of means with Nelson-Hsu Adjustment, dashed horizontal step plots represent lower and upper decision limit determined at 95^{th} percentile. If the level is significantly different from the average, then the corresponding vertical line crosses the decision limits. The results obtained are presented and interpreted in Table 4.2. The type

Table 4.2: Type 3 tests of fixed effects.

Type 3 Tests of Fixed effects				
Effect	**Num DF**	**Den DF**	**Wald Chi-square**	**p-value**
Breast feeding	9866	1	43.13	0.0001
HIV status	9866	1	7.18	0.0074
Birth order number	9866	2	29.51	0.0001
Breast feeding by Birth order number	9866	2	8.27	0.0003
Respondent age	9866	2	0.93	0.3937
Breast feeding by Respondent age	9866	2	9.42	0.0001
Children under five years	9866	2	4.17	0.0155
Children alive in a household	9866	2	60.57	0.0001
Breast feeding by Children alive	9866	2	7.53	0.0005

3 tests of fixed effects for the model fitted using Laplace method in GLMMs is shown in Table 4.2. The F-statistics which is used for the significant test for the fixed effects and corresponding p-value shows that all effects are important in the fitted model when tested at 5% level of significance. Only mother's age is not significant (p-value=0.3937). However, due to the hierarchical principal for the model with interaction effects which are significant, the main effect is retained in the model. The residual log pseudo-likelihood of the fitted model is given by 67122.62 and generalized chi-square statistics is 9671.98. The ratio of chi-square statistics to its degree of freedom which is the measure of variability in the marginal distribution of the data is 0.92. The variance of the random effect is estimated as $\sigma_u^2 = 0.06780$ given in Table 4.3 if the PQL method is used.

Table 4.3: Random effect and model information.

Random Effects			
	Laplace Estimate (S.E)	**Marginal Estimate (S.E)**	**Penalized Estimate (S.E)**
Variance(Intercept)	0.1063 (0.1028)	0.07419 (0.0855)	0.06780 (0.0815)
Model Information			
Number of parameters	11	11	11
-2loglikelihood	3144.74	67680.29	67122.62
AIC	3180.74		

Table 4.4, Table 4.5 and Table C.2 in Appendix C2 shows the solution for fixed effects. Estimated parameters, standard errors and fit statistics obtained from the MQL, PQL and Laplace in GLMMs for the fitted models are shown respectively in these tables. The fitted models are the random intercept models; one can observe that standard errors are smaller than those in the model in section 3.6.2. The parameters estimated for the model fitted using Laplace, MQL and PQL are almost the same, and parameter are found to be significant across all three methods shows consistency. The coefficients for fixed effects are interpreted in the same way as in the ordinary logistic regression model. The estimates are slightly lower than those in section 3.6.2; this is due to the fact that this model accounts for the random effects. The conclusion remains the same so the interpretation of the coefficients will also be done explicitly. In this section, we also use another form of presentation based on least-square means analysis for graphical and tabular then interpret odds ratios as before. Here the contrast is done on the logit scale.

Figure 4.1 illustrates adjusted comparison of breastfeeding by mother's age interaction least-square means for multiplicity. The lines that represent the significant difference between the least-square means of the level of breastfeeding by mother's age interaction effects are the ones centered. The line that crosses the 45-degree line shows that the under-five mortality is not significant between corresponding categories. The average of breastfeeding by mother's age interaction effects on logit scale is -3.5507 as given in Figure 4.2 below. One can observe that the differences in means of levels with the vertical lines that crosses 95% decision limits suggest that they are significant. This provides more insight of what is shown in Figure 4.1. One can refer to Table C.4 in Appendix C.

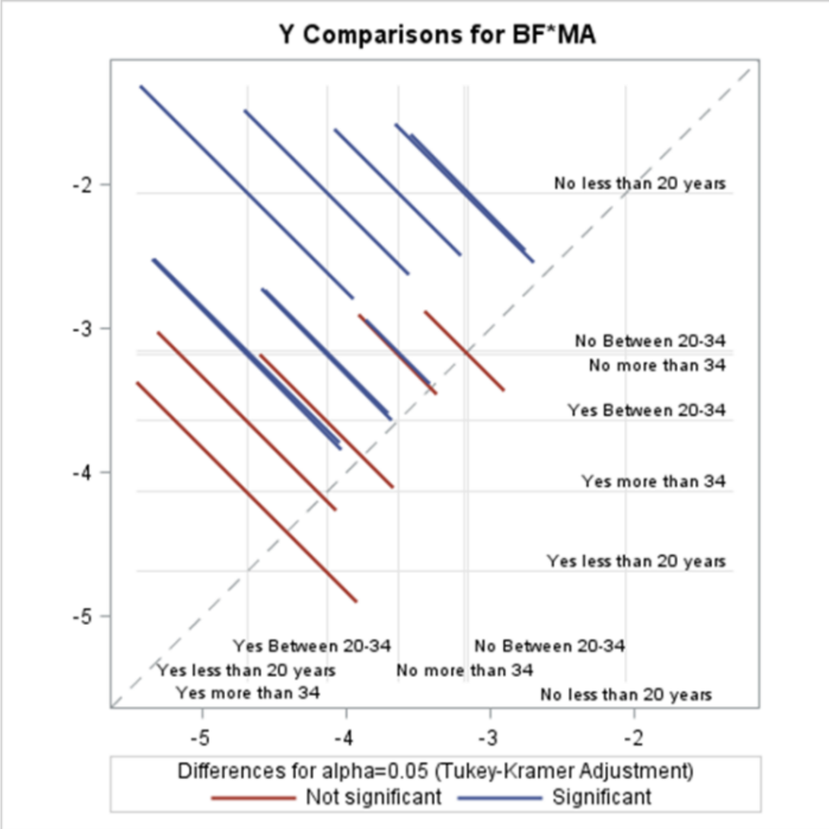

Figure 4.1: Diffogram for breastfeeding by mother's age

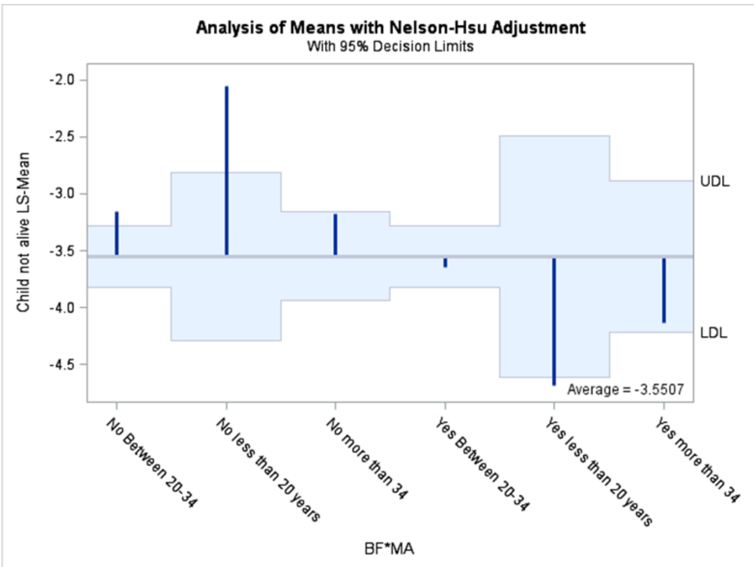

Figure 4.2: Analysis of means for breastfeeding by mother's age interaction effects

Figure 4.3 illustrates adjusted comparison of breastfeeding by birth order interaction least-square means for multiplicity. The blue lines represent the significant difference between the least-square means of level of breastfeeding by birth order interaction effects. These lines do not cross the 45-degree line. This figure shows that the lines centered at intersections and denoted by blue lines more than 4 birth order and two to four birth order represent a significant difference of least-square means of breastfeeding by birth order interaction effects. The blue line shows that the under-five mortality is significantly associated with corresponding categories.

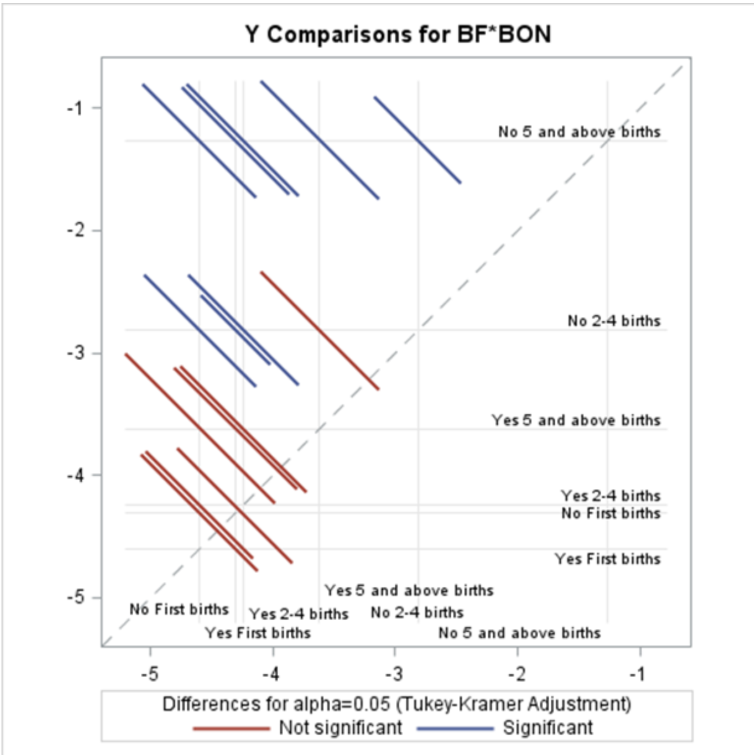

Figure 4.3: Diffogram for breastfeeding by birth order of the child.

Figure 4.4 displays the analysis of means for breastfeeding by child birth order. The blue line that crosses the decision limits shows that the under-five mortality is significantly associated with corresponding categories. The average of breastfeeding by birth order interaction effect on a logit scale is -3.63. One can observe that the differences in means of levels with the vertical lines that cross 95% decision limit suggest that they are significant. This figure agrees with Figure 4.3 and Table C.3 in Appendix C shows least square means with confidence intervals and standard errors.

Table C.2 in Appendix C2 shows standard errors, odds ratios and corresponding 95%

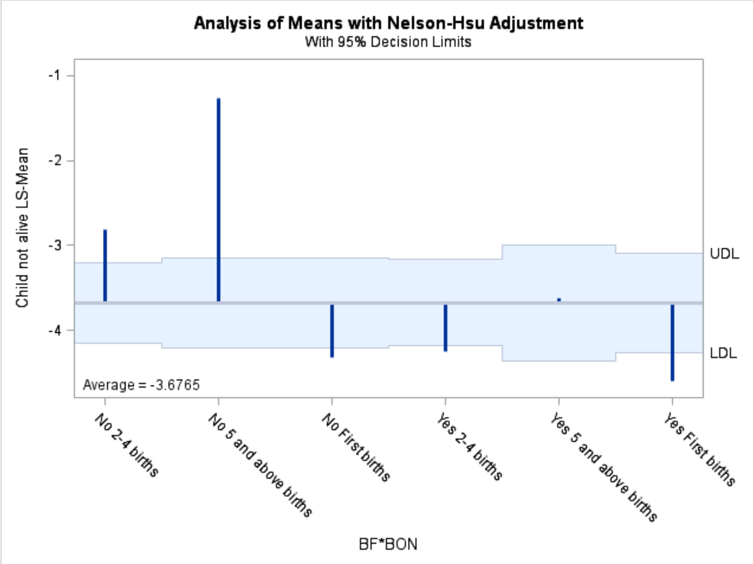

Figure 4.4: Analysis of means for breastfeeding by child birth order interaction effects.

confidence interval obtained using Laplace method. These odds ratios were obtained using the procedure PROC GLIMMIX and they were slightly different from those obtained using PROC LOGISTIC in GLMs. Some variables that were found to be significant in ordinary logistic are not significant in GLMMs. This can be the result of accounting for correlation by including random effects in the model. The covariate, breastfeeding, was not significantly associated with under-five mortality (p-value=0.6355). The corresponding odds ratio was 0,7811 (with 95% CI:0.2812-2.1698). The effects of HIV status of the mother which is HIV-positive was found to be positively associated with under-five mortality (p-value=0.0074). The corresponding odds ratio was 1.6551 (with 95% CI:1.1442-2.3834). The odds of death for a child from HIV-positive mother were 1.6551 times the odds of death for a child from HIV-negative mother. Mother's age was found to be insignificant. The effect of childbirth order number that is more than four was found to be significantly associated with under-five mortality(p-value=0.0223). The corresponding odds ratio was 2.6663 (with 95% CI:1.1499-6.1826). The number of children alive within two to four was found to be positively associated with under-five mortality (p-value=0.0002) and the number of children alive which is more than four was found to be negatively associated with the under-five mortality (p-value=0.0223). The corresponding odds ratios were 3.0114 (with 95% CI:1.6776- 5.4047) and 0.4492 (with 95% CI:0.2200-0.9170) respectively. The odds of death for a child from a mother with two to four children alive were 3.0114 times the odds of death for a child from a mother with less than two children alive. The odds of death for a child from a mother with more than four children alive were 0.4492 times the odds of death for a child from a mother with less than 2 children alive. The number of children (five years or under) in a household that is from two to four was found to be negatively associated with the under-five mortality (p-value=0.0104). The corresponding odds ratio was 0.6525 (with 95% CI:0.4707-0.9047). This implies that child is likely to survive especial when child is from a household with two to four children under five or under compared to a child from a household with less than two children five or under. The odds of death for a child from a mother who is from a household with two to four children five years or under were 0.6525 times the odds of death for a child from a mother who is from a household with less than two children five and under. The two-way interaction effects for breastfeeding (category "No") by mother's age less than 20 years is positively associated with the under-five mortality (p-value=0.0082). The corresponding odds ratio was 5.3351 (with 95% CI:1.5419-18.4593). The odds of death for a child from a mother who does not breastfeed and age less than 20 years were 5.3351 times the odds of death for a child from a mother who breastfeed and age over 34 years. Two-way interaction effects for breastfeeding (category "No") by a number of children alive within two to four was found to be positively associated with under-five mortality (p-value=0.0012). The corresponding odds ratio was 3.3848 (with 95% CI:1.6199-7.0728). The odds of death for a child from a mother who does not breastfeed by a number of children alive within two to four were 3.3848 times the odds of death for a child from a mother who breastfeeds by a number of children alive that is less than two. Two-way interaction of breastfeeding (category "No") by child birth order number that is within two to four and breastfeeding (category "No") by birth order, more than four were found to be positively associated with the under-five mortality. The corresponding odds ratios were 3.1265 (with 95% CI:1.5221-6.4231) and 7.9129 (with 95%

CI:2.8173-22,2251) respectively.

Table 4.4: Marginal quasi-likelihood (MQL), coefficients, standard errors, odds ratios, p-values and confidence intervals.

					95% confidence interval	
Effects	**Estimate**	**Odds ratio**	**Standard error**	**P-value**	**Lower limits**	**Upper limits**
Intercept	-4.5776		0.4294	0.0001		
Socio-demographic characteristics						
Breast feeding						
Yes(ref)						
No	-0.2316	0.7933	0.5205	0.4939	0.2860	2.2003
HIV status						
Negative(ref)						
Positive	0.4975	1.6446	0.1859	0.0011	1.1424	2.3676
Birth order number						
Less than 2 birth(ref)						
2 to 4 birth	0.3548	1.4259	0.3085	0.0632	0.7789	2.6103
More than 4 birth	0.9807	2.6663	0.4291	0.0020	1.1499	6.1826
Respondent age						
Over 34 years (ref)						
20 to 34 years	0.4941	1.6390	0.3226	0.8244	0.8716	3.0833
Less than 20 years	-0.5475	0.5784	0.5375	0.3172	0.2004	1.6484
Socio-economic characteristics						
Children alive						
Less than 2 children(ref)						
2 to 4 children	1.1002	3.0048	0.2985	0.0001	1.6739	5.3939
More than 4 children	-0.8032	0.4479	0.3648	0.0010	0.2191	0.9156
Children under-five years						
Less than 2 children(ref)						
2 to 4 children	-0.4206	0.6567	0.1659	0.0210	0.4744	0.9090
More than 4 children	-0.4861	0.6150	0.3090	0.5371	0.3356	1.1270
Interaction between Socio-demographic and Socio-economic characteristics						
Breast feeding by Birth order						
yes versus less than 2 birth(ref)						
No Versus 2 to 4 birth	1.1225	3.0725	0.3663	0.0001	1.4986	6.2994
No versus more than 4 birth	2.0446	7.7261	0.5261	0.0001	2.7551	21.6663
Breast feeding by Respondent age						
Yes versus Over 34 years(ref)						
No versus 20 to 34 years	-0.4751	0.6218	0.3751	0.0521	0.2981	1.2971
No versus Less than 20 years	1.6615	5.2672	0.6325	0.0080	1.5247	18.1960
Breast feeding by children alive						
Yes versus less than 2 children(ref)						
No versus 2 to 4 children	1.2041	3.3338	0.3748	0.0010	1.5992	6.9498
No versus more than 4 children	-0.7972	0.4506	0.4349	0.3610	0.1921	1.0568

Table 4.5: Penalized quasi-likelihood coefficients, standard errors, odds ratios, p-values and confidence intervals.

					95% confidence interval	
Effects	**Estimate**	**Odds ratio**	**Standard error**	**P-value**	**Lower limits**	**Upper limits**
Intercept	-4.5799		0.4285	0.0001		
Socio-demographic characteristics						
Breast feeding (BF)						
Yes(ref)						
No	-0.2329	0.7922	0.5202	0.2662	0.2858	2.1961
Mother's HIV status (HS)						
Negative(ref)						
Positive	0.4975	1.6446	0.1856	0.008	1.1431	2.3662
Birth order number (BON)						
Less than 2 birth(ref)						
2 to 4 birth	0.3574	1.4296	0.3078	0.2321	0.7820	2.6135
More than 4 birth	0.9857	2.6797	0.4279	0.001	1.1584	6.1990
Respondent age(MA)						
Over 34 years (ref)						
20 to 34 years	0.4972	1.6441	0.3220	0.0733	0.8747	3.0905
Less than 20 years	-0.5474	0.5785	0.5366	0.0553	0.2021	1.6559
Socio-economic characteristics						
Number of children alive (CL)						
Less than 2 children(ref)						
2 to 4 children	1.1007	3.0063	0.2979	0.001	1.6767	5.3902
More than 4 children	-0.8039	0.4476	0.3635	0.001	0.2195	0.9126
Children under-five years (C5)						
Less than 2 children(ref)						
2 to 4 children	-0.4201	0.6570	0.1651	0.0040	0.4754	0.9080
More than 4 children	-0.4859	0.6151	0.3069	0.0601	0.3371	1.1226
Interaction between Socio-demographic and Socio-economic characteristics						
Breast feeding by Birth order number						
yes versus less than 2 birth(ref)						
No Versus 2 to 4 birth	1.1232	3.0747	0.3654	0.0010	1.5023	6.2927
No versus more than 4 birth	2.0464	7.7400	0.5245	0.0001	2.7687	21.6373
Breast feeding by Respondent age						
Yes versus Over 34 years(ref)						
No versus 20 to 34 years	-0.4747	0.6221	0.3744	0.0760	0.2986	1.2958
No versus Less than 20 years	1.6623	5.2714	0.6311	0.0010	1.5301	18.1606
Breast feeding by children alive						
Yes versus less than 2 children(ref)						
No versus 2 to 4 children	1.2041	3.3338	0.3742	0.0090	1.6011	6.9416
No versus more than 4 children	-0.7978	0.4503	0.4334	0.0540	0.1926	1.0530

4.8 Summary of Generalized Linear Mixed Models

GLMMs are the extension of the GLMs. In these models, the linear predictor is the mixture of random effects and fixed effects. These models also relax the normality assumption made in the case of LMMs. GLMMs could be used to incorporate correlations in the model and identify sensitive subjects. For GLMMs the modeling is straightforward, one has to first identify the distribution of data, understand what need to be modeled then identify random and fixed effects. SAS procedure used to fit such models is PROC GLIMMIX and estimation method can be specified under the statement method. The methods that could be specified are Laplace, Penalized Quasi-Likelihood, and Marginal Quasi-Likelihood. The results obtained using PROC GLIMMIX procedure and the Laplace method shows that HIV status of the mother, number of children alive, and child birth order number are associated with under-five mortality. Furthermore, these results also show the two-way interaction that is associated with the under-five mortality. These two-way interaction includes breastfeeding by child birth order number, breastfeeding by a number of children alive and breastfeeding by mother's age. Other methods lead us to similar a conclusion as Laplace. The Penalized quasi-likelihood was the method with small standard errors for each parameter estimates compared to Laplace and Marginal quasi-likelihood. The GLMMs is attractive for use in modeling. Nonetheless, it still makes the assumption about linearity between log odds and predictors which may not always be true. The alternative is to use generalized additive models.

Chapter 5

Generalized Additive Models

5.1 Introduction

The statistical models which have been discussed so far assume linearity parametric form for the covariate effects. However, in some cases, this assumption of linear dependence of response on covariates may not hold. These parametric regression models discussed provide a powerful tool for modeling the relationship between response and set of covariates. However, these parametric models are not flexible for modeling a complicated relationship between response and set of covariates. The limitation of the parametric modeling is that it is restrictive in many cases. This section describes the flexible statistical non-parametric models that can be used to model complicated relationships between the response and a set of covariates. These models are known as the generalized additive models (GAMs) and they are non-parametric. They can be applied in the settings that include standard continuous response regression, count, dichotomous response, survival data and time series data. GAMs are suitable for exploring the data set and visualizing the relationship between the dependent variable and the independent variables (Liu, 2008). The parametric and non-parametric regression models should not be viewed as competing models, but as methods that complement each other (Hastie and Tibshirani, 1986, 1990; Wood, 2006). One can use non-parametric techniques to validate a parametric model. Using a combination of parametric and non-parametric methods is much more powerful than using only one of the two methods (Marx and Eilers, 1998; Wood, 2006). One of the discussed statistical models is the logistic regression for binary data which falls under the generalized linear models (with many other models). The logistic regression models the effect of covariates x_j in terms of a linear predictor of the form $x_j\beta_j$, where β_j are the model parameters. The GAMs generalizes the general linear models and GLMs by replacing $\beta_0 + \sum_{j=1}^{p} x_j\beta_j$ with $S_0 + \sum_{j=1}^{p} S_j(x_j)$, where S_j is unspecified ('non-parametric') function. This function can be estimated in a flexible manner using cubic spline smoother, in an iterative method called back-fitting algorithm (Hastie and Tibshirani, 1990; Liu, 2008). The name cubic spline is from the piecewise polynomial fit, with the order k=3 (Liu, 2008). We define a smoother as a tool for summarizing the trend of a dependent variable as a function of one or more independent variables. The smoother produces estimated known as smooth (Liu, 2008).The main property of smoother is its non-parametric nature. The estimate of the trend produced is less variable than response or log odds itself. The strength of GAMs is the ability to deal with highly non-linear and monotonic relationships between the log odds variable and one or more independent variables. Generalized additive models rely on the assumption that functions have to be additive and that the added component needs to be smooth. The GAMs were originally developed by Hastie and Tibshirani (1986) to match properties of GLM with additive models. We first begin with the overview of the methodology then discuss the form of the logistic regression in the generalized additive models setting.

5.2 Univariate Smooth Function

The smoother is the tool for summarizing the trend of response as a function of covariates (Liu, 2008; Wood, 2006). We first consider the simplest smooth function, where the model contains one smooth function of one covariate

$$y_i = S(x_i) + \epsilon_i \tag{5.1}$$

where y_i is the response variable, x_i is the covariate, $S(.)$ is the smooth function and ϵ_i are independent identically distributed random variables with mean zero and constant variance(σ^2). In order to approximate the smooth function, suppose we have a scatterplot of the points (x_i, y_i) where y_i is the response and x_i is the covariate value for a point. We want to fit the smooth curve which describes the relationship between y and x. The method of curve interpolation to determine the curve that simply minimizes $(\boldsymbol{y} - \boldsymbol{X\beta})^{'}(\boldsymbol{y} - \boldsymbol{X\beta})$ will not yield the smooth curve at all (Wood, 2006). However, the cubic spline smoother does forces smoothness on $S(x)$. The model is then fitted by minimizing the following penalized least-square function.

$$\sum_{i=1}^{n}(y_i - S(x_i))^2 + \lambda \int_a^b [S^{''}(x)]^2 dx \tag{5.2}$$

where λ is fixed constant(smoothing parameter) and $a \leq x_1 \leq \ldots \leq x_n \leq b$. We assume (a,b) includes all possible range. The functions S can be approximated by linear combination of basis functions $b_j(x)$ as $S(x) = \sum_{j=1}^{q} b_j(x)\beta_j$ and $\int [S^{''}(x)]^2$ measure the "Wiggliness" of the function S. If the $\int [S^{''}(x)]^2 = 0$ (indicate a straight line or perfect curve) then we have a function S that is a linear function. However, a non-linear function of S will produce values $\int [S^{''}(x)]^2 > 0$ (smoother S is highly non-linear). The smoothing parameter $\lambda > 0$ has to be chosen wisely by the analyst since its plays an important role in estimation. The parameter λ controls the tradeoff between the goodness of fit that is measured by $(y_i - S(x_i))^2$ and the model smoothness (Hastie and Tibshirani, 1990). The larger the value of λ the smoother S becomes and the penalty term becomes more important. Furthermore, the small values of λ yield a wiggly curves and penalty become unimportant (Liu, 2008; Yee and Mitchell, 1991). We now look at additive model by penalized least-square and general case.

5.3 Additive Models by Penalized Least-Squares

The function S is the linear combination of the parameters, and one can show that the penalty from penalized least square can be written as quadratic form of β.

$$\int [S^{''}(x)]^2 dx = \boldsymbol{\beta}'\boldsymbol{H}\boldsymbol{\beta}. \tag{5.3}$$

Suppose now the model has two smoothers as follow

$$Y_i = S_1(x_i) + S_2(x_i) + \epsilon_i. \tag{5.4}$$

The smoothers has the form $S_1(x) = \sum_{j=1}^{q_1} b_{1j}(x_i)\beta_j$ and $S_2(x) = \sum_{j=1}^{q_2} b_{2j}(z_i)\gamma_j$. Where x and z are two explanatory variables and for simplicity we assume that all x_i and z_i lie in [0, 1]. Here $b_{1j}(.)$ and $b_{2j}(.)$ are cubic spline basic functions of S_1 and S_2 respectively. When two smoothers are now used in place of one smoother then this the definition of $\boldsymbol{Y}$ as a function of q, $\boldsymbol{X}$ and $\boldsymbol{\beta}$. However, the general form does not (Wood, 2012). The optimization becomes

$$\sum_{i=1}^{n}(y_i - S(x_i))^2 + \lambda_1\boldsymbol{\beta}'\boldsymbol{H_1}\boldsymbol{\beta} + \lambda_2\boldsymbol{\beta}'\boldsymbol{H_2}\boldsymbol{\beta} \tag{5.5}$$

where $\boldsymbol{X}$ is a design matrix of covariates, λ_1 , λ_2 directly control the effective degree of freedom per smoothing term. The smoothing parameter can also be obtained by generalized cross validation (Wood and Augustin, 2002). Here $\boldsymbol{H} = \int d(x)d(x)'dx$ is the penalty matrix which consists of known coefficients and $d(x)$ is given by

$$d(x) = [b_1^{''}(x), b_2^{''}(x), b_3^{''}(x), \dots]'.$$

We then can argue that the penalized regression spline fitting problem is similar to minimizing

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}'\boldsymbol{H}\boldsymbol{\beta}. \tag{5.6}$$

This can also be written as

$$\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}'\boldsymbol{\beta}'\boldsymbol{y} + \boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{H})\boldsymbol{\beta}.$$

Taking the derivative with respect to β and equating to zero, we obtain

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{H})^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{5.7}$$

The parameter λ can be set by hand or selected automatically and penalized maximum likelihood could be used to estimate the unknown parameter β (Liu, 2008). The Hat or Influence matrix, $\boldsymbol{A}$ for this model is given as

$$\boldsymbol{A} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{H})^{-1}\boldsymbol{X}'. \tag{5.8}$$

We first require some method for choosing λ.

5.4 Selection of Smoothing Parameters λ

In order to minimize cubic spline smoother which is being considered, we have to choose a smoothing parameter, λ, wisely. If λ is much higher then the data will be over smoothed, but if λ is too low then the data will be under smoothed (Wood, 2006). It is possible to choose λ that is data driven. The penalized likelihood can be used to estimate model coefficients given λ. There are other approaches that are useful when the scale parameter is known instead of attempting to minimize expected mean square error which results into estimation by Un-Biased Risk Estimation(UBRE). If the scale parameter is unknown then attempting to minimize prediction error leads to ordinary cross validation or generalized cross validation (Wood, 2006).

5.4.1 Average Mean Square and Predictive Square Error

One can focus on the global measure known as Average Mean square Error (AMSE), instead of minimizing the Mean Square Error (MSE) at each covariate x_i (Liu, 2008; Wood and Augustin, 2002). The average mean square error is given by.

$$AMSE(\lambda) = \frac{1}{n}\sum_{i=1}^{n} E[\hat{S}_{\lambda}(x_i) - S(x_i)]^2 \tag{5.9}$$

where $\hat{S}_{\lambda}(x_i)$ is an estimator of $S(x)$ and $S(x_i = Y_i - \epsilon_i)$. We now consider the Average Predictive Square Error (PSE) which is given by

$$PSE(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[Y_i^* - \hat{S}_{\lambda}(x_i)]^2. \tag{5.10}$$

AMSE and PSE differ by a constant δ, where Y_i^* is the new observation at x_i, $Y_i^* = S(x_i) + \epsilon_i^*$ and ϵ_i^* is independent of $\epsilon^{'}s$. There are other procedures for estimating for selecting $\lambda$for example Cross Validation(CV) and Generalized Cross Validation(GCV).

5.4.2 Cross Validation

CV is a statistical approach for partitioning sample data into two subsets (Liu, 2008; Wood, 2006). This technique is sufficient when the sample is large. The data is recycled by switching the role of tests samples and training in CV. Cross-validation could be used in selecting λ, by minimizing

$$CV(\lambda) = \sum_{i=0}^{n}[y_i - \hat{S_{\lambda}^{-i}}(x_i)]^2 \tag{5.11}$$

where, $\hat{S_{\lambda}^{-i}}(x_i)$ indicates the fit at x_i which is computed by leaving out the i^{th} data point. This is the approach that is available in SAS and is similar to minimizing $PSE(\lambda)$.

5.4.3 Generalized Cross Validation

Another approach for selecting λ is known as GCV which is computationally intensive. However, there are some shortcuts available for many situations (Liu, 2008; Wood, 2006). The GCV is approximately the same as Mallow's C_p statistic and this shown in the study by Liu (2008). The GCV is given by

$$V_g = \frac{n \parallel y_i - X\hat{\beta}_\lambda \parallel}{[n - tr(F_\lambda)]^2} \tag{5.12}$$

where $tr(F_\lambda)$ is the effective degree of freedom of the model, and $\hat{\beta}_\lambda$ is the coefficient of the estimate that is obtained by direct minimization of

$$\parallel y - X\beta \parallel^2 + \sum_j \lambda_j \beta' H_j \beta.$$

5.4.4 Degrees of Freedom of a Smoother

The other way of expressing the required smoothness of the function other than in terms of λ are to use degrees of freedom. In SAS procedure PROC GAM one can select the value of a smoothing parameter through specifying the degrees of freedom of a smoother also known as an effective number of parameters. The effective number of parameters indicates the amount of smoothing. Suppose there is a linear smoother say F_λ, then the degrees of freedom is given by

$$df(Smoother) = tr(F_\lambda).$$

The more the smoothing the fewer degrees of freedom of the smoother. The degrees of freedom may be a decimal number (Liu, 2008).

5.5 Back-fitting and General Local Scoring Algorithm

The general idea of the generalized additive model is to plot the value of the response variable together with single covariate then compute the smooth curve that goes through the data. GAMs are designed to take advantage of the ability to fit the logistic regression and other GLMs. Its main focus is to explore the data set and visualize the relationship between response and set of covariates (Liu, 2008; Marx and Eilers, 1998). However, the GLMs focus specifically in estimation and inference. The data is divided into a number of a section called knots. The scatterplot smoother used in GAMs attempts to generalize data into a smooth curve by local fitting to the subsection of the data. One of the advantages of GAMs is that the error term is estimated precisely since curves are fitted algorithmically. The algorithms used are often iteratively, non-parametric, and do not show a great deal of complex numerical processing. The GAMs framework is based on back-fitting with linear smoothers, limitations arise in the difficulty that is presented by back-fitting in the selection of a model and inference(Marx and Eilers, 1998). There are different techniques for the formulation and estimation of additive models. The general algorithm for model formulation and estimation of the additive model is called back-fitting. Back-fitting can fit an additive model using any regression type fitting mechanism(Wood, 2006).

5.5.1 Back-fitting Algorithm

Define the partial residual as

$$R_j = Y - S_0 - \sum_{k \neq j} S_k(x_k)$$

with $E(R_j \mid X_j) = S_j(x_j)$. This observation provides a way for estimating each smooth function $S_j(.)$ given the estimate $[\hat{S}(.), i \neq j]$ for all others. The resulting iterative procedure is known as back-fitting.

Step 1. initialize:

$$S_0 = E(Y), S_1^1 = \ldots = S_p^1, m = 0.$$

Step 2. Iterate: $m = m + 1$ for $j = 1$ to p do:

$$R_j = Y - S_0 - \sum_{k=1}^{j-1} S_k^m(x_k) - \sum_{k=j+1}^{p} S_k^{m-1}(x_k)$$

$$S_j^m = E(R_j \mid X_j).$$

Step 3: Calculate

$$RSS = AVG(Y - S_0 - \sum_{j=1}^{p} S_j^m(x_j))^2$$

until fails to decrease. $S_j^m(.)$ denotes the estimate of $S_j(.)$ at the m^{th} iteration. RSS do not increase at any step and thus the algorithm always converges.

5.5.2 General Local Scoring Algorithm

Step1: Initialize:

$$S_0 = E(Y), S_1^1 = \ldots = S_p^1, m = 0.$$

Step 2: Update/iterate $m = m + 1$, from the adjusted dependent variable

$$z_i = \eta_i + (y_i - \mu_i)(\frac{\partial \eta_i}{\partial \mu_i}),$$

$$\eta^{m-1} = S_0 + \sum_{j=1}^{p} S_j^{m-1}(x_{ij}),$$

$\eta^{m-1} = g(\mu^{m-1})$ so $\mu^{m-1} = g^{-1}(\eta_i)$ construct the weight.

$$W_i = (\frac{d\eta_i^{m-1}}{d\mu_i^{m-1}})^2 V_i^{-1},$$

where $V_i = var(Y_i)$. Fit a weighted additive model to z_i using the back-fitting algorithm with weights W. We obtain estimated functions $S_i^m(.)$ and model η^m.

Step 3: Repeat: continue with step 1 and step 2 until deviance fails to decrease. Suppose the initial estimate of η is given, then the first order Taylor series expansion and fisher scoring method will yield an improved estimate according to Liu (2008).

$$\eta^{\text{est}}(x) = \eta^{\text{given}} + \delta. \tag{5.13}$$

Here;

$$\delta = \frac{\text{Score function}}{\text{Exected Information matrix}},$$

$$\begin{aligned} \delta &= \frac{\frac{\partial l}{\partial \eta}}{E(\frac{-\partial^2 l}{\partial \eta^2} \mid x)}, \\ &= E(\eta(x) - \frac{\partial l}{E(\frac{-\partial^2 l}{\partial \eta^2} \mid x)} \mid x). \end{aligned} \tag{5.14}$$

Using chain rule we have that

$$\frac{\partial l}{\partial \eta} = \frac{\partial l}{\partial \mu}\frac{\partial \mu}{\partial \eta},$$

$$\begin{aligned} \frac{\partial l}{\partial \mu_i} &= \frac{1}{\mu_i} - (1 - y_i)\frac{1}{(1 - \mu_i)}, \\ &= \frac{y_i - \mu_i}{(1 - \mu_i)\mu_i}. \end{aligned} \tag{5.15}$$

We know that $Var(Y_i) = E(Y_i^2) - (E(Y_i))^2$,

$$\begin{aligned} \text{Var}(Y_i) &= E(Y_i^2) - (E(Y_i))^2, \\ &= 1^2\mu_i + 0^2(1 - \mu_i)(-\mu_i), \\ &= \mu_i(1 - \mu_i) \end{aligned} \tag{5.16}$$

and

$$V_i^{-1} = \frac{1}{\mu_i(1 - \mu_i)}.$$

Thus

$$\frac{\partial l}{\partial \eta} = (y - \mu)V^{-1}\frac{\partial}{\partial \eta},$$

$$\frac{\partial^2 l}{\partial \eta^2} = (y - \mu)\frac{\partial}{\partial \eta}(V^{-1}\frac{\partial \mu}{\partial \eta}) - (\frac{\partial \mu}{\partial \eta})^2 V^{-1},$$

and

$$E(\frac{\partial^2 l}{\partial \eta^2} \mid x) = -(\frac{\partial \mu}{\partial \eta})^2 V^{-1},$$

$$\eta^{est}(x) = E[\eta(x) + (Y - \mu)\frac{\partial \eta}{\partial \mu} \mid x]. \tag{5.17}$$

Replacing the conditional estimation with smoothers we have the improved estimates

$$\eta^{\text{est}}(x) = \text{smoother}[\eta(x) + (Y - \mu)\frac{\partial \eta}{\partial \mu} \mid x]. \tag{5.18}$$

5.6 Estimation of the Parameter Estimate β

If the data is non-normal, one can apply the framework of the GLM. The linear predictor is modeled as the sum of the B-spline and iterative method (scoring) is used. The smoothness of the curve will be influenced by the number of B-spline, a value of the coefficient or amplitudes. If these are almost equal then the curve will be flat. The curve will show a lot of wiggles if the amplitude varies widely.

5.6.1 B-splines

There are other popular smoothing techniques besides cubic spline such as loess and kernel smoothers, where the graphical summaries of non-parametric fits are provided in them. However, despite the fact that non-parametric provides rich exploratory flexibility, it is not simple to use for future prediction (Wood, 2006; Marx and Eilers, 1998). The B-spline smooth basis is independent of the response variable but only dependent on:

- Range of the covariate,
- the number and position of knots (equally spaced), and
- the degree of the B-spline(often cubic).

The B-spline of q degree consists of $q + 1$ polynomial pieces of degree q; these pieces are joined at q inner knots at which the derivatives up to order $q - 1$ are continuous. The B-spline is positive on the domain spanned by $q + 2$ knots, for a given x $q + 1$ B-spline is non-zero. The fit to the data can be expressed as

$$S = \sum_{i=1}^{N} (y_i - \sum_{t=1}^{n} b_{it} a_t)^2 \tag{5.19}$$

where $b_{it} = B_t(X_i)$, the value of the B-spline t at X_i, $\sum_{t=1}^{n} b_{it} a_t$ is the sum of B-splines. The solution for the vector a is obtained from regression of y on the matrix $\boldsymbol{B}$ and $\boldsymbol{B}$ is known as B-spline matrix of dimension $N \times n_i$.

5.6.2 P-splines

This is another way of representing the cubic splines by the use of B-spline basis. The B-spline basis are strictly local so there are more appealing and each basis function is zero over intervals $m + 3$ adjacent knots (Wood, 2006). The $(m + 1)^{th}$ order spline can be expressed as

$$S(X) = \sum_{i=1}^{k} \beta_i^m (X) \beta_i. \tag{5.20}$$

The B-spline basis function is defined recursively as

$$\beta_i^m = \frac{X - X_i}{X_{i+m+1} - X_i} \beta(X)^{m-1} + \frac{X_{i+m+2} - X}{X_{i+m+2} - X_{i+1}} \beta_{i+1}^{m-1}, i = 1, \ldots, k \tag{5.21}$$

$$\beta_i^{-1}(X) = \begin{cases} 1 & \text{if } X_i \leqslant X < X_{i+1} \\ 0 & \text{otherwise.} \end{cases} \tag{5.22}$$

There are others spline such as cyclic cubic regression spline, cubic regression spline, thin plate regression spline and thin plate spline (seeWood (2006)).

5.6.3 Penalized Likelihood and Estimation

The penalized likelihood is an alternative way to find regression coefficients for categorical variable(s). The likelihood is maximized by using iterative method such as Newton-Raphson algorithm and Scoring method. Newton-Raphson method is a technique used to find the zero(s) of a function taking real values (Wood, 2006; Marx and Eilers, 1998).

Penalized Likelihood

The drawback for using B-spline is that one is required to optimize the number and position of knots. Given a wiggliness measure for each function, the penalized log-likelihood can be defined as

$$\begin{aligned} \log L_p(\boldsymbol{\beta}) &= \log L(\boldsymbol{\beta}) - \frac{1}{2}\sum_{j=1}^{p} \lambda_j \boldsymbol{\beta}' H_j \boldsymbol{\beta}, \\ &= \log L(\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}' \boldsymbol{S}\boldsymbol{\beta} \end{aligned} \tag{5.23}$$

where $\boldsymbol{S} = \sum_{j=1}^{p} \lambda_j H_j$, L denotes the usual likelihood function and λ_j are penalty factors or smoothing parameters, controlling the tradeoff between goodness of fit of the model smoothness. Assuming that λ_j values are known, then the likelihood is maximized in order to find $\hat{\beta}_j^{'}$s.

Estimation

The penalized log-likelihood in equation (5.23) can be maximized through iterative reweighted Least-Squares. Here we assumes that λ_j is known. To maximize this equation one needs to take its derivative with respect to β_j and equate to zero, that is to say,

$$\frac{\partial l_p}{\partial \beta_j} = \frac{\partial l}{\partial \beta_j} - [\boldsymbol{S}\boldsymbol{\beta}]_j = \phi^{-1} \sum_{i=1}^{n} \left\{ \frac{y_i - \mu_i}{V(\mu_i)} \right\} \frac{\partial \mu_i}{\partial \beta_j} - [\boldsymbol{S}\boldsymbol{\beta}]_j = 0. \tag{5.24}$$

The $[.]_j$ is the j^{th} row vector. The equation resulted in minimizing the likelihood are the same as those equations that would have to be solved to obtain β by non-linear weighted least square, given that weight $V(\mu_i)$ are known in advance and are independent of β (Wood, 2006). The Least-Square objective would be

$$S_p = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{Var(Y_i)} + \boldsymbol{\beta}' \boldsymbol{S}\boldsymbol{\beta} \tag{5.25}$$

where, μ_i depends non-linearly on β, but the weights $V(\mu_i)$ are treated fixed. The assumption made here is that the $Var(Y_i)$ terms are known. In order to find the least-square estimates, one can take a derivative with respect to β_j and equating to zero. This System of equations will be as in equation (5.24). If $var(y_i)$ terms were fixed. The iterative method is required to solve these equations. It can be shown that in the vicinity of some coefficient vector estimate $\hat{\beta}^{|k|}$ (Wood, 2006).

$$S_p \simeq \|\sqrt{w^{[k]}}(z^{[k]} - \boldsymbol{x\beta})\|^2 + \boldsymbol{\beta}'\boldsymbol{S\beta} \tag{5.26}$$

The pseudo-data is defined as

$$z_i^{[k]} = g'(\mu^{[k]})(y_i - \mu_i^{[k]}) + \boldsymbol{X_i}\hat{\boldsymbol{\beta}}^{[k]} \tag{5.27}$$

where $\boldsymbol{z}^k$ is a vector of pseudo-data with elements $z_i^{[k]}$ and $\boldsymbol{W}^{[k]}$ is the diagonal weight matrix with elements $w_i^{[k]}$ given by

$$w_i^{[k]} = [V(\mu_i^{[k]}g'(\mu_i^{[k]})^2)]^{-1} \tag{5.28}$$

where g is the model link function. Assuming the smoothing parameters are known, then the maximum penalized likelihood estimates, $\hat{\beta}$, are obtained through iterating the following steps:

Step1: Use current $\beta^{[k]}$, compute the pseudo-data $z^{[k]}$ and iterative weights $W^{[k]}$.

step 2: Minimize equation (5.26) with respect to β, then obtain $\hat{\beta}^{[k+1]}$; and so that $\eta^{[k+1]} = \boldsymbol{X\beta}^{[k+1]}$. Increase value of k by one unit.

The converged $\hat{\beta}$ solves equation (5.24).

5.7 Generalized Additive Logistic Regression Model

In Chapter 3 logistic regressions was discussed as one of the popular technique for modeling binary data since we have a dichotomous response variable.

$$Y_i = \begin{cases} 1, & \text{if child is not alive (with probability } \pi(x)) \\ 0, & \text{if child is alive (with probability } 1 - \pi(x)) \end{cases}$$

$\boldsymbol{X} = (x_{i1}, x_{i2}, \dots, x_{ip})$ is a vector of covariates, and Y_i is the binary response variable. The ordinary logistic model was discussed in Chapter 3 and is given as

$$\text{logit}(\pi(x)) = \eta_{\text{L}}(\text{x}) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}. \tag{5.29}$$

It can as well be written as

$$\pi(x) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}.$$

The basic idea of the GAMs is to replace the linear predictor with an additive predictor. The assumption for logistic regression still applies except the linearity assumption. The GAM logistic model is given by

$$\text{logit}(\pi(x)) = \eta_{\text{A}}(\text{x}) = log(\frac{\pi(x)}{1 - \pi(x)}) = S_0 + \sum_{j=1}^{p} S_j(x_{ij}). \tag{5.30}$$

Alternatively, it can be written as

$$\pi(x) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} S_j(x_{ij}))}{1 + \exp(S_0 + \sum_{j=1}^{p} S_j(x_{ij}))}.$$

The functions $S_1, S_2, \dots, S_p$ are estimated using the procedures described above. One can also have a semi-parametric generalized additive model. This happens when the model consists of parametric and non-parametric terms. The interaction effects can also be incorporated to the generalized additive model. This model with two parametric and two non-parametric predictors is of the form.

$$\text{logit}(\pi(x)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + S_1(x_3) + S_2(x_4). \tag{5.31}$$

In general the semi-parametric logistic model is written as

$$\text{logit}(\pi(x)) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{j=p+1}^{q} S_j(x_{ij}). \tag{5.32}$$

Let $E(Y \mid X) = \mu$ so that

$$\eta(x) = g(\mu) = \log \left\{ \frac{\pi(x)}{1 - \pi(x)} \right\} \tag{5.33}$$

where η is a function of p variables. Suppose $Y = \eta(x) + \epsilon$, given some initial estimate of $\eta(x)$, one can construct the adjusted dependent variable

$$Z_i = \eta_i + (y_i - \mu_i) \frac{\partial \eta_i}{\partial \mu_i}. \tag{5.34}$$

We fit an additive model to the $Z_i's$, where it is treated as a response variable Y in $\mu = S_0 + \sum_{j=1}^{p} S_j(x_{ij})$. This algorithm is the same as one mentioned earlier namely Local scoring algorithm. For more details one can see Liu (2008).

5.8 Fitting a Logistic GAM Model using the GAM Procedure

The GAMs are useful in finding a predictor-response relationship in several kinds of data without using a specific model. They combine the ability to explore non-parametric relationships together with the distributional flexibility of generalized linear models. The SAS PROC GAM scales well the increasing dimensionality and yields interpretable models (Wood, 2006). Carrying out exploratory modeling with PROC GAM could inspire parsimonious parametric models. In this section, we assume that some of the covariates have a linear relation with the log odds and in some we assume non-linearity, this yields the semi-parametric model. Using the SAS procedure PROC GAM, under model option some variable are included in the keyword spline (in this cases non-linearity assumption is made for them).

5.8.1 Observing Correlation Among Predictors

Table 5.1 shows correlation among continuous predictors considered and p-values. The p-values can be used to test if two variables are correlated or not. Figure 5.1 also shows the relationship among continuous predictors. Both Figure 5.1 and Table 5.1 suggest that there is an issue of correlation between variables. There might be an impact of multicollinearity on parameter estimates which is a concern.

Table 5.1: Pearson correlation matrix for continuous predictors.

Pearson Correlation Coefficients, N = 11013				
$Prob >\mid r \mid$ **under H0:** $\rho = 0$				
	BN	**NLC**	**RCA**	**NC5U**
Birth order number(BN)	1.0000	0.90499	0.7766	0.09905
P-value		0.0001	0.0001	0.0001
Number of living children(NLC)	0.90499	1.0000	0.74084	0.21446
P-value	0.0001		0.0001	0.0001
Respondent's current age(RCA)	0.7766	0.74084	1.0000	-0.06454
P-value	0.0001	0.0001		0.0001
Number of children 5 or under(NC5U)	0.09905	0.21446	-0.06454	1.0000
P-value	0.0001	0.0001	0.0001	

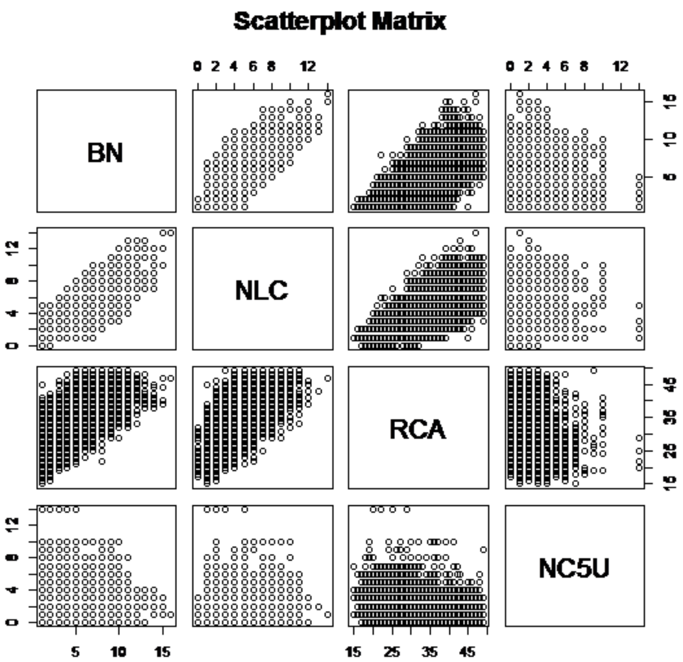

Figure 5.1: Scatter plot matrix of continuous predictors.

5.8.2 Fitting the Logistic Additive Model

Consider first part of the output that is obtained using PROC GAM procedure. Table 5.2 shows a summary for the back-fitting and local scoring algorithms. The deviance for the final estimate is also provided in Table 5.2. This value of deviance for final estimate can be used in computing the AIC as shown below

$$\begin{aligned}\text{AIC} &= \text{Deviance} + 2\text{pf},\\ &= 2749.12 + 2 \times 15 \times 1, \qquad (5.35)\\ &= 2779.12\end{aligned}$$

where p is the model degrees od freedom and f is the scale parameter(f = 1 for binomial and poison). The model degrees of freedom is 1 + 14 = 15. This AIC value can be used to compare models fitted by PROC GAM. One can not compare models fitted by PROC GAM and PROC GENMOD using AIC. In PROC GENMOD the AIC value is calculated as

$$\text{AIC} = -2\text{LL} + 2\text{p}$$

where LL is the log likelihood of the fitted model.

Table 5.2: Summary for algorithms used in fitting the model

Iteration Summary and Fit Statistics	
Number of local scoring iterations	9
Local scoring convergence criterion	1.26E-09
Final Number of Backfitting Iterations	1
Final Backfitting Criterion	1.67E-09
The Deviance of the Final Estimate	2749.115676

The critical part of PROC GAM results is the "Analysis of Deviance" shown in Table 5.3. For each smoothing effect in the model, this table provides a Chi-Square(χ^2) test comparing the deviance between full model and the one without non-parametric component variable. The analysis of deviance results shows that non-parametric effects of all four continuous predictors are significant at 5% significant level since their corresponding p-values are less than 0.05.

Table 5.3: Analysis of deviance.

Smoothing Model Analysis Analysis of Deviance				
Source	**DF**	**Sum of Squares**	**Chi-Square**	**P-value**
Spline(BN)	3	37.178636	37.1786	0.0001
Spline(NLC)	3	126.409501	126.4095	0.0001
Spline(RCA)	3	15.08642	15.0864	0.0017
Spline(NC5U)	3	28.194937	28.1949	0.0001

Note: BN: Child birth order number, NLC: Number of children alive
RCA: Mothers Age NC5U: Number of children five or under in a household

Table 5.4 shows the linear portion and parameter estimates for parametric part of the model, standard errors, t-values, and p-values. This table also shows smoothing parameters, degrees of freedom, a number of unique observation and value of GCV for each predictor. The breastfeeding is negatively associated with under-five mortality (p-value=0.0017). The HIV status of the mother was not significant at 5% significant level (p-value=0.0917). However, it was significant at 10% significant level. This suggests that both HIV status of a mother and breastfeeding were associated with the under-five mortality. The predictor mother's age in linear portion was not significant (p-value=0.9812). This might have been the result of some part of significance being taken by non-linear part. Other predictor variables such as Childbirth order, a number of children alive and a number of children 5 and under in a household are found to be significantly associated with under-five mortality since their corresponding p-values are less than 0.05. The degree of freedom is an indication of the amount of smoothing. The more the smoothing means less degree of freedom or higher span. The smoothing parameter was almost equal to one and the corresponding degree of freedom is 3. Figure 5.2 shows plots of the partial prediction for each of the continuous predictor considered. These plots can be used to

Table 5.4: Analytical information about fitted model.

Regression Model Analysis Parameter Estimate				
Effects	**Parameter Est**	**STD error**	**t-value**	**p-value**
Intercept	-1.2092	0.37268	-3.24	0.0012
HIV status of respondent				
negative	-0.33153	0.19654	-1.69	0.0917
positive(reference)	1			
Breastfeeding				
Yes	-0.37883	0.12081	-3.14	0.0017
No(reference)	1			
Birth order number				
Linear(BN)	0.73046	0.05084	14.37	0.0001
Number of children alive				
Linear(NLC)	-1.1232	0.06117	-18.36	0.0001
Respondent age				
Linear(RCA)	-0.00031019	0.01318	-0.02	0.9812
Number of children 5 or under				
Linear(NC5U)	-0.21251	0.04794	-4.43	0.0001
Smoothing Model Analysis				
Fit Summary for Smoothing Components				
Component	**Smoothing Parameter**	**DF**	**GCV**	**NUO**
Birth order number				
Spline(BN)	0.999842	3	153.455	16
Number of children alive				
Spline(NLC)	0.999684	3	597.253	15
Respondent age				
Spline(RCA)	0.998885	3	0.71918	35
Number of children 5 or under				
Spline(NC5U)	0.994058	3	22.8856	12

Note: NUO: number of unique observation, STD: standard and Est:estimate

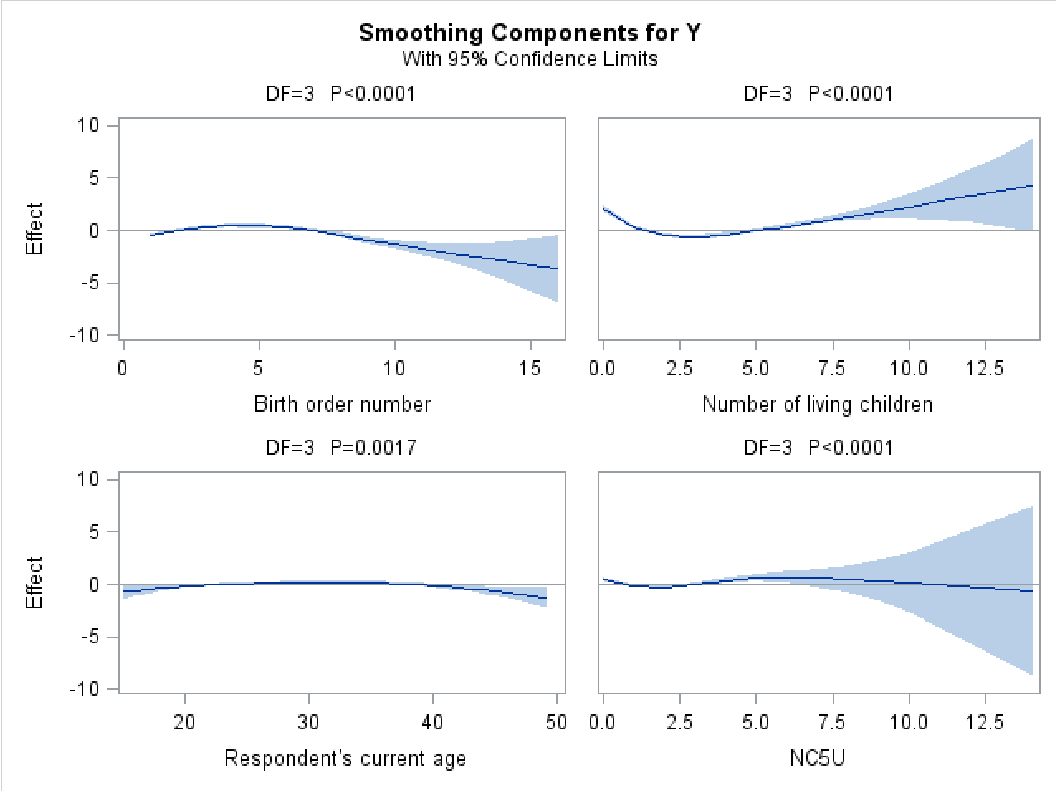

Figure 5.2: Partial prediction for each predictor.

investigate as to why PROC GAM and PROC GENMOD provide different results. These plots are produced by including the option PLOTS = COMPONENT(COMMONAXES) which gives curve-wise Bayesian confidence band to each smoothing component and plot share the same vertical axis limits. These confident interval might be wider towards the end as a result of lack of data. The plots show that the partial predictions corresponding to child birth order number, a number of children alive have quadratic pattern and a number of children 5 and under in a household does not have a quadratic pattern. This suggested that under-five mortality was associated with a quadratic pattern for child birth order number and a number of children alive. The number of children 5 and under in a household have 95% confidence limits that contain the zero axes suggesting no effect of quadratic pattern or non-linear on the survival of the child. The mother's age have 95% confidence limits containing zero axes and a line was almost straight this means that mother's age had no quadratic effect on the child survival status.

5.9 Summary of the Generalized Additive Model

Logistic regression is often used when the response is dichotomous. However, the assumption about linearity between link function (logit) and predictors need to be made. This assumption may not hold thus an alternative method is required such as GAMs. Generalized additive models are the generalization of additive models (AMs) by relaxing normality assumption. The first step to fitting GAMs is to turn GAMs into penalized generalized linear model (P-GLMs) with coefficient β and smooth parameter λ. This can be done by choosing basis and wiggliness measures for the smooth terms. Secondly is to select smoothing parameters in which one can use either GCV or UBRE. The parameter estimates β are then obtained by using penalized iteratively re-weighted least-square(P-IRLS). The confidence interval can be obtained by the use of Bayesian smoothing model (Wood, 2006). One can test the hypothesis through the use of GLM methods on un-penalized GAM. With the use of PROC GAM to fit the model; we have noticed that under-five mortality was associated with a quadratic pattern of childbirth order, a number of children alive and a number of children five and under in a household. The under-five mortality was also associated with a linear pattern of mother's age. It was also found to be significantly associated with breastfeeding and mother's HIV status and linear pattern of mother's age.

Chapter 6

Discussion and Conclusion

The objective of this study was to identify risk factors associated with the under-five mortality in the United Republic of Tanzania. The identified factors can be used to guide policy makers on speeding up the provision of better life to people and evaluate progress made towards achieving the MDG4. Generalized linear models, Survey logistic regression models, generalized linear mixed models, and generalized additive models were used to identify the risk factors. Firstly, a generalized linear model called logistic regression model that assumes survey data was obtained through simple random sampling was used. The interaction effects considered was up to the second order. Due to a large number of variables, stepwise selection procedure was adopted to eliminate non-significant variables. When logistic regression was used breastfeeding and interaction terms, breastfeeding by child birth order number, breastfeeding by mother's age, breastfeeding by a number of children alive and HIV status of a mother were significantly associated with the under-five mortality. However, a number of children alive, mother's age and birth order number were not significantly associated with the under-five mortality but due to the hierarchical principle of the model with interaction terms number of children alive, mother's age and birth order number were retained in the model. The Model checking and goodness of fit using Hosmer-Lemeshow test failed to reject the selected model. The model was refitted through the survey logistic regression model and generalized linear mixed models. Both models seem to be the good alternative since they account for the complexity of the survey design. The conclusion reached from the survey logistic was similar to the one reached by generalized linear mixed models. The risk of child death for a mother who was HIV-positive was higher compared to the incidence of child death for a mother who was HIV negative. The risk of child death for a mother who did not breastfeed was higher than the incidence of child death for a mother who did breastfeed. The incidence of child death was high for a child whose birth order number was more than one compared to the incidence of death for a child whose birth order number was less than two. The risk of child death for a mother who did not breastfeed and at middle age group (20 to 34 years) was higher compared to the incidence of child death for a mother who did breastfeed and at old age group (over 34 years). The risk of child death for a mother who did not breastfeed and at young age group (less than 20 years) was lower compared to the incidence of child death for a mother who did breastfeed and at old age group (over 34 years). The risk of child death for a mother who did not breastfeed and child whose birth order number above four was high compared to the risk of child death for a mother who did breastfeed and with a child whose birth order number less than two. The risk of child death for a mother with more than four children alive was found to be lower than the risk of child death for a mother with less than two children alive. The risk of child death for a mother with 2 to 4 children alive was high compared to the risk of child death for a mother with less than two children alive. The results from survey logistic regression and logistic regression were shown in Chapter 4 and 3 respectively. From the results, we observed that standard errors for logistic regression model are smaller compared to standard errors for survey logistic for each parameter estimate, suggesting under-estimation of variance. This shows that assumption we made in order to use logistic regression resulted in an invalid conclusion. We obtained appropriate estimates by taking into account, for the sampling design features. The parameter estimates and odds ratios for both models are almost the same. However, the confidence intervals for odds ratios are narrower for logistic regression. This has resulted in underestimation of the variance. The survey logistic regression and generalized linear mixed model are useful since they account for the complexity of the survey design.

Logistic regression, survey logistic regression, and generalized linear mixed models are often used when the response is dichotomous. However, the assumption about linearity between log (odds) and independent variables need to be made. If this assumption does not hold the generalized additive models could be used as an alternative. Using generalized additive models the under-five mortality was found to be significantly associated with the quadratic pattern of childbirth order number, a number of children alive and has no quadratic effect of a number of children five or under in a household. Under-five mortality was also found to be significantly associated with mother's HIV status and breastfeeding at 10% level of significant. We also found that under-five mortality has no quadratic effect of mother's age

The findings of this study imply that the child survival status is likely to improve in Tanzania. If breastfeeding is done by mothers it is likely to reduce the risk of death for a child under five, more especially mothers in younger age group (less than 20 years). The reduction of mothers who are infected with HIV will also improve the child survival status. The children will survive if their birth order of the child is two and above, more especially if the number of children alive not more than four. The improvement could be achieved by creating an enabling environment for improvement of socio-economic development programs, well-controlled number of children each mother should have, the improvement of awareness campaigns on health issues and an importance of breastfeeding in a growth of the child.

Study by Lemani (2013) found that mother HIV status was significantly associated with infant mortality. Other factors found to be significantly associated with infant and child mortality were: mother's education, wealth index, sex of the child, mother's age and child birth order. None of the environmental factors were found to be significantly associated with both infant and child mortality. The current study also found that none of the environmental factors were associated with under-five mortality. Factors such as mother's education, wealth index and sex of child were found to be insignificantly associated with under-five mortality. The study also found that child care variable such as breastfeeding was significantly associated with under-five mortality while in the study by Lemani (2013) this factor was not included. This child care variable breastfeeding was found to be significantly associated with child mortality in different settings (Mekonnen, 2011; Mustafa and Odimegwu, 2008). This study further found that HIV status of a mother was significantly associated with the under-five mortality.

There are avenues for further work on this study. Future studies could be done is focus on the major occurrence of the under-five mortality contributing to the community in Tanzania by considering spatial analysis. We hope to extend this study by considering the generalized additive mixed model to include random effects in the generalized additive model and also to account for the missing values than refit the models. The joint modeling may also be considered, such as considering malnutrition, education and other variables to be modeled simultaneously with under-five child mortality.

References

Agwanda, A. and Amani, H. (2014). Population growth, structure and momentum in tanzania.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Allison, P. D. (2012). *Logistic regression using SAS: Theory and application*. SAS Institute.

An, A. B. (2002). Performing logistic regression on survey data with the new surveylogistic procedure. In *Proceedings of the twenty-seventh annual SAS® users group international conference*, pages 258–27. SAS Institute Inc. Cary, NC.

Archer, K. J., Lemeshow, S., and Hosmer, D. W. (2007). Goodness-of-fit tests for logistic regression models when data are collected using a complex sampling design. *Computational Statistics & Data Analysis*, 51(9):4450–4464.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., and White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, 24(3):127–135.

Commission, Z. A. (2013). Tanzania hiv/aids and malaria indicator survey 2011/2012.

Coovadia, H. M., Rollins, N. C., Bland, R. M., Little, K., Coutsoudis, A., Bennish, M. L., and Newell, M.-L. (2007). Mother-to-child transmission of hiv-1 infection during exclusive breastfeeding in the first 6 months of life: an intervention cohort study. *The Lancet*, 369(9567):1107–1116.

Czepiel, S. A. (2002). Maximum likelihood estimation of logistic regression models: theory and implementation. *Available at czep. net/stat/mlelr. pdf*.

Dagne, T. (2011). *Democratic Republic of Congo: Background and current developments*. DIANE Publishing.

David, K. and Mitchel, K. (1994). *Logistic regression: A self learning text*. New York: Springer–Verlag Inc.

Dobson, A. J. and Barnett, A. (2008). *An introduction to generalized linear models*. Chemical Rubber Company press.

Ettarh, R. and Kimani, J. (2012). Determinants of under-five mortality in rural and urban kenya. *Rural Remote Health*, 12:1812.

Factbook (2015). *Tanzania Demographics Profile 2014*. `http://www.cia.gov/library/publications/the-world-factbook/geos/tz.html`.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical science*, 43:297–310.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*. Chemical Rubber Company Press.

Hawkes, C. and Ruel, M. (2006). The links between agriculture and health: an intersectoral opportunity to improve the health and livelihoods of the poor. *Bulletin of the World Health Organization*, 84(12):984–990.

Hosmer, D. W. and Lemeshow, S. (2004). *Applied logistic regression*. John Wiley & Sons.

Jen, T.-H., Tam, H.-P., and Wu, M. (2011). An estimation of the design effect for the two-stage stratified cluster sampling design. *Journal of Research in Education Sciences*, 56(1):33–65.

Kalton, G., Brick, J. M., and Lê, T. (2005). Estimating components of design effects for use in sample design. *Household Sample Surveys in Developing and Transition Countries*, pages 95–121.

Kincaid, C. (2005). Guidelines for selecting the covariance structure in mixed model analysis. In *Proceedings of the Thirtieth Annual SAS Users Group International Conference*, number 198-30. SAS Institute Inc Cary, NC.

Kwesigabo, G., Mwangu, M. A., Kakoko, D. C., and Killewo, J. (2012). Health challenges in tanzania: Context for educating health professionals. *Journal of public health policy*, pages S23–S34.

Lee, E. S. and Forthofer, R. N. (2005). *Analyzing complex survey data*. Sage Publications.

Lemani, C. (2013). *Modelling covariates of infant and child mortality in Malawi*. PhD thesis, University of Cape Town.

Lemeshow, S. and Hosmer, D. (2000). *Applied Logistic Regression (Wiley Series in Probability and Statistics*. Wiley-Interscience Hoboken.

Littell, R. C., Pendergast, J., and Natarajan, R. (2000). Tutorial in biostatistics: modelling covariance structure in the analysis of repeated measures data. *Statistics in medicine*, 19(1793):1819.

Liu, H. (2008). Generalized additive model. *University of Minnesota Duluth, Duluth*.

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

Manning, C. (2007). Generalized linear mixed models (illustrated with r on bresnan et al. datives data). *Unpublished handout. http://nlp. stanford. edu/˜ manning/courses/ling289/GLMM. pdf*.

Marx, B. D. and Eilers, P. H. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis*, 28(2):193–209.

Masanja, H., de Savigny, D., Smithson, P., Schellenberg, J., John, T., Mbuya, C., Upunda, G., Boerma, T., Victora, C., Smith, T., et al. (2008). Child survival gains in tanzania: analysis of data from demographic and health surveys. *The Lancet*, 371(9620):1276–1283.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models.* Cyclic Redundancy Check press.

McCulloch, C. E. and Neuhaus, J. M. (2001). *Generalized linear mixed models.* Wiley Online Library.

Mekonnen, D. (2011). Infant and child mortality in ethiopia. *The role of Socioeconomic, Demographic and Biological factors in the previous five years period of 2000 and 2005.*

Moeti, A. (2010). *Factors affecting the health status of the people of Lesotho.* PhD thesis, University of KwaZulu-Natal.

Molenberghs, G. and Verbeke, G. (2006). *Models for discrete longitudinal data.* Springer Science & Business Media.

Mustafa, H. E. and Odimegwu, C. (2008). Socioeconomic determinants of infant mortality in kenya: analysis of kenya dhs 2003. *J Humanit Soc Sci*, 2(8):1934–722.

Olsson, U. (2002). Generalized linear models. *An applied approach. Studentlitteratur, Lund*, 18.

Park, I. and Lee, H. (2001). The design effect: Do we know all about it. In *Proceedings of the Annual Meeting of the American Statistical Association, August*, pages 5–9.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Shackman, G. (2001). Sample size and design effect. *Albany Chapter of the American Statistical Association.*

Siller, A. B. and Tompkins, L. (2006). The big four: analyzing complex sample survey data using sas®, spss®, stata®, and sudaan®. In *Proceedings of the Thirty-first Annual SAS® Users Group International Conference*, pages 26–29.

Šimundić, A.-M. (2008). Measures of diagnostic accuracy: basic definitions. *Med Biol Sci*, 22(4):61–5.

Sukaina, N. (2009). *Socio-demographic, socio-economic and household environmental characteristics associated with diarrheal disease among children under-five years of age in ethiopia.* http://www.statistics.su.se/polopoly_fs/1.178916.1401264593!/menu/standard/file/Master_thesis_Sukaina_Nasser.pdf.

Susuman, A. S. (2012). Child mortality rate in ethiopia. *Iranian journal of public health*, 41(3):9.

TACAIDS, Z. and NBS OCGS, I. (2013). Tanzania hiv/aids and malaria indicator survey 2011–12. *Dar es Salaam, Tanzania. Dar es Salaam, Tanzania: Tanzania Commission for AIDS (TACAIDS), Zanzibar AIDS Commission (ZAC), National Bureau of Statistics (NBS), Office of the Chief Government Statistician (OCGS), and ICF International.*

TACAIDS, Z. A. (2013). Commission, national bureau of statistics, office of the chief government statistician, icf international (2013) hiv/aids and malaria indicator survey 2011-12. dar es salaam, tanzania. *Dar es Salaam, Tanzania: Tanzania Commission for AIDS, ZAC, NBS, OCGS, and ICF International.*

UNICEF (2012). Levels & trends in child mortality, estimates developed by the un inter-agency group for child mortality estimation, report 2010.

Vittinghoff, E., Glidden, D. V., Shiboski, S. C., and McCulloch, C. E. (2011). *Regression methods in biostatistics: linear, logistic, survival, and repeated measures models.* Springer Science & Business Media.

Waagepetersen, R. (2007). *Computation of the likelihood function for GLMMs.* http://people.math.aau.dk/~rw/Undervisning/Topics/Handouts/6.hand.pdf.

Walker, C. L. F., Perin, J., Aryee, M. J., Boschi-Pinto, C., and Black, R. E. (2012). Diarrhea incidence in low-and middle-income countries in 1990 and 2010: a systematic review. *BMC public health*, 12(1):220.

Wood, S. (2006). *Generalized additive models: an introduction with R.* Chemical Rubber Company press.

Wood, S. N. (2012). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467).

Wood, S. N. and Augustin, N. H. (2002). Gams with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological modelling*, 157(2):157–177.

Yee, T. W. and Mitchell, N. D. (1991). Generalized additive models in plant ecology. *Journal of vegetation science*, pages 587–602.

Appendix A

Different Codes Used

A.1 Logistic Regression SAS Code

The variable used to fit the models are described below in full names.
Wi: Mother's wealth index, SDW: Source of drinking water, AHSH: Age of household head, BF: Currently breastfeeding, MS: Current marital status of the mother, HS: HIV status of a mother, BON: Birth order number of the child, BOC: Number of children ever born, MA: Mother's age, MF: Main floor material, EL: Mother's education level, C5: Number of children 5 and under in a household, S: sex of the child, CW: Currently working, CL: Number of children alive and L: Type of place of residence.

The Output Delivery System (ODS) is used to create the output and can be displayed graphically or in Hypertext Markup Language (HTML). PROC LOGISTIC fits the linear logistic regression for binary response assuming data is from a simple random sample using maximum likelihood. The option DATA=Tanzania specifies the dataset name that is of interest. The option DESCENDING is used to model 1,s instead of 0's, and by default SAS PROC LOGISTIC model 0's. The statement CLASS informs SAS of variables with categorical (An, 2002). Under the statement model, the link function is specified as LOGIT since response is binary and LACKFIT request Hosmer-Lemeshow test statistics to be produced and @2 request two-way interaction. The statement SELEC-

TION=STAPWISE allows automatic selection of variable to be included in the model and another option is to use backward or forward selection. The SAS code is as follow.

```
ods graphics on; ods HTML; PROC LOGISTIC DATA = data ; CLASS WI AHSH
BF SDW MS HS BON BOC MA MF EL C5 S CW CL L ;MODEL
Y(event='yes')=WI|AHSH|BF|SDW|MS|HS|BON|BOC|MA|MF|EL|C5|S|CW|CL|L@2/LINK=LOGIT
SELECTION=STEPWISE LACKFIT expb; RUN; ods HTML close; ods graphics
off;
```

A.2 Survey Logistic Regression SAS Code

The survey logistic procedure is used as an alternative to logistic regression procedure to capture survey design. PROC SURVEYLOGISTIC fits linear logistic regression model for binary response survey data using a method of maximum likelihood. This procedure incorporates survey design such as stratification, clustering, and unequal weighting. The option descending is included here to model 1s rather than 0s. All categorical variables are included in class the class statement. This informs SAS about the variables with different levels. The link function is LOGIT and EXPB give the odds ratios.

```
ods graphics on; ods HTML; PROC SURVEYLOGISTIC DATA = data ; ods
output oddsratios=domainors ; STRATUM V023; CLUSTER V021; Weight
V005; CLASS WI AHSH BF SDW MS HS BON BOC MA MF EL C5 S CW CL L ;
MODEL Y(event='yes')=  BF HS BON BF*BON BOC BF*BOC MA BF*MA C5 CL
BF*CL BON*CL / LINK=LOGIT EXPB; RUN; ods HTML close; ods graphics
off;
```

A.3 Generalized Linear Mixed Model SAS Code

Generalized linear mixed model can be fitted using GLIMMIX or NLMIXED SAS procedures. Both of these procedures offer similar syntax. PROC GLMMIX fits a linear logistic model with random and fixed effects. The method is specified under the statement method (METHOD=laplace) and option PLOTS=All produces required plots. The CLASS statement is also used to specify categorical variables. The distribution is specified as BINOMIAL and option SOLUTION and ODDS RATIO request SAS to produce solution for fixed effects and corresponding odds ration. The random statement is used to specify variables that are considered as random. The code for GLMM is given as.

```
proc glimmix data=data Method=LAPLACE; Class WI AHSH BF SDW MS HS
BON BOC MA MF EL C5 S CW CL L ; MODEL Y(event='yes')=  BF HS BON
BF*BON BOC BF*BOC MA BF*MA C5 CL BF*CL BON*CL / LINK=LOGIT
ODDSRATIOS Solution; LSMEANS BF HS BON BF*BON MA BF*MA C5
CL BF*CL / PLOT=DIFFPLOT ADJUST=TURKEY ALPHA=0.05;\\
LSMEANS BF HS BON BF*BON MA BF*MA C5 CL BF*CL / PLOTS=ANOMPLOT
ADJUST=NELSON ALPHA=0.05; RANDOM INT/ SUBJECT=cluster; run;
```

A.4 Generalized Additive Model SAS Code

The methodology behind GAM procedure relaxes the linearity assumption, this allows the hidden structure of the relationship between dependent variable and independent variables to be discovered. PROC GAM fit a logistic additive model with binary response variable child survival status and other predictors. Each term is fitted using B-spline smoother with default degrees of freedom which is 3. The class statement is also used here as before. However, in the model they are in included inside key word PARAM and continuous predictors are included inside keyword SPLINE. The output statement is used to obtain estimated functions and confidence intervals. The code for GAM is given as.

```
ods graphics on; ods html proc gam data=Data desc
plots=components(clm commonaxes); class HS BF; model Y =param(BF)
param(HS) spline(BN) spline(NLC) spline(RCA)
spline(NC5U)/dist=binomial; run; ods html close; ods graphics off;
```

Appendix B

Derivation of Some Properties of the Exponential Family

B.1 Properties of the Exponential Family

It is possible to get the general expression for the mean and the variance of the exponential distribution in terms of a, b and ϕ.

$$f(y, \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + C(y, \phi)\right\}$$

where $f(y, \theta, \phi)$ is the density function.

$$\int f(y, \theta, \phi) dy = 1.$$

Differentiating both side with respect to θ we get

$$\frac{\partial}{\partial\theta}\Bigg[\int \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + C(y,\phi)\right\}dy\Bigg] = 0,$$
$$\int \frac{\partial}{\partial\theta} \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + C(y,\phi)\right\}dy = 0,$$
$$\int \left[\frac{y - b^{'}(\theta)}{a(\phi)}\right] f(y,\theta,\phi)dy = 0,$$
$$\int \frac{yf(y,\theta,\phi)}{a(\phi)}dy - \int \frac{b^{'}(\theta)f(y,\theta,\phi)}{a(\phi)}dy = 0,$$
$$\int \frac{yf(y,\theta,\phi)}{a(\phi)}dy = \int \frac{b^{'}(\theta)f(y,\theta,\phi)}{a(\phi)}dy,$$
$$\int yf(y,\theta,\phi)dy = \int b^{'} f(y,\theta,\phi)dy,$$
$$\int yf(y,\theta,\phi)dy = b^{'}(\theta)\int f(y,\theta,\phi)dy,$$
$$E(y) = b^{'}(\theta)\times 1, \text{since} \int f(y,\theta,\phi)dy = 1,$$
$$E(y) = b^{'}(\theta)\text{is the mean of y.}$$

Taking the second derivative with respect to θ we obtain

$$\int \left[\frac{y - b^{'}(\theta)}{a(\phi)}\right] f(y,\theta,\phi)dy = 0,$$
$$\int \left\{\left[\frac{y - b^{'}(\theta)}{a(\phi)}\right]^2 f(y,\theta,\phi) - \frac{b^{''}(\theta)}{a(\phi)} f(y,\theta,\phi)\right\}dy = 0,$$
$$\int \left[\frac{y - b^{'}(\theta)}{a(\phi)}\right]^2 f(y,\theta,\phi)dy = \frac{b^{''}(\theta)}{a(\phi)}\int f(y,\theta,\phi)dy,$$
$$\frac{1}{a(\phi)^2}\int [y - b^{'}(\theta)]^2 f(y,\theta,\phi)dy = \frac{b^{''}(\theta)}{a(\phi)},$$
$$\frac{\mathrm{Var}(y)}{a(\phi)^2} = \frac{b^{''}(\theta)}{a(\phi)},$$
$$\mathrm{Var}(y) = a(\phi)b^{''}(\theta).$$

B.2 Statistical Inference

$$L(\theta, y) = \prod_{i=1}^{N} \exp\left\{\frac{y_t\theta_i - b(\theta_i)}{a_i(\phi)} + C(y,\phi)\right\}. \tag{B.1}$$

$$l(\theta, y) = \sum_{i=1}^{N} \left\{\frac{y_t\theta_i - b(\theta_i)}{a_i(\phi)} + C(y,\phi)\right\}. \tag{B.2}$$

Taking partial derivatives with respect to θ, we obtain the score function given by

$$\mathrm{U} = \frac{\partial l}{\partial \theta} = \sum_{i=1}^{N} \frac{y_i - b^{'}(\theta_i)}{a_i(\phi)}. \tag{B.3}$$

Taking the expected value of the score function and equate to zero, we have

$$\mathrm{E(U)} = \sum_{i=1}^{N} \frac{\mathrm{E}(y_i) - b^{'}(\theta_i)}{a_i(\phi)} = 0.$$

The information which is the variance of the score function is given by

$$I = \mathrm{Var(U)} = \sum_{i=1}^{N} \frac{\mathrm{Var}(y_i)}{a_i(\phi)^2} = n\frac{a(\phi)b^{''}(\theta)}{a(\phi)^2} = n\frac{b^{''}(\theta)}{a(\phi)}$$

where the derivative of the score function with respect to θ is given by

$$\mathrm{U}^{'} = \frac{\partial \mathrm{U}}{\partial \theta} = -n\frac{b^{''}(\theta)}{a(\phi)}.$$

This means that

$$\mathrm{Var(U)} = -\mathrm{U}^{'} = -\frac{\partial \mathrm{U}}{\partial \theta}.$$

For generalized linear models (GLMs) $y_i, i = 1, 2, \ldots, n$ is distributed as

$$f(y_i; \theta, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + C(y_i,\phi)\right\}. \tag{B.4}$$

The score function is given by

$$\mathrm{U} = \frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \frac{(y_i - \mu_i)}{a_i(\phi)} \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \frac{(y_i - \mu_i)}{a_i(\phi)} \frac{\mathrm{G}^{'}(\eta_i)}{\mathrm{Var}(\mu_i)} \boldsymbol{X_i}$$

where, $\eta_i = \mathrm{g}(\mu_i) = \boldsymbol{X_i}$. The information matrix is obtained by finding the second derivative and is given by

$$\mathbf{I} = \mathrm{Var}(\boldsymbol{U}) = \sum_{i=1}^{N} \frac{\mathrm{G}'(\eta_i)}{\mathrm{Var}(\mu_i)} \boldsymbol{X_i X_i'} \tag{B.5}$$

where score vector can be viewed as $\boldsymbol{U} \sim \mathrm{MVN}_p(\mathbf{0}, \boldsymbol{I})$. Thus

$$Q = \boldsymbol{U I^{-1} U'} \sim \chi^2(p).$$

Sampling distribution of the Maximum Likelihood Estimator(MLE)

The Taylor series expansion of the function $f(x)$ about $x = a$ is given by

$$\begin{aligned} f(x) &= f(a) + (x-a)f'(a) + \frac{1}{2}(x-a)^2 f''(a) + \frac{1}{3}(x-a)^3 f'''(a) + \ldots \\ &\approx f(a) + (x-a)f'(a) \end{aligned} \tag{B.6}$$

so that the Taylor series expansion of the score vector $\boldsymbol{U(\beta)}$ about $\hat{\boldsymbol{\beta}}$ becomes

$$\boldsymbol{U(\beta)} \approx \boldsymbol{U(\hat{\beta})} + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\frac{\partial U(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} = \boldsymbol{U(\hat{\beta})} + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\boldsymbol{U'(\hat{\beta})}.$$

However, $\boldsymbol{U(\hat{\beta})} = 0$ we have that $\boldsymbol{U(\beta)} \approx (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\boldsymbol{U'(\hat{\beta})}$. If $\boldsymbol{U'}$ is approximated by $E(\boldsymbol{U'}) = -\mathrm{Var}(\boldsymbol{U}) = -\boldsymbol{I}$ then we have that $\boldsymbol{U(\beta)} \approx (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\boldsymbol{I}$ which is

$$\boldsymbol{I}^{-1}\boldsymbol{U(\beta)} \approx (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \tag{B.7}$$

Taking the expected value in equation B.7 we get

$$\mathrm{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \boldsymbol{I}^{-1}\mathrm{E}(\boldsymbol{U(\beta)}) = 0.$$

This implies that $\mathrm{E}(\boldsymbol{\beta}) = \boldsymbol{\beta}$, so $\boldsymbol{\beta}$ is the consistent estimator of $\boldsymbol{\beta}$. The variance is thus given by

$$\begin{aligned} \mathrm{Var}(\boldsymbol{\beta}) &= \mathrm{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{'}], \\ &= \mathrm{E}[(\boldsymbol{I}^{-1}\boldsymbol{U}(\boldsymbol{\beta}))(\boldsymbol{I}^{-1}\boldsymbol{U}(\boldsymbol{\beta}))^{'}], \\ &= \boldsymbol{I}^{-1}\mathrm{E}[\boldsymbol{U}(\boldsymbol{\beta}))(\boldsymbol{U}(\boldsymbol{\beta}))^{'}]\boldsymbol{I}^{-1}, \\ &= \boldsymbol{I}^{-1}\mathrm{Var}(\boldsymbol{U}(\boldsymbol{\beta}))\boldsymbol{I}^{-1}, \\ &= \boldsymbol{I}^{-1}\boldsymbol{I}\boldsymbol{I}^{-1}, \\ &= \boldsymbol{I}^{-1}. \end{aligned} \tag{B.8}$$

So $\boldsymbol{\beta} \sim \mathbf{MVN}(\boldsymbol{\beta}, \boldsymbol{I}^{-1})$ and we can have that

$$Q = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{'} \sim \chi^2(p)$$

which is known as the Wald statistics.

Appendix C

Additional Results

C.1 Checking Multicolinearity

Table C.1: Checking the presence of multicolinearity.

Variable	Estimate	Standard Error	p-value	Tolerance	VIF
Intercept	0.00399	0.01939	0.8371	.	0
WI	-0.00361	0.00313	0.2475	0.45917	2.17782
AHSH	0.00102	0.00428	0.8109	0.82673	1.20959
BF	-0.03372	0.00397	0.0001	0.90208	1.10855
SDW	0.00144	0.0039	0.7117	0.92567	1.08029
MS	-0.00933	0.00451	0.0385	0.94112	1.06256
HS	0.02762	0.0092	0.0027	0.97097	1.02989
BON	-0.00985	0.00526	0.061	0.17365	5.75872
BOC	0.10066	0.00641	0.0001	0.12493	8.00429
MA	-0.00412	0.00483	0.3935	0.55471	1.80274
MF	-0.00318	0.00622	0.6092	0.4714	2.12133
EL	0.00477	0.00356	0.1805	0.81454	1.22768
C5	-0.00952	0.0035	0.0066	0.85816	1.16528
S	-0.00406	0.00375	0.2783	0.99808	1.00193
CW	-0.00981	0.00598	0.101	0.93317	1.07162
CL	-0.0662	0.00363	0.0001	0.1661	6.0204
L	-0.00437	0.00609	0.4724	0.72982	1.3702

C.2 Generalized Linear Mixed Model Results

Table C.2: Laplace, estimated coefficients, odds ratios, standard errors, p-values and confidence interval.

					95% confidence interval	
Effects	**Estimate**	**Odds ratio**	**Standard error**	**P-value**	**Lower limits**	**Upper limits**
Intercept						
Socio-demographic characteristics						
Breast feeding (BF)						
Yes(ref)						
No	-0.2471	0.7811	0.5213	0.4939	0.2812	2.1698
Mother's HIV status (HS)						
Negative(ref)						
Positive	0.5016	1.6514	0.1872	0.001	1.1442	2.3834
Birth order number (BON)						
Less than 2 birth(ref)						
2 to 4 birth	0.3548	1.4259	0.3085	0.0630	0.7789	2.6103
More than 4 birth	0.9807	2.6663	0.4291	0.0010	1.1499	6.1826
Respondent age(MA)						
Over 34 years (ref)						
20 to 34 years	0.4943	1.6394	0.3223	0.7199	0.8716	3.0833
Less than 20 years	-0.5537	0.5748	0.5375	0.4071	0.2004	1.6484
Socio-economic characteristics						
Number of children alive (CL)						
Less than 2 children(ref)						
2 to 4 children	1.1024	3.0114	0.2985	0.0001	1.6776	5.4057
More than 4 children	-0.8003	0.4492	0.3641	0.002	0.2200	0.9170
Children under-five years (C5)						
Less than 2 children(ref)						
2 to 4 children	-0.4269	0.6525	0.1667	0.0010	0.4707	0.9047
More than 4 children	-0.4948	0.6097	0.3097	0.4071	0.3323	1.1188
Interaction between Socio-demographic and Socio-economic characteristics						
Breast feeding by Birth order number						
yes versus less than 2 birth(ref)						
No Versus 2 to 4 birth	1.1400	3.1268	0.3673	0.0010	1.5221	6.4231
No versus more than 4 birth	2.0685	7.9129	0.5269	0.0001	2.8173	22.2251
Breast feeding by Respondent age						
Yes versus Over 34 years(ref)						
No versus 20 to 34 years	-0.4760	0.6213	0.3754	0.0550	0.2977	1.2966
No versus Less than 20 years	1.6743	5.3351	0.6333	0.0010	1.5419	18.4593
Breast feeding by children alive						
Yes versus less than 2 children(ref)						
No versus 2 to 4 children	1.2193	3.3848	0.3760	0.0200	1.6199	7.0728
No versus more than 4 children	-0.8060	0.4466	0.4246	0.4600	0.1943	1.0266

Table C.3: Breastfeeding by child birth order least square means.

BF*BON Least Squares Means				
breastfeeding	**birth order**	**Estimate**	**Standard Error**	**Lower**
No	2-4 births	-2.8162	0.2246	-3.2564
No	5 and above births	-1.2617	0.2292	-1.711
No	First births	-4.311	0.2463	-4.7937
Yes	2-4 births	-4.2463	0.2966	-4.8276
Yes	5 and above births	-3.6203	0.323	-4.2536
Yes	First births	-4.6011	0.3014	-5.192

Table C.4: Breastfeeding by mother's age least square means.

BF*MA Least Squares Means					
Breastfeeding	**Respondent age**	**Estimate**	**Standard Error**	**Lower**	**Upper**
No	Between 20-34	-3.1576	0.1774	-3.505	-2.81
No	less than 20 years	-2.0553	0.315	-2.673	-1.4379
No	more than 34	-3.1759	0.2112	-3.59	-2.762
Yes	Between 20-34	-3.6418	0.1751	-3.985	-3.2986
Yes	less than 20 years	-4.6898	0.4572	-5.586	-3.7935
Yes	more than 34	-4.1361	0.3163	-4.756	-3.5161

Table C.5: Breastfeeding by number of children alive least square means.

BF*CL Least Squares Means					
Breastfeed	**number of children alive**	**Estimate**	**Standard Error**	**Lower**	**Upper**
No	2 to 4 children	-0.7131	0.2126	-1.1299	-0.2963
No	above 4	-4.641	0.2765	-5.183	-4.0991
No	less than 2 children	-3.0348	0.2177	-3.4615	-2.608
Yes	2 to 4 children	-3.1542	0.3082	-3.7584	-2.5501
Yes	above 4	-5.0569	0.3392	-5.7217	-4.392
Yes	less than 2 children	-4.2566	0.2791	-4.8038	-3.7094