UNIVERSITY OF
KWAZULU-NATAL

STATISTICAL MODELS TO DETERMINE FACTORS AFFECTING UNDER-FIVE CHILD MORTALITY IN SOUTH AFRICA

By

Andisiwe Bovu

A thesis submitted to the

University of KwaZulu-Natal

In fulfilment of the academic requirements for the degree

of

MASTER OF SCIENCE IN STATISTICS

School of Mathematics, Statistics and Computer Science

College of Agriculture, Engineering and Science

February 2020

Declaration

I, Andisiwe Bovu, declare that this thesis submitted to the University of KwaZulu-Natal is my original work, except where otherwise indicated.

- This thesis does not contain other person's data, pictures, graphs or other information, unless specifically acknowledge as being sourced from the person.
- This thesis does not contain other person's writing unless specifically acknowledged as being sourced from other researchers.
- Where other written sources have been quoted, then:

  Their words have been re-written, but the general information attributed to them has been referenced

  Where their exact words have been used, then their writing has been placed in italics and inside quotation marks, and referenced

- This thesis does not contain text, graphics or tables copied and pasted from the internet, unless specifically acknowledged, and the source being detailed in the thesis and in the references section

30/10/2019

_____ _____

Mr Andisiwe Bovu (Student) Date

30/10/2019

_____ _____

Dr. SF Melesse (Supervisor) Date

Acknowledgements

I would like to thank my supervisor Dr. Sileshi Fanta Melesse, for giving me academic support to complete this thesis. Thank you for your invaluable guidance, time, advice and patience. This work would not be complete without your continued support.

Thank you to the National Research Foundation (NRF) for the financial assistance towards completion of my degree.

Lastly, I wish to thank all my friends and family especially my Mother, Brothers and Sister who were always there supporting and encouraging me to complete my degree. Special thanks to Bulelani Nangamso Pepeta and Mziwanda Mangwane for always being available to proofread and share constructive ideas with me, thank you a lot brothers, this is a start of journey of great things. To all of you this is the fruit of your moral and financial support.

Abstract

The level of under-five child mortality is an important indicator of economic, social and health development of the nation. In the last two decades, substantial progress has been made in improving under-five child mortality globally, with deaths dropping among children under the age of five years from approximately 12 million in 1990 to about 6.3 million in 2015. However, significant strides to address the key risk factors are still needed in the Sub-Saharan Africa region if they are to achieve the Sustainable Development Goals 2030. The key objective of the study is to identify key factors associated with mortality of children under the age of five years in South Africa. In order to identify these factors, the study used different statistical models that accommodate a binary response variable. Models used include Logistic Regression, Survey Logistic Regression, Generalized Linear Mixed Models and Generalized Additive Models. Although logistic regression is useful in modelling data with a dichotomous outcome, it is not suitable for modelling data obtained through a complex survey that incorporates weights, stratification and clustering. Survey logistic regression is used to model the relationship between binary dependent and the set of explanatory variables by making use of the sampling design information. In this case, the inclusion of random effects in the model results in generalized linear mixed models (GLMM). These models are an extension of linear mixed models that allow response variable from different distributions, such as binary responses. One can think of GLMM as an extension of generalized linear models (e.g. logistic regression) that combine both features of fixed and random effects. These statistical models assume linearity parametric form for the explanatory variable. However, this assumption of linear independence of response on covariates may not hold. Hence, we introduce generalized additive models (GAM). The GAM models show some non-linear relationship between the response variable and some covariates. The results showed that, the size of child at birth, breastfeeding, birth order number, ethnicity, number of children 5 under, total children ever born, source of drinking water and province were significantly associated with under-five child mortality. The study concludes that prolonged breastfeeding, improved health services and source of water are among the main factors to decline under-five child mortality further. Therefore, the study suggests that there is a need to strengthen child health interventions in South Africa to reduce the under-five mortality rate even more in order to achieve sustainable development goals (SDG) 2030.

Keywords: Under-five child mortality, Survey Logistic Regression, Generalized Linear Mixed Models and Generalized Additive Models.

Contents

List of Tables

List of Figures

Abbreviations

U5MR	Under five child mortality
WHO	World Health Organization
SADHS	South Africa Demographic and Health Survey
MDG	Millennium Development Goals
SDGs	Sustainable Development Goals
NDP	National Development Plan
HIV	Human Immunodeficiency Virus
AIDS	Acquire Immune Deficiency Syndrome
UN	United Nations
UNICEF	United Nations International Children's Education Fund
Stats SA	Statistics South Africa
DHS	Demographic Health Survey
SAMRC	South Africa Medical Research Council
NDoH	National Department of Health
MSF	Master Sample Frame
EAs	Enumeration Areas
PSUs	Primary Sampling Units
DUs	Dwelling Units
EDA	Exploratory Data Analysis
GLM	Generalized Linear Model
FS	Fisher Scoring
NR	Newton-Raphson
IWLS	Iterative Reweighted Least Squares
AIC	Akaike Information Criterion
BIC	Bayesian information criterion

SC	Schwarz Criterion
ROC	Receiver Operating Characteristics
AUC	Area under the curve
GOF	Goodness-of-fit
DEFF	Design effects
GLMMs	Generalized Linear Mixed Models
LMMs	Linear Mixed Modes
GAMs	Generalized Additive Models

Chapter 1

Introduction

1.1 Background of the study

African children face a higher risk of death as compared to European children. Under-five child Mortality rate (U5MR) was reported to be 5.5% of live births in Sub-Saharan Africa by World Health Organization (WHO), while in European countries it was found to be 1.0% of live births (Khodaee *et al*., 2015). In the early 1990s, South Africa had the majority of young people who had been systematically deprived access to proper services by apartheid policies. About 5.9% children would die earlier than their fifth birthday because of completely preventable causes (South Africa Demographic and Health Survey (SADHS), 1998), which amongst others include diarrhea, pneumonia and HIV/AIDS. The occurrences of these diseases were high and provision of essential preventative interventions such as immunization were limited (Maluleke and Chola, 2015). More than 30% of children under the age of one year were not immunized against measles, and even those that received the full suite of foremost vaccinations to protect them against preventable diseases were fewer than 66.4% of the children population.

South Africa was amongst 189 countries that supported the United Nation's Millennium Declaration (United Nations, 2000), committing to meeting eight goals, referred to as Millennium Development Goals (MDG) set to be attained by 2015, using 1990 as a base year (Statistics South Africa and National Treasury, 2007). The MDG 5 called for countries to improve maternal health and reduce child mortality by three-quarters (Ronsmans and Graham, 2006). Millennium Development Goals 4 called for a reduction in the child mortality rate by two-thirds (Statistics South Africa and National Treasury, 2007). Nevertheless, Bhutta *et al.* (2010) reported that maternal mortality and under-five child mortality in South Africa have both increased since the MDG baseline of 1990. The under-five mortality rate for South Africa was very high and the country was far from reaching the universal goal of 2.0% of deaths. An evaluation of South African provinces confirmed that Limpopo had the highest under-five child mortality rate of 11% which contribute 55% to the overall under-five mortality (MDG Country Report, 2010). Like many Sub-Saharan nations, South Africa by 2010 was failing to reach the target to minimize child mortality.

Sub-Saharan countries face a challenge of reducing under-five child mortality rate. Better health care services are not available to many South Africans living below breadline and government struggles to supply these primary services. In 2015, it was mentioned that globally under-five child mortality rates had declined between the period of 1990 and 2015. Over the 25-year period, global child mortality lowered from 9% to 4.3% (WHO and UNICEF, 2015). Reports from the United Nations (UN) have shown a decline in child mortality rates in South Africa from 6,1% of the total live births in 1990 to 4,5% of the total live births in 2015. However, in this regard, the overall performance of the country is low compared to the performance of many other countries. For instance, the world has made tremendous progress in reducing the under-five child mortality rate by 47% from 1990 to 2015. On the other hand, South Africa has attained a

reduction of about 26%, which makes the country very unlikely to obtain the MDG goal number 4. The objective was to have under-five mortality rates of 1.8% and 2.0% births respectively through 2015 (Stats SA, 2013). Although HIV/AIDS is typically quoted as the essential reason for this poor performance, poverty and inequality are some of the factors to which need not to be overlooked. It has been reported that the health of under-five children in South Africa is largely influenced by socioeconomic conditions under which they live and about 66% of kids in the country live in poverty, with a monthly family income of much less than R1200 (SADHS, 2016).

1.2 Problem statement

South Africa remains the country with the largest number of people living with HIV/AIDS (Kerber *et al*., 2013). Hence, South Africa was one of the four countries worldwide where the under-five child mortality rate was higher in 2005 than 1990 the MDG (Kerber *et al*., 2013). However, between the space of 2006 and 2010, South Africa reduced under-five child mortality rates by approximately 40%. Progress on reducing child mortality has taken place due to numerous policies and programme adjustments that have expanded coverage of child health interventions. These consist of the increase in Prevention of Mother-to-Child Transmission (PMTCT), Anti-Retroviral Therapy (ART) and immunization coverage (Nannan *et al*., 2012; UNICEF, 2013; Dorrington *et al*., 2014).

Even though South Africa has made developments towards lowering under-five child mortality in the past decade classically in 2005 to 2015 space of time frame, the country has not met the under-five child mortality target for the MDGs. Despite having invested substantially in programmes and policies aimed to achieve these targets (Mathews *et al.*, 2016). The levels of under-five child mortality reflect about the state of public health and hygiene of the environment people live in. Above all, people's attitude towards the dignity and value of human life itself necessitate research on the subject matter regarding their impact on under-five child mortality rates (Kyei, 2011).

Despite the progress and achievements in tackling under-five child mortality rates, challenges remain unconquered. Many under-five children mortalities are still affected by malnutrition, pneumonia, diarrhea and AIDS whereas there is a lot of funding to support research on these areas. This calls for multi-displinary collaboration as a first step to tackle the issue. While the infant mortality rate (IMR) has shown improvement, there has been very little progress with the neonatal mortality rate (NMR). Almost 40% all under-five deaths occur in neonates, and about 20 000 babies are stillborn every year (Nannan *et al*., 2012; Dorrington, 2014; Health Systems Trust, 2012).

1.3 Justification of the study

Under-five child mortality rate is a key indicator of both poor and unimproved population health and development. Generally, this reflects the socio-economic and environmental conditions in which children live. Nevertheless, since 1994, South Africa has developed policies, programs and special committees in effort to forestall childhood morbidity and mortality, and

with such life circumstances of South Africans have been improving. The identification of determinants of under-five child mortality is important for formulating appropriate health programmes and policies. A significant pace of decline in under-five child mortality can only be achieved if the strategies and policies against mortality are directed towards those factors associated with high mortality rates (Mustafa and Odimegwu, 2008). Therefore, this study will add on an ongoing body of research of under-five child mortality to help the South African government, non-governmental organizations and other partners in the health sector to know and understand the important areas they need to focus on. This research will also help government to develop more policies, programmes and projects to reduce under-five child mortality rates.

1.4 Objectives

The general objective of this study is to determine the factors affecting under-five child mortality in South Africa.

The specific objectives are:

- To determine the effect of socio-economic factors in under-five child mortality in South Africa.
- To determine the effect of demographic factors in under-five child mortality in South Africa
- To investigate the effect of the environmental factors in under-five child mortality in South Africa.

Chapter 2

Literature review

Chapter 2 provides an overview of child mortality of children under five years of age. The first section looks at the general overview of child mortality while the second one concentrates on worldwide overview of child mortality and then, the third one focuses on child mortality in South Africa. In the fourth section, the Mosley and Chen (1984) analytical framework of factors affecting child mortality: socio-economic, demographics and environmental factors are reviewed.

2.1 Under-five Child mortality

Under-five child mortality study has been researched by demographers due to the fact for a population to grow it is vital that children born live to tell the tale and develop to reproduce children of their own. Under-five child mortality rate is defined as the probability expressed as a rate per 1000 live births of a child born in a specific year dying before reaching the age of five years (United Nations, 2003*)." The under-five child mortality is categorized by neonatal (0 – 27 days), comprising of early neonatal (0-6 days) and late neonatal (7–27 days); post-neonatal (1–11 months); and children (1–4 years) (Nannan et al., 2012). The first 28 days of life is the most vulnerable time for a child's survival and this period is significantly important because a large portion of the under-five child death occurs during the neonatal period. Demographically, the key age groups are infants (0–12 months) measured by the infant mortality rate (IMR) and children (1–4 years) measured by the child mortality rate (CMR), both expressed as values per 1000 live births. Both age groups are inclusive of and measured by the under-five mortality rate"* (Nannan *et al*., 2012).

2.2 Worldwide overview on child mortality

The level of child mortality is a significant indicator of economic, social and health development of the nation. In 2000, world leaders united aiming to decrease child mortality by two thirds, in between 1990 to 2015, which was named as MDGs target 4. There has been notable progress in decreasing child mortality worldwide in the last 25 years. The chart below shows the notable progress between 1990 and 2015 in under-five child mortality by regions.

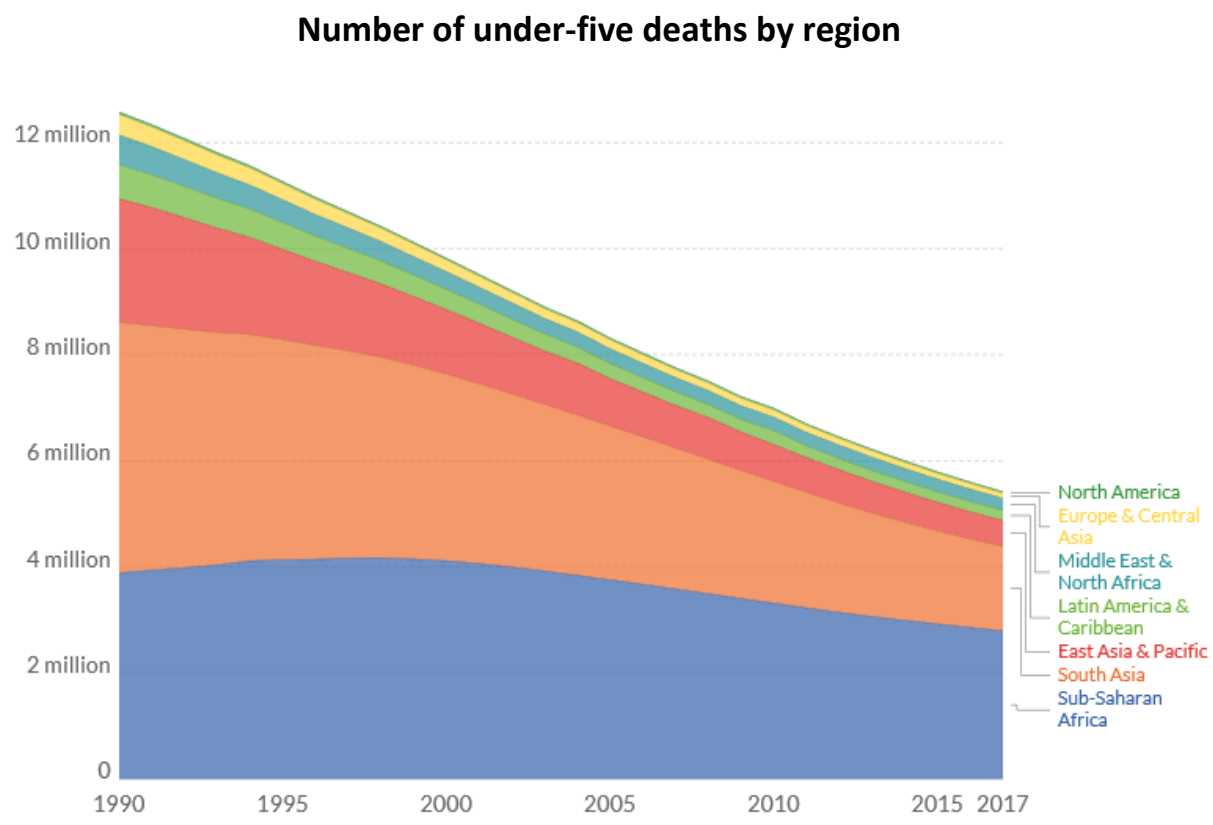

Figure 2. 1 Child deaths by region

Source: UN Inter-Agency Group for Child Mortality Estimation

In this period, under-five mortality reduced from 9.1% live births to 4.3% live births, which is a 53% drop. In the same period, the annual child mortality rate decreased from 12.7 million to 5.9 million. "*Since 1990, developed and developing countries both lowered child mortality significantly. Developed countries had low child mortality with 15 deaths per thousand live births in 1990, which was reduced to six in 2015, which is 60% decline from 1990, while developing countries had 100 deaths per thousand live births which was reduced to 47 in 2015, with 54% decline from 1990. The MDG target to reduce child mortality to five deaths per thousand live births was not met in 2015. Similarly, in developing countries, the target to reduce child mortality to 3.3% deaths per thousand live births was not achieved in 2015. Globally, the annual decline rate of child mortality increased from 1.8% in 1990 to 3.9% in 2015. Between these periods, accelerated progress was seen which was important for reducing child mortality (UN Inter-agency, 2015).*

Despite this progress, Sub-Saharan Africa still has high under-five child mortality, where one child in 12 dies before his or her fifth birthday. Another region with high under-five child mortality is Southern Asia, where one child in 19 had suffered death before reaching the age five. Whilst, high-income countries have an average ratio of 1 in 147 deaths before their fifth birthday (UN Inter-Agency, 2015). East Asia, the Pacific and Latin America met region wise MDG 4 target,

and the Caribbean based on point estimation. Globally, 62 countries met the target, among which 24 countries belong to low and middle income. It is estimated that in certain countries between 2016 and 2030, 94.4 million children will have to suffer death before reaching the age of five. Even if the countries will meet the sustainable development goals (SDGs) of reducing under-five mortality to 25 or below per thousand live births by 2030, the projected 56 million deaths will still stand. Sub-Saharan Africa needs to accelerate the progress along with South Asia where urgent action is needed to improve the scenario "(You *et al*., 2015).

2.3 Child mortality in South Africa

South Africans are now facing challenges for which the highest caliber of leadership, vision and commitment is needed. The large increase in under-five child mortality and morbidity are threatening to overpower the health system and undermine the potential of South Africa to attain the (MDGs) (UNICEF, 2009).

South Africa falls into group of countries in which the vital statistics are not yet of a high enough quality to produce reliable estimates of child mortality directly (Nannan *et al*., 2012). However, compared to many other countries, in which there have generally been little or no improvements in the vital registrations (Setel *et al*., 2007), great strides have been made in improving population health statistics in South Africa. In terms of vital registration, a new death notification form was introduced in 1998, which complied with WHO standards for the certification of cause of death and was accompanied by efforts to extend death registration to all areas of the country.

This improved availability of cause of death statistics in South Africa has occurred during a period of deep change in under-five child mortality due to the unfolding HIV/AIDS epidemic. Bradshaw and Nannan (2006) noted that HIV/AIDS has resulted in increased under-five child mortality rate in South Africa. However, reliable data are absent, as there has been no way to assess the extent to which improved death registration has contributed to the apparent increase in the number of child deaths.

The level of under-five child mortality is one of the indicators of the level human development, hence its inclusion in the construction of human development indices (HDI) and the multi-dimensional approach to combat high level of poverty (Motshwaedi, 2011). (Lagerdien, 2005) argues that under-5 child mortality is the most crucial element of a child's right in South Africa. Child mortality indicators are universally accepted measures of whether a country is meeting its obligation towards children. The country's socio-economic status can be derived from the level of indicators for the children under-five mortality. Demographic Health Survey 1998 indicated that three-quarters of all under-five deaths occur in the first year of life and one thirds occur in the first month of life.

Since a good start in life is critical to the physical, intellectual and emotional development of every individual, poverty in early childhood can prove to barrier for life. Poverty denies children

their rights to basic education, primary health, adequate nutrition and safe water and sanitation (UNICEF, 2007).

The government of South Africa has succeeded in reducing income poverty considerably since 1994 through a three-fold increase in social grants expenditure, the number of beneficiaries, and introducing the child support grant in 1998 (UNICEF, 2007). The prevention of the mother to child transmission programmes reduced HIV transmission from mother to child; however, this has not been sustained owing to limitations in community-based child health care services.

"Attempts have been made to quantify child poverty in South Africa. According to the report South African Child Gauge 2006 by the Children's institute at the University of Cape Town:

- *Fifty-five percent of children belong to the households living under the ultra-poverty line of R800 or less a month and this amount to 10 million children.*
- *The Eastern Cape and Limpopo presented the most poverty-stricken profiles, with close to three quarters of children living under the ultra-poverty line.*
- *The poorest provinces were found to be those with large rural populations and little access to employment opportunities*
- *63 percent of African children lived in ultra-poor households*" (UNICEF, 2007).

The under-five child mortality has declined, however; children continue to die from preventable and treatable causes of death. The next section focuses on factors affecting child mortality using analytical framework by considering the socio-economic level of households, environmental and the demographic factors.

2.4 Factors affecting child mortality

Many factors contribute to the level of child mortality. Mosley and Chen (1984) revealed that all social and economic determinants of child mortality necessarily operate through a common set of proximate determinants, to exert an impact on mortality. In this framework, a set of proximate determinants or intermediate variables that directly influence the risk of morbidity and mortality are identified. All social and economic determinants must operate through these variables to affect child survival. This study adopted the Mosley and Chen (1984) approach to the analysis of child mortality.

Figure 2.2 illustrates the path to a healthy child or a sick child and eventual child mortality. The analytic framework proposed by socioeconomic factors operates through demographic factors, environmental factors leading to a healthy child or sick child. However, with modern medical intervention (through prevention or treatment), a child may remain healthy, the sick child could recover and become healthy or treatment may fail and the child dies. Each of the factors is discussed below.

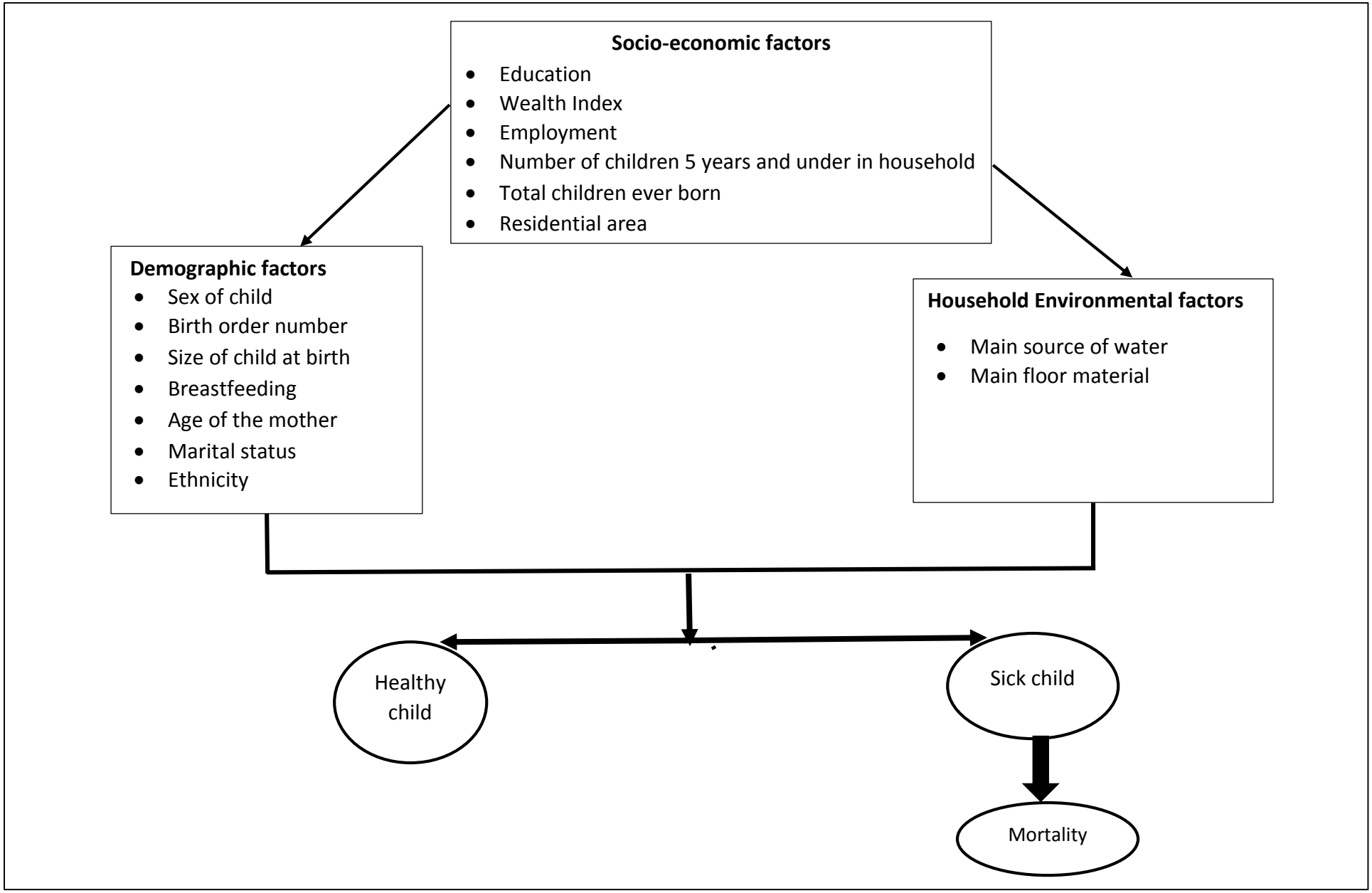

Figure 2. 2 Analytical framework for child survival

Adapted from Mosley and Chen (1984)

2.4.1 Socio-economic factors

Quite a number of researchers (Cleland, 1989; Hobcraft, 1993; Sastry, 1996; Wagstaff, 2000; Bawah and Zuberi, 2005; Mustafa and Odimegwu, 2008) has extensively reviewed the relationship between child survival and socio-economic factors. The present study ascertain research considers socio economic factors that has been adopted from the Mosley and Chen (1984) framework, which are mother's education, wealth index, employment and place of residence

Education

Mother's education is regarded as the most important socio-economic variable affecting child survival. It may affect child survival by influencing her alternatives and enhancing her ability on different health care practices. The relation has gained the attention of many researchers. According to Hobcraft (1993), an increased chance of child survival can even be associated with a small increase in the education level of mothers. Cleland (1989) demonstrated that the enhancement of child survival due to mother's education is because of the modest effect of education on health knowledge and beliefs. As births to teenage mothers are related to a high risk of mortality at the early ages of the child, mother's education has an advantage by delaying the age at first birth. Moreover, it may be related to factors that form and alter the economic choices and health related practices of individuals. However, some studies have shown that the degree of association in Sub-Saharan Africa is weaker than in other regions. The reason for this is the weak health infrastructure in the sub-continent (Hobcraft, 1993).

Bello and Joseph (2014) concluded using Twum-Baah *et al*., (1994) findings that children born to mothers with higher educational levels are associated with lower hazard of under-five child mortality as compared to children born to mothers with primary educational level or non-educated. Wambugu (2014) outlined that the educational status of mothers had a significant impact on child mortality in relation to results obtained from Goro (2007) when using a multiple regression model. It was evident that the educational status of mothers had a significant impact on child mortality. Kumar and Gemechis (2010) stipulated that a mother's educational status has a significant impact on lowering the risk of child mortality.

Wealth index

Improving the health of the poor and reducing health disparities between the poor and non-poor have become central goal of certain international organizations, including the World Bank and WHO (Wagstaff, 2000). Since 1997, the top priority of the World Bank has been to work with countries like South Africa to improve the health, nutrition and population outcomes of the world's poor, to protect the population from the impoverishing effects of illness, malnutrition and high fertility (Wagstaff, 2000). In South Africa, urban households are more likely than non-urban households to fall into the higher wealth quintiles, while non-urban households are more likely to fall into the lower wealth quintiles (SADHS, 2016). The reason is that; South Africa is still

a developing country. Rural mothers and children are often disadvantaged in term of access to basic health services that can lead to under-five child mortality.

Employment

The household income level is also an important factor in determining the level of child mortality amongst children under-five. In circumstances where there are low income earnings or no income, there will be limited or no access to the basic needs such as hygiene, shelter and food. Hence, in a society where inequality is rife and gap between the rich and poor is widening, the disparities in affordability and accessibility to fundamental human needs may also result into rising rates of malnutrition and under-five child mortality. For example, in Kenya most socio-economic factors are not related with high risk of under-five child mortality though children born in the richest households automatically have a lower probability of mortality relative to children born in the poorer households (Mustafa and Odimegwu, 2008).

Number of children 5 years and under in household and Total children ever born

Family size has been found to influence under-five child mortality. According to (Woldemicael, 2001; Mambugu, 2014) when many children live together, the chance of contact with germs increases and hygiene may deteriorate. Many children in a household increases the likelihood of having disease like infections because of crowding and competition for the mother's time. (Woldemicael, 2001 and Wambugu, 2014) have found that in Eritrea the probability of having diarrhea is about 60% higher if there are six or more children living in the household than if the number is less than three. While in Ethiopia, the odds of having infections are linked with the number of children remained significant even after controlling for all environmental, behavioral and other socio-economic variables considered in a study (Wambugu, 2014).

Residential area

Place of residence influences under-five child mortality especially with rural and urban area difference. Some studies found the risk of death of children is lower in the urban areas compared with rural areas (Kabir *et al*., 2001; Kembo and Ginneken, 2009). This is the general expectation considering that the level of development is more advanced for urban than for rural areas. One of the major studies done from DHS data in fifty-four low and middle-income countries from 2005 to 2013 showed under-five mortality was higher in rural areas. Among fifty-four countries, half of them had under-five child mortality of 84 per 1000 live births in rural areas and 61 per 1000 live births in urban areas. The magnitude of difference in under-five child mortality in rural and urban areas of those countries varied ranging from 16 per 1000 live births to 50 per 1000 live births in countries like Cameron, Burundi. Certain countries like Ukraine, Jordan, showed low rural-urban differences in under-five mortalities 3 per 1000 live births (WHO, 2016). One of the DHS studies from Brazil had shown urban areas had low child mortality. The differences were not clarified through urban life advantage, but community variables such as ecological setting, political economy and health system, played an important role through socioeconomic characteristics (Fotso, 2006). However, some studies that have found

contradictory results, where children in rural areas have a lower risk of dying than their urban counterparts have (Manda, 1998). Such findings are quite unusual considering that urban areas are associated with better socioeconomic and environmental factors that contribute to the reduction of under-five child mortality. A study of DHS on forty-seven developing countries had shown that the poor population in urban areas had higher under-five child mortality than rural areas in some countries, which was linked with high population growth in urban areas (Poel *et al*., 2007).

In Zimbabwe, child mortality variations exist between urban-rural residence because of regional disparities and availability of healthcare infrastructure (Zimbabwe Central Statistical Office/ Macro International Inc., 2007). Sahn and Stifle (2003) utilized data from DHS of 24 African countries and concluded that under-five child mortality in urban areas is lower relative to in rural areas. However, it should be noted that the HIV/AIDS epidemic is partially responsible for the high risk of child mortality in Africa, especially in sub-Saharan African countries. Generally, it is assumed that under-five child mortality in urban areas is lower than observable levels in rural areas.

2.4.2 Demographic factors

Several demographic factors have been considered in studies of under-five child mortality. These factors, also referred to as maternal factors have an independent direct influence on pregnancy and child survival, as indicated in the Mosley and Chen framework. These affect the health of the child before and after delivery and in some cases affect the mother's caretaking behavior towards the child (Mahy, 2003). These incorporate the age of the mother, sex of the child, Birth order, Breastfeeding, marital status and ethnicity.

Age of the mother

In studies conducted in different parts of the world, it has been uncovered that birth to women younger than the age of 18 and older than the age of 35 have a very high risk of under-five child mortality during first and older births. It is believed that a young mother is not biologically matured and a much older mother experiences complication, thus the possibility of pregnancy-related complications is high (Jolly *et al.*, 2000). Children born to younger mothers, aged 15-19 years old, and older mothers, aged 35-49 years, were more likely to die as compared to children born to in the age category 25-34 years (Ezra and Gurum, 2002). A study of determinants of infant and child mortality in Tanzania used the data from 1991/92 DHS and concluded that demographic and biological factors inclusive of teenage pregnancies had a more pronounced impact on infant and child mortality. On the other hand, determinants of interest such as socioeconomic determinants of child mortality were not as significant as elsewhere in Africa (Mturi and Curtis, 1995).

Sex of the child

According to (Kosher, 1993 and Raji, 2010) the sex of a child has been known to be a determinant and a cause of gender differential in mortality. Male mortality generally surpasses female mortality in the neonatal period. However, this differential is reversed in the post-neonatal period. Studies by Chen *et al*. (1981); Bhuiya and Streatfield (1991), and Arokiasamy (2002) supported that higher female than male mortality continued through childhood. The reversal of the sex differential of mortality, noticeably during childhood and persisting through adolescence, was postulated to be reflective of sex-biased health and nutrition related behavior favoring male children (Chen *et al*., 1981). Moreover, they conclude that son preference in parental care, intra family food distribution, feeding practices, and utilization of health services are some of the behavioral mechanisms by which sex-biased attitudes may have led to the observed mortality pattern. In East Asia, South Asia, Middle East and North Africa son preference is the most prevalent. Hesketh and Xing (2006) point out that son preference is manifest prenatally, through sex determination and sex-selective abortion, and post-natal through neglect and abandonment of female children, which leads to higher female mortality.

One would expect the mother's education to intervene in sex discrimination. However, the positive effect of mother's education on child survival is not analogous for boys and girls in Bangladesh (Bhuiya and Streatfield, 1991). They showed that for boys a change in mother's education from no schooling to 1-5 years of schooling resulted in a reduction in the predicted risk of 45 percent, while for girls it was only seven percent. Furthermore, a change from no schooling to six or more years reduced the risk of dying by 70 percent for boys and by only 32 percent for girls. However, Eswaran (2002) concluded that the empowerment of women, which increases the bargaining power of wives relative to their husbands, results in a decline in fertility and in the mortality rate of children under-five.

Even though most studies show discrimination bias towards girls, Pande (2003) recognized the sex composition of siblings as a factor in selective discriminatory practices that affect the health of surviving children. He identified that in rural India not all girls face the same level of discrimination; the first girl born after two or more boys may face less discrimination than a boy who has two or more older brothers. On the other hand, girls who were born into a family that already has two or more surviving daughters and no surviving sons are among the most likely to be severely stunted and are less likely to be immunized than are first daughters.

Birth order number

Birth order can be considered as one of the factors, and it has been the subject of a great deal of interest. Elliot (1992) came with reasons why birth order is likely related to mortality risk. To start with, the pool of parental resources, including both time and material resources, available to each child decreases as the sibling size increases. First and early born children will spend early years having exclusive or close to the exclusive attention of parents while later born will have to compete with siblings over resources right after they are born. Moreover, younger

siblings are likely to be introduced to developmentally inappropriate activities by older siblings. Finally, many siblings increase the likelihood of communicable diseases being introduced into the family, and younger siblings may be more susceptible to these diseases (Elliott, 1992).

Size of child at birth

Birth weight is an important determinant of perinatal, neonatal and post neonatal outcomes. Specifically, low birth weight, which are those infants born weighing less than 2 500g. Several studies have shown low birth weight is closely associated with under-five child mortality and affect child development and future risk of chronic disease. Low birth weight (with or without prematurity) decrease the odds of the children in the first month (Suparmi *et al*., 2016). Pojda and Kelly (2000) categorized birth weight for children who weigh less than 2 500g as children with a very small weight, average weight are children who weigh between 3000g to 3 500g and very large weight as children who weigh 4 500g or more at birth. The children weighing 2000 – 2 499g at birth are 4 times likely to die during their first 28 days of life than children who weigh 2 500 – 2 999g and 10 times more likely to die than children weighing 3000-3 499g. This because Low birth weight is associated with impaired immune function, poor cognitive development, and high risk of developing acute diarrhea of pneumonia. Bangladesh estimated that almost half of the under-five child mortality from pneumonia and diarrhea could be prevented if low birthweight were eliminated (Pojda and Kelley, 2000).

Breastfeeding

The relationship between breastfeeding and under-five child mortality has been fully documented in studies from many countries of the world. Although the magnitude of the estimates differs from study to study and across cultures, most research in developing countries attest to the significance of breastfeeding as a determinant of child survival. In general, the literature indicates that breastfed children are less vulnerable to the risk of under-five child death than are artificially fed children. In addition, even among breastfed children, both the duration and the intensity of breastfeeding are positively associated with child survival. Hence, entirely breastfed under-five children tend to have a lower risk of dying than partly breastfed ones (Akwara, 1994)

In several traditional societies, women breastfeed their children for extended periods. According to Akwara (1994) and Buchanan (1975) prolonged breastfeeding is said to have the effect of safeguarding the health of the child. Some clinical and epidemiological studies have shown that mother's milk has at least three properties, which help to protect the health of infants. First, it is nutritious breast milk appears to meet the nutritional requirements for the normal growth of an infant for at least six months (Wray, 1978). Consumed in sufficient quantities, it provides protection against malnutrition syndromes such as kwashiorkor and marasmus (Akwara, 1994 and Kleinman, 1984). (Akwara, 1994; Barros and Victora, 1990) have found that absence of breastfeeding is related to an excess in the incidence of diseases, such as diarrhea and gastrointestinal infections that are worsened by malnutrition.

Marital status

According to study, results of DHS data analysis from five countries in sub-Saharan Africa (Ethiopia, Kenya, Zimbabwe, Malawi, and Tanzania) shows children born from unmarried women are usually at disadvantage than the children born from married women. Particularly in unmarried mothers, the variation of effect was seen across different studied countries. In Ethiopia, the serious impact was seen while in Kenya it was much less. The study reports that the probability of children dying before the age of five for married women was 41.3% and for unmarried women was 75.5%. Furthermore, Clark & Hamplova (2013) suggests where single motherhood is less common, and they are stigmatized, they face more challenges in fulfilling the needs of children. Another study published by Clark & Hamplova (2013) in Sub-Saharan Africa stated that maternal care and economic status had more impact on child death among single mothers.

Ethnicity

Anderson *et al*. (2002) points out the overwhelming consensus on the differentials in the under-five child mortality rate among various population groups in South Africa. For example, the 1998 SADHS estimated that in 1996 the under-five child mortality for Africans was 47%, for Coloured people 19% and 11% for White people. Heaton and Amoateng (2007); Yach (1994) estimated the very similar pattern to this, with 51% for Africans,38% for Coloured people,8% for Asians and 7% for White people. Many researchers interpret racial difference in under-five child mortality as an expression of differential access to health care and socio-economic resources.

2.4.3 Household Environmental Factors

The environment in which the children lives has long been considered to have an impact on child survival status. Source of water and floor material are among the environmental factors, which highly affect under-five child mortality. As indicated by Mosley and Chen (1984) better water supply and the provision of sanitation facilities are important for child survival. Many studies have included household environmental variables as determinants of under-five child mortality, and have found a strong association (Kabir *et al*., 2001). However, such studies found out that environmental factors may not have an independent effect on childhood mortality but are influenced by some socioeconomic factors.

Main source of water

The risk of potentially fatal diarrheal diseases is expected to increase among households with no clean drinking water or with no safe sanitation. Mahmood (2002) has shown the relationship between access to clean water and sanitation to under-five mortality. According to Anderson *et al*. (2002) black and coloured populations showed a hierarchy of needs in which without clean water, sanitation matters little. In their analysis, they considered household social economic characteristic, access to and use of health care, environmental conditions and age of the mother. Mahmood (2002) also found that families living in households with piped water

connected in their houses have a significantly lower post neonatal mortality than those families that depend on wells for drinking water. However, the results did not show evidence of improved child survival in households that had flush toilets compared to those that did not have.

Main floor material

The floor material is strongly influenced by the socio-economic level of the household. This situation has become even worse where there is overcrowding, children have become more disposed to infectious diseases. Shehzad (2006) found that, in Pakistan, child illnesses such as diarrhea, acute respiratory infections and fever are affected by family size, flooring material, parental education and cleanliness of the area around the house. Anderson *et al.* (2002); Jacobs *et al*. (2009) and Shehzad (2006) has established the relationship between the type of dwelling and child mortality in their studies. This is expected, brick houses are likely to be more hygienic than those built from informal material, as is often the case in informal settlements in South Africa.

Chapter 3

Methodology

Introduction

This chapter discusses in detail the research methodology that has been adopted in this study and focus on the source of the data, and the research instrument used. The basic statistical and advanced analytical tool were employed to perform exploratory data analyses, logistic regression, survey logistic regression, generalized linear mixed models and generalized additive models, and present the strategy used in the analysis of the data. The variables included in this study were also presented.

3.1 Source of the data and Research instrument

The study uses secondary data from South Africa Demographic and Health Survey, 2016 (SADHS, 2016) source: https://www.dhsprogram.com/data/dataset_admin, which was conducted by Statistics South Africa (Stats SA) in partnership with the South African Medical Research Council (SAMRC) at the request of the National Department of Health (NDoH). *"SADHS 2016 data were collected using five questionnaires, the Household Questionnaire, the individual Woman's Questionnaire, the individual Man's Questionnaire, the Caregiver's Questionnaire, and the Biomarker Questionnaire. These questionnaires were adopted based on the DHS Program's standard of Demographic and Health Survey questionnaires to reflect the population and health issues relevant to South Africa. The input was solicited from various stakeholders representing government ministries and agencies, nongovernmental organizations, and international donors. After the preparation of the questionnaires in English, the questionnaires were translated into South Africa's 10 other official languages. In addition, information about the fieldworkers for the survey was collected through a self-administered Fieldworker Questionnaire."*

3.2 The sample design

"The sampling frame used for the SADHS 2016 is the Statistics South Africa Master Sample Frame (MSF), which was created using Census 2011 enumeration areas (EAs). In the MSF, EAs of manageable size were treated as primary sampling units (PSUs), whereas small neighboring EAs were pooled together to form new PSUs, and large EAs were split into conceptual PSUs. The frame contains information about the geographic type (urban, traditional, or farm) and the estimated number of residential dwelling units (DUs) in each PSU. The sampling convention used by Stats SA is DUs. One or more households may be located in any given DU; recent surveys have found 1.03 households per DU on average.

Administratively, South Africa is divided into nine provinces. The sample for the SADHS 2016 was designed to provide estimates of key indicators for the country as a whole, for urban and non-urban areas separately, and for each of the nine provinces in South Africa. To ensure that

the survey precision is comparable across provinces, PSUs were allocated by a power allocation rather than a proportional allocation. Each province was stratified into urban, farm, and traditional areas, yielding 26 sampling strata.

The SADHS 2016 followed a stratified two-stage sample design with a probability proportional to size sampling of PSUs at the first stage and systematic sampling of DUs at the second stage. The Census 2011 dwelling unit count was used as the PSU measure of size. A total of 750 PSUs was selected from the 26 sampling strata, yielding 468 selected PSUs in urban areas, 224 PSUs in traditional areas, and 58 PSUs in farm areas".

3.3 Exploratory Data Analysis

Exploratory data analysis (EDA) is a critical step in analyzing data from an experiment. The main purpose of EDA is to help understand the data into detail before the modelling and inferences tasks. We use EDA to:

- Determining the relationship between the dependent variable and the explanatory variable
- Detection of mistakes and checking assumptions
- Preliminary selection of appropriate models

The 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.

3.4 Study variables

3.4.1 Dependent variable

The child is alive variable was the dependent variable. Since this is a dichotomous variable, it was treated as such. Therefore, 1 was coded for the survival of a child and zero for death.

3.4.2 Independent variable

The survey captured a vast range of variables. However, this study employed 17 variables. The selection of independent variables in this study was guided by the reviewed literature and by the theoretical foundation established from the reviewed literature. Below are the lists of predictor variables used in this study.

Table 3. 1: Definition of variables

No.	Variable	Definition	Coding
	Demographics		
1.	Sex	Sex of child	Male (1) Female (2)
2.	Mother's age	Age of mother	< 20 years (1), 20-35 years (2), >35 years (3)
3.	Birth order	Birth order number	First birth (1) 2-3 births (2) More than 3 births (3)
4.	Birth size	Size of child at birth	Very small (1) Average (2) Very large (3)
5	Breastfeeding	Breastfeeding	No (0) Yes (1)
6.	Age of the head	Age of the household head	Less than 30 years (1) 30 – 39 years (2) More than 40 years (3)
7.	Marital status	Marital Status	Never Married (0) Married (1) Living with partner (2)
8.	Ethnicity	Ethnicity	Black African (1) Coloured (2) Others (3)
	Socio-Economics		
9.	Education Level	Highest level of Education	No education (0) Primary (1) Secondary (2) Higher (3)
10.	Employment	Mother currently working	No (0) Yes (1)
11.	Wealth index	Wealth index	Poor (1) Middle (2) Rich (3)
12.	Number of children under 5	Number of children 5 years and under in household	Less than 2 children (1) 2 or more children (2)
13.	Children ever born	Total children ever born	Less than 2 children (1) 2 or more children (2)
14.	Residential area	Place of residence	Urban (1) Rural (2)

	Household Environment		
15.	Water source	Main source of water	Safe water (1), Not safe water (2)
16.	Main floor	Main floor material	Finished (1) Unfinished (2)
17.	Province	Province	(1) Western cape (2) Eastern cape (3) Northern cape (4) Free state (5) Kwazulu-Natal (6) North west (7) Gauteng (8) Mpumalanga (9) Limpopo

3.5 Preliminary data analysis

The purpose of the present chapter is to summarize, and present results of the descriptive statistics used to describe the variables that are affecting under-five child mortality in South Africa. To perform this, we need to explore the demographics, socio-economics, fertility and household environment variables further.

These variables are categorical variables; now will look at frequency tables to see how much influence these variables have in under-five child mortality.

The distribution of the study population by demographic characteristics is presented in Table 3.2. The results from the dataset showed that male and female children were almost of the same proportion 51.6% and 48.6% respectively. Table 3.2: further showed that most children were from mothers aged 20-35 (48.4%) followed by mothers aged 35 and above that accounted for 34.3 %. The results indicated that 17.3% of children were from mothers aged 20 years and less. With respect to birth order, almost half of the children 49.5% were of birth order 2-3 births. Slightly more than one-third of the children 36.0% were of first birth then 14.5% were of birth order of more than three births. Considering self-rated child's size at birth by mothers, more than half of the respondents 58.2% rated their children as average while only 25.3% and 16.5% described their children as very large or very small. The distribution revealed that most children 73.2% did not receive breastfeeding from their parents while 26.8% did receive breastfeeding. More than half 58.6% of the heads of the households were above 40 years of age. This followed by household heads of age between 30-39 years who accounted for 27.1% and then followed by household heads age less than 30 years that accounted for 14.3%.

Results further showed that most mothers were never married as 54.3% of children were children of single mothers, followed by children of married women who accounted for 25.4%. Then 20.3% were children of mothers who live with their partners. Considering ethnicity affiliation, the results showed that 89.4% were children of Black African, which are the majority ethnic group in the country while 8.5% were children of Coloured, and others 2.2% were children of Whites and Indian/Asian combined.

Table 3. 2: Percentage distribution of demographic characteristics

Characteristics	Frequency	Percentage
Sex of child		
Male	1832	51.6
Female	1716	48.4
Mother's age		
<20 years	615	17.3
20-35 years	1717	48.4
>35 years	1216	34.3
Birth order		
First birth	1278	36.0
2-3 births	1755	49.5
More than 3 births	515	14.5
Birth size		
Very small	586	16.5
Average	2065	58.2
Very large	897	25.3
Breastfeeding		
Yes	951	26.8
No	2597	73.2
Age of the household head		
Less than 30 years	507	14.3
30 – 39 years	962	27.1
More than 40 years	2079	58.6
Marital status		
Never Married	1926	54.3
Married	901	25.4
Living with partner	721	20.3
Ethnicity		
Black African	3171	89.4
Coloured	300	8.5
Others	77	2.2

Table 3.3: displays the distribution of socio-economic characteristics. The results revealed the highest percentage distribution of children 78.9% belonged to mothers with secondary education. 9.7% and 1.5% of children are from mothers with primary education and no schooling respectively. Only 9.9% of children had mothers with higher education. With respect to employment, more than two third of children 70.3% were of mothers who were not working while 29.7% of children were children of mother who are working. With respect to wealth index, about 49.0% of the children were of poor mothers, and then 23.0% and 27.9% were of middle and rich mothers respectively. Considering the number of children 5 years and under in household, results revealed that 88.4% number of children 5 years and under in household were from households of less than two children while 11.6% were from households of two or more children. The distribution by total children ever born showed that most children 62.6% were from households of less than two children while 37.4% were from households of two or more children. Most children are from mothers who live in urban areas 52.5% and the rest 47.5% are from mothers who live in rural areas.

Table 3. 3: Percentage distribution of socio-economic characteristics

Characteristics	Frequency	Percentage
Education Level		
No education	53	1.5
Primary	344	9.7
Secondary	2800	78.9
Higher	351	9.9
Employment		
Yes	1052	29.7
No	2496	70.3
Wealth index		
Poor	1740	49.0
Middle	817	23.0
Rich	991	27.9
Number of children 5 years and under in household		
Less than 2 children	3137	88.4
2 or more children	411	11.6
Total children ever born		
Less than 2 children	2221	62.6
2 or more children	1327	37.4
Residential area		
Urban	1863	52.5
Rural	1685	47.5

Table 3.4: demonstrates the distribution by house environment characteristics. Considering the type of dwelling, the results showed that most children 90.4% were children from mothers who live in a finished dwelling floor material while 9.6% lived in an unfinished floor material. With respect to the household source of water, more than two-thirds 68.5% households had access to safe water while 31.5% did not have access to safe water. Percentage distribution of the province showed that most of the children 15.6% were children from mothers residing in KwaZulu-Natal and 14.1% were children from mothers residing in Mpumalanga. Followed by children of mothers residing in Limpopo 13.2%, Eastern Cape 12.7%, North West 11.1%, Gauteng 10.4%, Free State 9.0%, Northern Cape 8.1% and Western Cape 5.8%, respectively.

Table 3. 4: Percentage distribution of household environment characteristics

Characteristics	Frequency	Percentage
Main floor material		
Finished	3208	90.4
Unfinished	340	9.6
Water source		
Safe water	2430	68.5
Not safe water	1118	31.5
Province		
Western cape	206	5.8
Eastern cape	450	12.7
Northern cape	286	8.1
Free state	318	9.0
Kwazulu-Natal	555	15.6
North west	395	11.1
Gauteng	370	10.4
Mpumalanga	501	14.1
Limpopo	467	13.2

3.6 Chi-Square test of association

The aim of this section is to determine if there is a significant association between two categorical variables, the response and predictor variable using cross-tabulation procedure.

In Table 3.5 Table 3.6 and Table 3.7, we infer that the variables with p-value less than 5% level of significance were significantly associated with the response variable. The proportion of children in each category of the covariates and the results of the chi-square tests of association are presented in Table 3.5, Table 3.6 and Table 3.7 respectively.

Results showed that the proportion of children dying is higher for respondents with 2-3 births in birth order 49.5% as compared to the respondents with first birth and more than three births. Considering the birth size, the proportion of dying was higher for children born with average size 58.2 % than the other sizes. With respect to breastfeeding, the child from a mother who does not breastfeed had a higher proportion of dying 73.2% than children who received breastfeeding. The proportion of children dying was higher 54.3% for mothers that were never married than the child born by partner living together. Table 3.5 further showed that the proportion of children dying was higher 89.4% in the Black African population than the child born in other population groups was. Table 3.6: With respect to the wealth index, children born by

poor families had a higher proportion of dying 49% than the children born by middle and rich families. Considering the number of children 5 years and under in household, households with less than two children had a higher proportion of dying children 88.4% than the households of two or more children. With respect to the total number of children ever born, the results further showed that the proportion of dying children was more from households of less than two children 62.6% than the households of two or more children. Table 3.7: The proportion of children dying was higher 68.5% for children born drinking safe water than those who were not. It could be observed that KwaZulu-Natal had a higher proportion of dying children 15.6% than that of other provinces.

In summary we can say that Table 3.5, Table 3.6 and Table 3.7: shows that birth order, birth size, breastfeeding, ethnicity, marital Status, wealth index, number of children 5 years and under in household, total children ever born, water source and the province are significantly associated with child survival status.

Table 3. 5: Bivariate analysis of associations with under-five mortality demographic characteristics.

Demographic Characteristics					
Covariate	**Sample Size**	**DF**	**Proportion**	**Chi-Square**	**P-Value**
Sex of child	3548	1		2.663	0.103
Male			0.516		
Female			0.484		
Mother's age	3548	2		0.388	0.824
<20 years			0.173		
20-35 years			0.484		
>35 years			0.343		
Birth order	3548	2		12.884	0.002
First birth			0.36		
2-3 births			0.495		
More than 3 Births			0.145		
Birth size	3548	2		40.894	<0.001
Very small			0.165		
Average			0.582		
Very large			0.253		
Age of the household head	3548	2		0.555	0.758
Less than 30 years			0.143		
30 – 39 years			0.271		
More than 40 years			0.586		
Breastfeeding	3548	1		19.318	<0.001
Yes			0.268		
No			0.732		
Marital status	3548	2		11.049	0.004
Never Married			0.543		
Married			0.254		
Living with partner			0.203		
Ethnicity	3548	2		12.431	0.002
Black African			0.894		
Coloured			0.085		
Others			0.022		

Table 3. 6: Bivariate analysis of associations with under-five mortality socio-economic characteristics

Socio-Economic Characteristics					
Covariate	**Sample Size**	**DF**	**Proportion**	**Chi-Square**	**P-Value**
Education Level	3548	3		7.266	0.064
No education			0.015		
Primary			0.097		
Secondary			0.789		
Higher			0.099		
Employment	3548	1		2.379	0.123
Yes			0.297		
No			0.703		
Wealth index	3548	2		16.660	<0.001
Poor			0.49		
Middle			0.23		
Rich			0.279		
Number of children 5 years and under in household	3548	1		6.984	0.008
			0.884		
Less than 2 children			0.116		
2 or more children					
Total children ever born	3548	1		11.166	0.001
Less than 2 children			0.626		
2 or more children			0.374		
Residential area	3548	1		0.737	0.391
Urban			0.525		
Rural			0.475		

Table 3. 7: Bivariate analysis of associations with under-five mortality household environment characteristics

Household Environment Characteristics					
Covariate	**Sample Size**	**DF**	**Proportion**	**Chi-Square**	**P-Value**
Water source	3548	1		10.878	0.001
Safe water			0.685		
Not safe water			0.315		
Main floor material	3548	1		2.278	0.131
Finished			0.904		
Unfinished			0.096		
Province	3548	8		16.013	0.042
Western cape			0.058		
Eastern cape			0.127		
Northern cape			0.081		
Free state			0.009		
Kwazulu-Natal			0.156		
North west			0.111		
Gauteng			0.104		
Mpumalanga			0.141		
Limpopo			0.132		

3.7 Generalized Linear Model

Introduction

Generalized Linear Model (GLM) are some of the most widely used statistical techniques to analyze data sets. Basically, the GLM can be used to test almost any hypothesis about the response variable or independent variables (Miller and Haden, 2006).

In Chapter 1, we stated that the main objective of this study is to identify factors associated with the under-five mortality in South Africa. Now, this chapter will focus on modeling the relationship between the response variable and the predictor variables using logistic regression. The response variable is a binary outcome, which is assumed to follow the Bernoulli distribution. Bernoulli distribution is a member of the exponential family. To make valid statistical inference all the covariates that affect the child survival are assumed to have fixed effects. In the following section, the reviews for the theory of Generalized Linear Models is presented.

3.7.1 Review of Generalized Linear Models

Generalized Linear Models (GLM) are a flexible class of non-linear models for non-normally distributed response data. GLMs includes normal linear models as special case, but also cater for other error distributions, in particular, error distributions catering for discrete data such as the Poisson and binomial distributions. GLMs originated from a variety of different analysis problems including dilution assay to determine infective organism concentration, probit analysis in toxicology experiments and log-linear models for cross-tabulation (McCullagh &Nelder, 1989). The GLMs are used to accommodate non-normal responses and provide a unified approach to modeling all types of response variables (McCullagh and Nelder, 1989; Olsson, 2002; Dobson and Barnett, 2008). Basically, the GLMs can be described as a unified mathematical way of describing the relationships between a response variable and a set of covariates. Generalized linear models are an extension of the linear models that are given by.

$$y = X\beta + e \tag{3.1}$$

where, X is the design matrix of covariates, β is the vector of coefficients and $\boldsymbol{\epsilon}$ is the vector of error terms. Let $\eta = X\beta$, here η is the linear predictor part of the model. Since a generalized linear model extends the general linear models by relaxing the assumption that response 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(μ), so the model becomes

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

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

3.7.2 Exponential family of distribution

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

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

where $a(\emptyset)$ and $b(\theta_i)$ are known functions and $c(y_i, \emptyset)$ is some function of y_i and $\emptyset$. The parameter θ_i is called the canonical parameter,$\emptyset$ is the dispersion parameter. The mean, $\mu = E(y) = b'(\theta)$, and the variance, $var(y) = \emptyset b"(\theta)$, can be obtained by differentiation in Appendix C.

3.7.3 The structure of Generalized Linear Models

Generalized Linear Model consists of three components namely, random component, link function and systematic component. The random component refers to the probability distribution of the response variable Y given the values of the explanatory variables in the model. 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 observation is dichotomous 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 $x_1, x_2, x_3, \dots. , x_p$ that leads to the linear predictor η given by $\eta = \alpha + \sum_{j=1}^{n} x_j\beta_j$. The following section describes the concept of logistic regression.

3.8 Logistic regression

The Logistic Regression technique is frequently used for the analysis of data collected retrospectively. It commonly used statistical modeling technique that describes the relationship of several covariates to a binary response variable. The main purpose 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 binary response variable and one or more explanatory variables (Keinbaum *et al.*, 2002). The response variable is often taking two or more possible values when an individual is to be classified into two or more groups.

Linear logistic regression technique fits the model for binary or ordinal response data using the method of maximum likelihood and this model has been in use of statistical analyses for many years (Harrell, 2015). In this study, we focus on predicting a binary response using multiple predictors.

3.8.1 Model

Suppose the explanatory variables of interest $X = (X_1; \ldots\ldots; X_p)$ are p predictors for the i-th individual. Let the probability that the outcome is present be denoted by $P(Y_i = 1) = \pi_i$ for the i-th individual and let the outcome of being absent be denoted by $P(Y_i = 0) = 1 - \pi_i$. The ratio of these two events is defined as the odds. Logistic regression does not make any assumption of linear regression and general linear model that are based on ordinary least squares algorithms. The method is based on the log transformation of the odds and is given by the

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_p \tag{3.4}$$

Eq. (3.4) can also be written like this

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_p)} \tag{3.5}$$

which is the probability of the event occurring, and the probability of the non-occurrence 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 $\beta_0, \beta_1, \ldots, \beta_p$ that best fit the data relating child survival to number of covariates using SADHS 2016. The logistic regression enables researchers to overcome many of the linear regression assumptions that are too restrictive. The model holds the following assumptions.

Firstly, the linear relationship between dependent and independent variables is not assumed. However, a linear relationship between log(odds) and the independent variables is assumed. Secondly, the dependent variable does not need to be normally distributed. In addition, normally distributed error terms are not assumed. Thirdly, homoscedasticity is not needed. However, logistic regression does not need variances to be heteroscedastic for each level of the independent variables. Lastly, response variable is required to be binary and the observation should be independent of each other. We now look at how the parameters can be estimated using the maximum likelihood

3.8.2 Parameter estimation

To obtain the parameter estimates set the first derivative of log-likelihood with respect to each β equal to zero, so the maximum likelihood estimates for β can be obtained by setting each of the $K + 1$ equation obtained to zero and solving for each β_k. Each such solution, 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 every element on

the diagonal of the matrix is less than zero. One more 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 element of β, denoted by β_k leads to the variance covariance matrix (Czepiel, 2002).

The goal of logistic regression is to estimate the $K+1$ unknown parameter β in Eq. (3.6)

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{k=0}^{K} x_{ik}\beta_k \qquad i = 1,2,\dots,N \tag{3.6}$$

Therefore, the values of maximum likelihood estimates are the values for β that maximize the likelihood function in Eq. (3.7)

$$(\beta|y) = \prod_{i=1}^{N} \frac{n_i!}{y_i!\,(n_i - y_i)!} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i} \tag{3.7}$$

After rearranging terms, the equation to be maximized can be written as:

$$= \prod_{i=1}^{N} \left(\frac{\pi_i}{1-\pi_i}\right)^{y_i} (1-\pi_i)^{n_i} \tag{3.8}$$

After taking exponential e on both sides Eq. (3.6),

$$\left(\frac{\pi_i}{1-\pi_i}\right) = e^{\sum_{k=0}^{k} x_{ik}\beta_k} \tag{3.9}$$

which, after solving for π_i becomes,

$$\pi_i = \left(\frac{e^{\sum_{k=0}^{K} x_{ik}\beta_k}}{1 + e^{\sum_{k=0}^{K} x_{ik}\beta_k}}\right) \tag{3.10}$$

Substituting Eq. (3.9) for the first term and Eq. (3.10) for the second term, Eq. (3.11) becomes

$$\prod_{i=1}^{N}\left(e^{\sum_{k=0}^{K}x_{ik}\beta_k}\right)^{y_i}\left(1-\frac{e^{\sum_{k=0}^{K}x_{ik}\beta_k}}{1+e^{\sum_{k=0}^{K}x_{ik}\beta_k}}\right)^{n_i} \tag{3.11}$$

By simplifying the first product and the second product Eq.(3.11), becomes,

$$\prod_{i=1}^{N}\left(e^{y_i\sum_{k=0}^{K}x_{ik}\beta_k}\right)\left(1+e^{\sum_{k=0}^{K}x_{ik}\beta_k}\right)^{-n_i} \tag{3.12}$$

Thus, taking the natural log of Eq. (3.12) yields the log likelihood function:

$$l(\beta)=\sum_{i=1}^{N}y_i\left(\sum_{k=0}^{K}x_{ik}\beta_k\right)-n_i.\log\left(1+e^{\sum_{k=0}^{K}x_{ik}\beta_k}\right) \tag{3.13}$$

To find the critical points of the log likelihood function, set the first derivate with respect to each β equal to zero. In differentiating Eq. (3.13), we note that

$$\frac{\partial}{\partial\beta_k}\sum_{k=0}^{K}x_{ik}\beta_k=x_{ik} \tag{3.14}$$

In differentiating the second half of Eq. (3.13), we take note of the general rule $\frac{\partial}{\partial x}logy=\frac{1}{y}\frac{\partial y}{\partial x}$. Thus, differentiating Eq. (3.13) with respect to each β_k,

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

$$=\sum_{i=1}^{N}y_ix_{ik}-n_i\ .\frac{1}{1+e^{\sum_{k=0}^{k}x_{ik}\beta_k}}.e^{\sum_{k=0}^{k}x_{ik}\beta_k}.\frac{\partial}{\partial\beta_k}\sum_{k}^{K}x_{ik}\beta_k$$

$$= \sum_{i=1}^{N} y_i x_{ik} - n_i \cdot \frac{1}{1 + e^{\sum_{k=0}^{K} x_{ik}\beta_k}} \cdot e^{\sum_{k=0}^{K} x_{ik}\beta_k} \cdot x_{ik}$$

$$= \sum_{i=1}^{N} y_i x_{ik} - n_i \pi_i x_{ik} \tag{3.15}$$

The maximum likelihood estimates for β can be found by setting each of the $K + 1$ equations in Eq. (3.15) equal to zero and solving for each β_k. The general form of the matrix of second partial derivatives is,

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

$$= \frac{\partial}{\partial \beta_{k'}} \sum_{i=1}^{N} -n_i x_{ik} \pi_i$$

$$= -\sum_{i=1}^{N} n_i x_{ik} \frac{\partial}{\partial \beta_{k'}} \left(\frac{e^{\sum_{k=0}^{K} x_{ik}\beta_k}}{1 + e^{\sum_{k=0}^{K} x_{ik}\beta_k}} \right) \tag{3.16}$$

By applying the two general rules. First, a rule for differentiating exponential functions. Second, the quotient rule for differentiating the quotient of two function. Thus, we can write Eq. (3.16) as:

$$-\sum_{i=1}^{N} n_i x_{ik} \pi_i (1 - \pi_i) x_{ik'} \tag{3.17}$$

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. In the next section, we look at how non-linear equations can be solved iteratively using Raphson method.

3.8.3 Newton-Raphson Method

Setting the $K + 1$ equations from the first derivative of log-likelihood equate these equations to zero, this results into a system of non-linear equations with each $K + 1$ unknown variable. The solution to these equations is the vector with elements,β_k. After verifying that the matrix of second partial derivatives is negative definite and that the solution is global maximum rather than 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 and Dlamini, 2016).

Generalizing Newton's method to a system of equations is not difficult. In our case, we need to solve Eq. (3.15) the first derivative of the log likelihood function. Since Eq. (3.15) is really a system of $K + 1$ equations, whose root we want to find simultaneously. Hence, it is convenient to use matrix notation to express each step of the Newton Raphson method. Thus, the first step of Newton Raphson can be expressed as:

$$\beta^{(1)} = \beta^{(0)} + [-l"(\beta^{(0)})]^{-1}.\, l'\left(\beta^{(0)}\right) \tag{3.18}$$

Using matrix multiplication, we can show that

$$l'(\beta) = X^T(y - u) \tag{3.19}$$

Therefore, we can verify that

$$l"(\beta) = -X^T W X \tag{3.20}$$

The system of equation to solve for β is given by the following

$$\beta^{(1)} = \beta^{(0)} + [X^T W X]^{-1}.\, X^T(y - u) \tag{3.21}$$

Applying Eq. (3.12) continuously until there is essentially no change between elements of β from one iterative to the next. At that point, the maximum likelihood is said to have converged.

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 because 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.

3.9 Model Selection and Diagnostics

In a regression model, the problem of finding good predictors and discarding irrelevant ones becomes increasingly hard as the number of possible covariates increases. The regression uses as many variables as possible to minimize the sampling error, frequently leading to overfitting. Penalizing models with many parameters, allowing a compromise between complexity and fit, can solve this. Different forms of information criteria have been developed to help choose between models with different sets of covariates. The two most common are the Akaike information criteria (AIC) (Akaike, 1974) and the Bayesian information criterion (BIC) (Schwarz, 1978).

3.9.1 Akaike's Information Criterion

The only way to assess a model fit is to use Information Criterion. This criterion quantifies how well the model has predicted the data. The Akaike's Information Criterion (AIC) is a very useful statistic for comparing the relative fit of different models. This statistic was proposed by Akaike (1974) and is defined as

$$AIC = -2(Log\text{-}likelihood) + 2k \tag{3.22}$$

where k is the number of estimated parameters included in the model. The AIC penalizes for the addition of parameters, and thus selects a model that fits well but has a minimum number of parameters (Akaike, 1974). A model with the lowest AIC is preferred. The method is particularly useful when comparing non-nested.

3.9.2 Schwarz Criterion

Schwarz Criterion (SC) is another important criterion for model selection that measures the trade-off between model fit and the complexity of the model (Stone,1979), which provides more parsimonious model than AIC does. SC is also called Bayesian Information Criterion (BIC) or Schwarz Information Criterion (SIC) because Schwarz (Schwarz, 1978) gave a Bayesian argument on it. This criterion is closely related to Akaike's Information Criterion. SC is given by

$$SC = -2Log\text{-}likelihood + klog(n) \tag{3.23}$$

where *k* is the number of independent parameters, and *n* is the sample size. SC produces more severe penalization on the likelihood for estimating more parameters (Dlamini, 2016; Allison, 2012). BIC tends to favor parsimonious models. The model achieving the lowest BIC value is preferable model. When doing a model selection, we narrow down the options before comparing models. This is done by building the regression model systematically 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. The procedure will continue until

all non-significant variables are discarded 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 can be excluded from the model each time the new covariate is added to 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 (Dlamini, 2016).

3.10 Model Checking

3.10.1 Goodness-of-fit Test

Goodness-of-fit or lack-of-fit tests are designed to determine formally the adequacy or inadequacy of the fitted logistic regression model. Therefore, after fitting the logistic model to data set, it is reasonable to determine how well the fitted values under the model compare with the observed. The Pearson and Deviance goodness-of-fit statistics used for assessing the goodness-of-fit of the model.

3.10.2 Deviance

Nelder and Wedderburn (1972) first proposed the deviance as a measure of goodness-of-fit. The deviance is used to assess the fit of the model. It compares the models that are nested and in order to define the deviance we let $l(\hat{\mu}, \emptyset, y)$ be the log-likelihood of the reduced model at the maximum likelihood estimate, and let $l(y, \emptyset, y)$ be the log-likelihood estimate of the full or saturated model. The deviance is then given by

$$D = 2\big(l(y, \emptyset, y) - l(\hat{\mu}, \emptyset, y)\big) \tag{3.24}$$

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

$$D^* = \frac{Deviance}{\emptyset} \tag{3.25}$$

The Binomial and Poisson distribution has deviance and scaled deviance that are identical because $\emptyset = 1$ 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 the degrees of freedom $df_1 - df_2$.This will result 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.

3.10.3 Pearson Chi-Square Statistics

Karl Pearson (Pearson 1900) proposed chi-square goodness-of-fit test. This is another statistic for testing and comparing models and is defined as

$$\chi^2_{Pearson} = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu_i})^2}{\widehat{Var(\widehat{\mu_i})}} \tag{3.26}$$

where $\widehat{Var(\widehat{\mu_i})}$ 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.

3.11 Testing hypothesis

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. When the model is fitted, one can test for the significance of each parameter. The distribution $\hat{\beta}$ in Appendix C is $\beta \sim MVN(\beta, I^{-1})$ and can be used to test for the significance of $\hat{\beta}_j (j = 1,2, \dots, p)$ in the model. The Wald Chi-Square is given by

$$\chi^2{}_{wald} = \left(\frac{\widehat{\beta_J}}{\sqrt{\widehat{V_J}}} \right)^2 \tag{3.27}$$

where V_j ' s are the diagonal elements of I^{-1}. Chi-square distribution can be used with 1 degree of freedom and compare it with Wald Chi-square statistics. The hypothesis being tested is H_0: $\beta = 0$ against the alternative H_a: $\beta \neq 0$. If the Wald Chi-square statistics is greater than the table value of Chi-Square, H_0 is rejected that means the explanatory variables are significantly adding something in the model.

3.11.1 Odds ratio

Logistic regression quantifies the relationship between the binary dependent variable and the set of covariates using odds ratios (Kleinbaum; Kupper; Nizam and Muller, 2008). Odds ratios are used to compare the relative odds of the occurrence of the outcome of interest given exposure to the variable of interest divided by the probability that the event will not happen. Dlamini (2016) and Czepiel (2002) defined odds ratio as the ratio of the odds for those with risk

factor($X = 1$) to the odds for those without the risk factor($X = 0$). The log of the odds ratio is given by

$$log(\widehat{OR}) = log(OR(\, x = 1, x = 0))$$
$$= logit(x = 1) - logit(x = 0),$$
$$= (\hat{\beta}_0 + 1 \times \hat{\beta}_1) - (\hat{\beta}_0 + 0 \times \hat{\beta}_1),$$
$$= \hat{\beta}_1 \qquad (3.28)$$

The odds ratio then obtained by taking exponent of both sides of equation 3.28

$$OR = \exp(\hat{\beta}_1) \qquad (3.29)$$

The parameter β_1 associated with X represents the change in the log (odds) when X changes from $X = 0$ to $X = 1$. Then 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 therefore say X increases by unit, the odds of risk factor increase by $\exp(\hat{\beta}_1)$.

3.11.2 Confidence Interval for the Odds Ratio

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 (Dlamini, 2016; Allison, 2012). The confidence interval for slope and intercept is based on Wald tests. The $100\left(1 - \frac{\alpha}{2}\right)\%$ confidence interval for intercept given by

$$\hat{\beta}_j \pm Z_{1-\frac{\alpha}{2}}\sqrt{V_0} \qquad (3.30)$$

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

$$\hat{\beta}_j \pm Z_{1-\frac{\alpha}{2}}\sqrt{V_j} \qquad (3.31)$$

where $\sqrt{V_j}$ is the standard error of β_j. Here $Z_{1-\frac{\alpha}{2}}$ is the upper $100\left(1 - \frac{\alpha}{2}\right)\%$ 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 the corresponding $100\left(1 - \frac{\alpha}{2}\right)\%$

$$\exp\left(\hat{\beta}_j \pm Z_{1-\frac{\alpha}{2}}\sqrt{V_j}\right) \qquad (3.32)$$

This is the confidence interval for odds ratio associated with β_j where $j = 1,2,3, \ldots, p$

3.12 Logistic Regression Diagnostics

3.12.1 Pearson Residuals

The Pearson residual is the difference between observed and fitted values and divided by an estimate of the standard deviation of the observed value. This residual measure the relative deviations between observed and fitted values (Hosmer *et al.*, 2013). The Pearson residual is defined as follows

$$p_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i(n_i - \hat{\mu}_i)/n_i}} \tag{3.33}$$

where $\hat{\mu}_i$ is the fitted value and the denominator follows from the fact that $var(y_i) = n_i\pi_i(1-\pi_i)$. The result is called the Pearson residual since the square of p_i is the contribution of the i^{th} observation to Pearson's chi-squared statistic. With grouped data, the Pearson residuals are approximately normally distributed, but this is not the case with individual data. In both cases, observations with a Pearson residual exceeding two in absolute value may be worth a closer look.

3.12.2 Deviance Residuals

An alternative residual is based on the deviance or likelihood ratio chi-squared statistic. The deviance residual is defined as

$$d_i = \sqrt{2\left[y_i log\left(\frac{y_i}{\widehat{\mu_i}}\right) + (n_i - y_i) log\left(\frac{n_i - y_i}{n_i - \widehat{\mu}_i}\right)\right]}, \tag{3.34}$$

with the same sign as the raw residual $y_i - \hat{y}_i$. Squaring these residuals and summing over all observations yields the deviance statistic. Observations with deviance residual in excess of two may indicate lack of fit.

3.12.3 Influential observations

We now focus on detecting potential observations that 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 of the 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 *et al.*, 2013). The same methods have been developed for logistic regression.

3.12.4 Leverage of an observation

This is alternative measure where the observation with an extreme value on the predictor variable is known as a point with high leverage (Hosmer *et al.*, 2013). The leverage is defined as a measure of how far an independent variable deviate 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 *et al.*, 2013)

3.12.5 Predictive Accuracy

To check the predictive accuracy SAS procedure, PROC LOGISTIC produces other model statistics namely, Somers' D, Gamma, Tau-a and C. All these statistics range between zero and one. In all larger values correspond to a strong association between predicted and observed values. These measures of association are given by

$$Somers'\ D = \frac{C - D}{C + D + T},$$

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

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

$$C = 0.5(1 + Somers'D).$$

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 is said to 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 subject have the same classification on predictor and response variable. The total number of pairs is given by N. The value of C corresponds to the receiver operating characteristics (ROC) curve in the case of the binary response, which is defined below (Simundic, 2008).

3.12.6 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 pilot the ROC curve, one needs to plot sensitivity versus 1-specificity. Sensitivity measures the proportion of correctly classified positive outcome or event of interest (mortality), and the specificity measures the proportion of correctly classified event free outcome (no mortality). ROC provides a complete description of classification accuracy and

can be used as a graphical display of the prediction accuracy of the model (Dlamini, 2016; Vittinghoff *et al*., 2011; Simundic, 2008).

The shape of a ROC curve and the area under the curve (AUC) helps us estimate how high the discriminative power of a test is. The closer the curve is located to upper-left hand corner and the larger the area under the curve, the better the test is at discriminating between mortality and no mortality. The area under the curve (AUC) is between zero and one as shown in Figure 3.1. The ROC curve gives the measure of the model ability to classify between subjects, which have experienced the outcome versus those who did not.

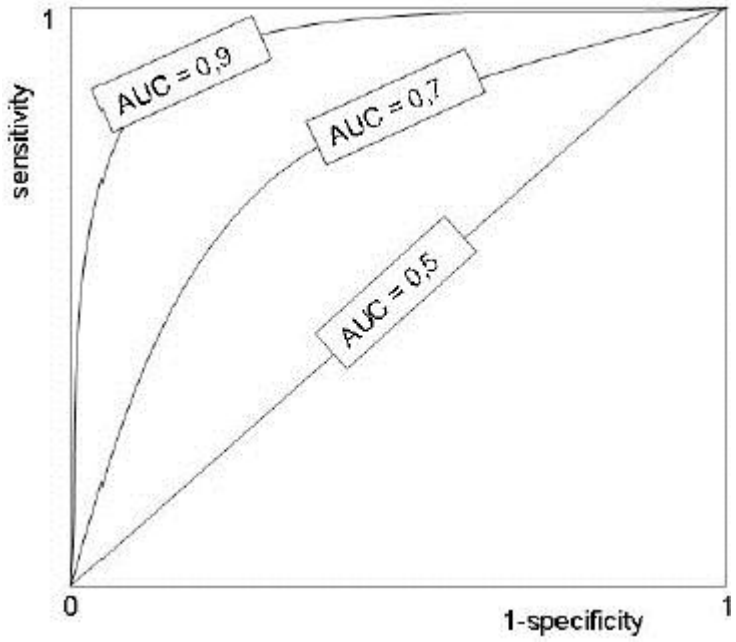

Figure 3. 1: Receiver Operating Characteristic (ROC) curve

Source: Simundic (2008).

Area under the curve (AUC) is a global measure of diagnostic accuracy. This area measures the prediction accuracy of the model and It tells us nothing about individual parameters, such as sensitivity and specificity (Simundic, 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 one 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 described in the Table 3.8 below

Table 3. 8: Relationship between the 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: Simundic.2008

3.13 Fitting the logistic Regression Model

The model was fitted using PROC LOGISTIC in SAS. Multivariate model was fitted to identify the association between the response variable and the covariates. The logistic regression was then fitted with all the variables that were identified as significant in the multivariate 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.

3.13.1 Interpretation of Logistic Regression coefficients, standard error and odds ratios

Table 3.9: shows the parameter estimates, Standard errors, p-values and odds ratios for the multivariate models. Under ethnicity, I have three categories, namely Black African, Coloured and Others. I have combined Whites and Indians to form the category called others. The variable is considered to have a significant effect if the p-values associated with the test of the hypothesis of the regression coefficient is less than 0.05.

The effect of size of child at birth (low birth weight) was positively associated with under-five mortality (p-value=0.0001). The corresponding odds ratio was 3.173 with (95% CI: 2.110; 4.771). The odds of death for a child born with very small weight was 3.173 times the odds of death for a child born with average weight. The effect of not breastfeeding was found to be positively associated (p-value=0.0002) with under-five child mortality. The corresponding odds ratio was 0.335 with (95% CI: 0.144; 0.782). The odds of death for a child from a mother who does not breastfeed was 0.335 times the odds of death for a child from a mother who breastfeed. The effect of birth order number above three was negatively associated (p-value=0.0312) with under-five child mortality. The corresponding odds ratio was 1.468 with (95% CI: 0.885; 2.132). The odds of death for a child whose birth order number is more than three was 1.468 times the odds of death for a child whose birth order number is between 2 - 3 births. The effect of ethnicity was negatively associated with under-five child mortality (p-value =0.0305). The corresponding odds ratio was 0.188 times with (95% CI: 0.041;0.855). The odds death of a child born from Coloured population group was estimated to be 0.188 times the odds of death for a child born from Black Africa population group.

Table 3. 9: Logistic Regression Model Coefficients, Standard errors, P-values and Odds ratios

Analysis of Maximum Likelihood Estimates and Odds Ratio estimates						
Demographic characteristics						
Parameter	**Estimate**	**Standard Error**	**P-Value**	**Odds ratio**	**95%Confidence interval**	
					Lower	**Upper**
Intercept	-3.9536	0.3685	<0.0001			
Mother's age (ref 20-35 years)						
<20 years	-0.1999	0.2641	0.4491	0.819	0.488	1.374
>35 years	-0.1740	0.2311	0.4513	0.840	0.534	1.322
Size of child at birth (ref. average)						
Very small	1.1546	0.2082	<0.0001	3.173	2.110	4.771
Very large	0.2289	0.2330	0.3260	1.257	0.796	1.985
Currently breastfeeding (ref. Yes)						
No	-1.0942	0.2912	0.0002	0.335	0.144	0.782
Birth order number (ref. 2 – 3 births)						
First birth	0.2635	0.2631	0.3165	1.302	0.777	2.180
More than 3 births	0.3840	0.2584	0.0312	1.468	0.885	2.436
Current marital status (ref. never married)						
Married	-0.3035	0.2637	0.2498	0.738	0.440	1.238
Living with partner	0.3200	0.2229	0.1512	1.377	0.890	2.132
Ethnicity (ref. Black African)						
Coloured	-1.6702	0.7721	0.0305	0.188	0.041	0.855
Others	-1.9225	0.8280	0.9809	0.146	0.029	0.741
Socio-Economic characteristics						
Wealth index (ref. Poor)						
Middle	-0.3288	0.2430	0.1760	0.720	0.447	1.159
Rich	-0.5520	0.2849	0.0527	0.576	0.329	1.006
Number of children 5 and under (ref.< 2 Children)						
2 or more children	-1.0923	0.4317	0.0114	0.335	0.144	0.782
Total children ever born (ref. less than 2 Children)						
2 or more children	0.5397	0.2747	0.0435	1.715	1.001	2.939
Household Environment characteristics						
Source of drinking water (ref. safe water)						
Not safe water	0.5133	0.2109	0.0149	1.671	1.105	2.526
Province (ref. KwaZulu-Natal)						
Eastern Cape	0.2932	0.3561	0.4104	1.341	0.667	2.694
Free State	0.7492	0.3897	0.0545	2.115	0.986	4.540
Gauteng	0.2728	0.4036	0.4990	1.314	0.596	2.897
Limpopo	-0.3357	0.3941	0.3943	0.715	0.330	1.548
Mpumalanga	0.7446	0.3218	0.0207	2.106	1.121	3.957
North West	0.2724	0.3620	0.4518	1.313	0.646	2.670
Northern Cape	0.5532	0.4619	0.2310	1.739	0.703	4.299
Western Cape	0.2741	0.5991	0.6473	1.315	0.407	4.256

The effect of number of children 5 year and under in household that is above two was found to be positively associated (p-value=0.0114) with under-five child mortality. The corresponding odds ratio was 0.335 with (95% CI: 0.144; 0.782). The odds of death for a child from a mother with two or more children alive was 0.335 times the odds of death for a child from a mother who has less than two children alive. The effect of total number of children ever born

that is above two was found to be positively associated (p-value=0.0435) with under-five child mortality. The corresponding odds ratio was 1.715 with (95% CI: 1.001; 3.939). The odds of death for a child from a mother who gave birth to two or more children alive was 1.715 times the odds of death for a child from a mother who has less than two children alive. The effect of not drinking safe water was found to be positively associated (p-value=0.0149) with under-five child mortality. The corresponding ratio was 1.671 with (95% CI: 1.105;2.526). The odds of death for child who doesn't not drink safe water was 1.671 times the odds death of a child who is drinking safe water. The effect of a province was found to be negatively associated (p-value=0.0207) with under-five child mortality. The correspondence ratio was 2.106 with (95% CI: 1.121;3.957). The odds of death for a child from a mother who lives in Mpumalanga was 2.106 times the odds death of a child from a mother who lives in KwaZulu-Natal provinces.

3.13.2 Model selection for logistic regression

Stepwise, forward and backward selection procedure were used to select significant variables associated with the response variable (child survival) in South Africa. All three procedures provided similar variable that were identified to be significant. Table 3.10 shows model statistics that was used to compare the two models.

Table 3. 10: Model fit statistics for logistic regression

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	1149.390	1063.219
SC	1155.564	1205.120
-2 Log- Likelihood	1147.390	1017.115

3.13.3 Hosmer and Lemshow Goodness-of-fit test

The Hosmer-Lemeshow goodness-of-fit statistic is another test used to assess the model fit. The test compares the predicted values against the actual values of the dependent variable. The method is similar to the chi-square goodness of fit. The Hosmer-Lemeshow test involves grouping the sample into groups based on the percentiles of estimated probability (Hosmer and Lemeshow, 2000).

To test for the goodness of fit of the model one can use the Hosmer Lemeshow test. The goodness-of-fit Chi-square statistics for Hosmer and Lemeshow is 2.5983 with 8 degrees of freedom and the corresponding p-value is 0.9570 as shown below in Table 3.11.

Table 3. 11: Hosmer and Lemeshow Goodness-of-Fit Test

Hosmer and Lemeshow Goodness-of-Fit Test	
Number of observations	3548
Number of groups	10
Hosmer-Lemeshow Chi-Square	2.5983
P - value	0.9570

This shows that there is no sufficient evidence to claim that the model does not fit the data adequately.

3.13.4 Prediction accuracy of the model

Model checking is important to check how much predicted probability associated with the response. The main objective is to have a model that maximizes the chance and sensitivity of identifying individuals that need justified intervention (Dlamini, 2016; Moeti, 2007). The table 3.5 shows the association of predicted probability and observed responses with the area under the curve being c = 0.761 and a concordant rate of 76.1 which tells us how good the model is for separating the 0's and 1's with a chosen model. Figure 3.2 displays the ROC curve of the fitted model and the area under the curve c = 0.761, which indicates that 76% probabilities were predicted correctly and that shows the model is a good prediction accuracy. The model assigned higher probability to child status (not alive) correctly. The measures Sumers' D, Gamma, and Tau-a are the summaries of the table of concordant and discordant pairs. These measures are likely to lie between zero and one where the large values indicate better predictive ability of the model, and these can be viewed as the measure of strength and direction of the relationship between pairs. The value Gamma is 0.523, which suggest that there is no perfect association. It is interpreted as 52% fewer errors are 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 response. The value for Somers'D is 0.523. This shows that not all pairs are concordant, and one may use to compare the model.

Table 3. 12: Association of Predicted Probabilities and Observed Responses

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	76.1	Somers' D	0.523
Percent Discordant	23.8	Gamma	0.523
Percent Tied	0.1	Tau-a	0.038
Pairs	460755	c	0.761

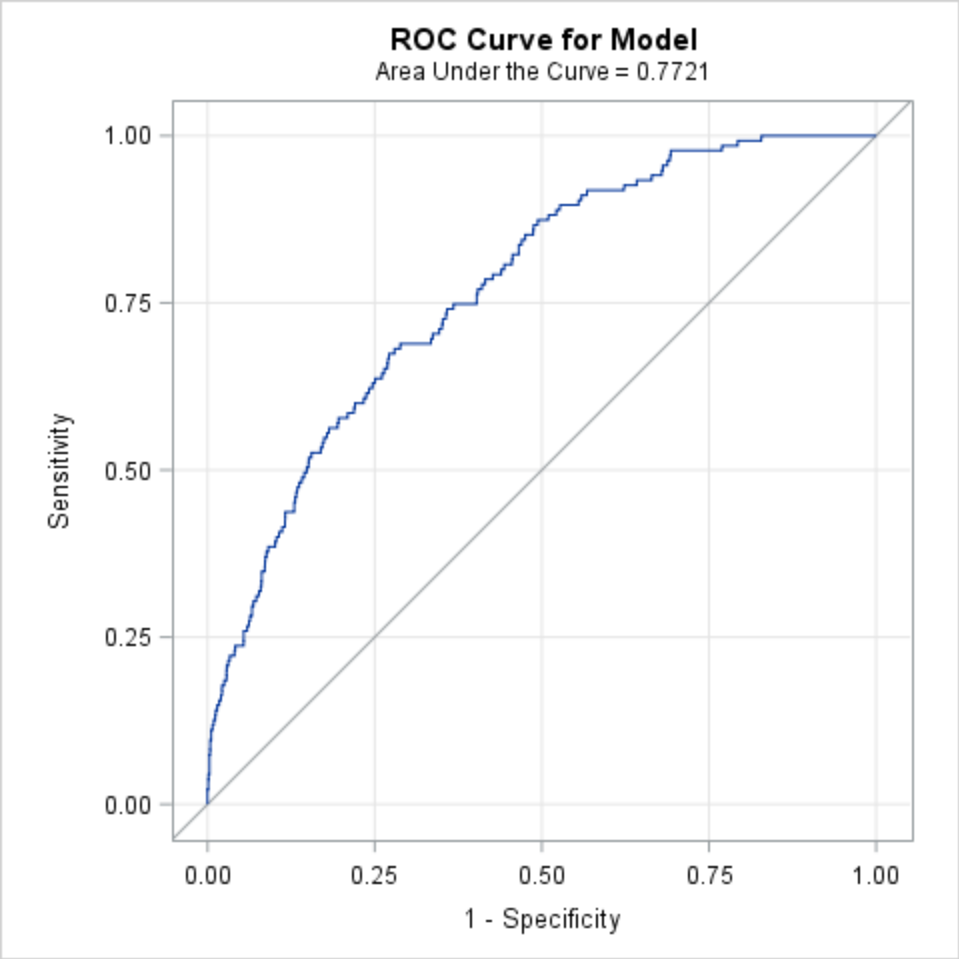

Figure 3. 2: Receiver Operating Characteristic (ROC) curve for model

3.13.5 Logistic regression diagnostic plots

We have discussed different techniques of diagnostics in section 3.12 above, now will focus on detecting potential observation that have significant impact on the model. The importance of diagnostic plots is to help us to detect if the was any error in data entry that may badly influence or skew the regression estimation. The residual and influence measures that help us to understand how the observations behave in the model, includes Standardized Pearson residuals, Standardized residuals, and Deviance residuals. Figure 3.3 shows influence diagnostics, which were produced by using, plots all 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 said to be influential observation.

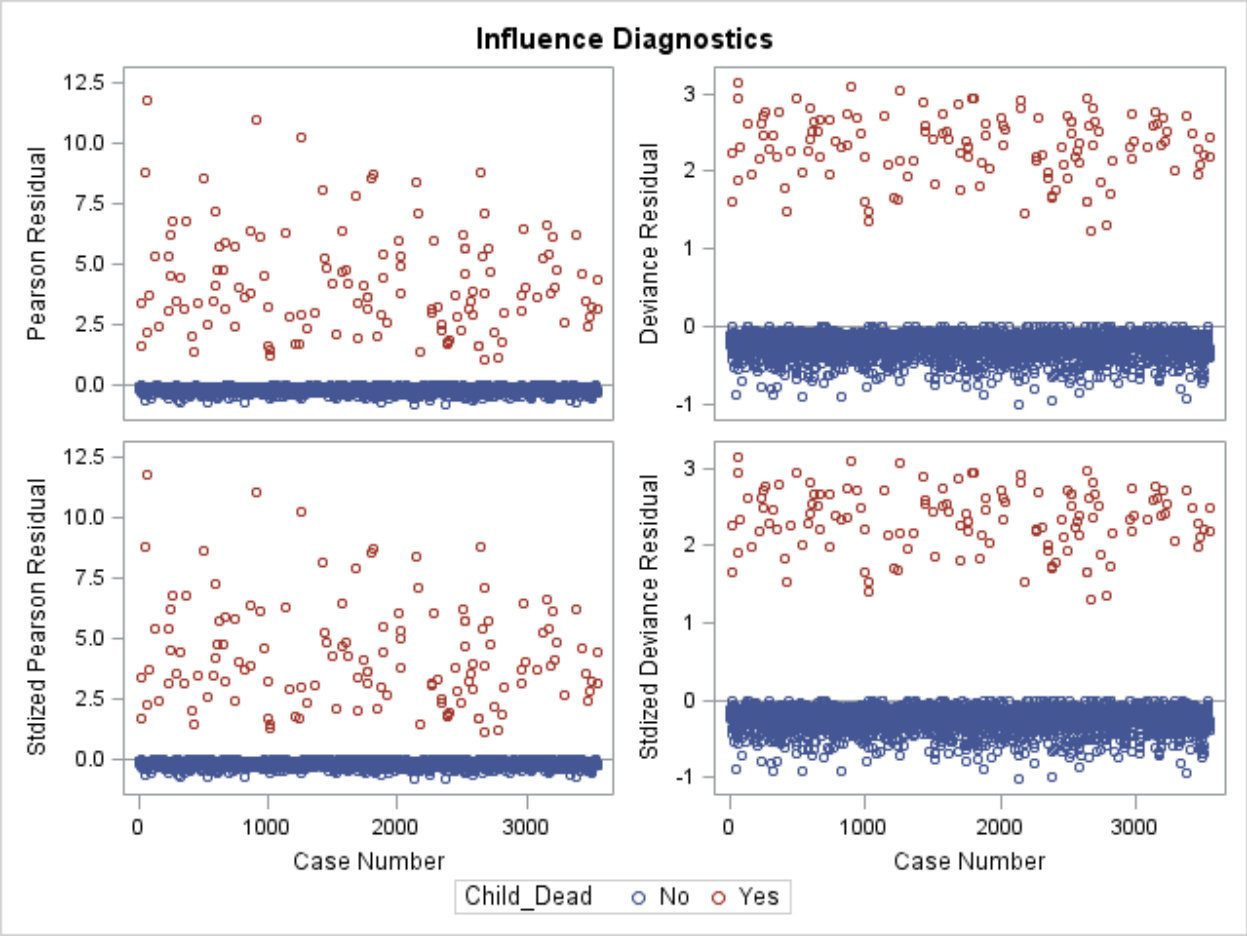

Figure 3. 3: Logistic Regression diagnostic plots

3.14 Limitations of the Logistic regression

In logistic regression, the goal is to find the best fitting model to describe the relationship between the dichotomous response variable and the set of independent variables. There is no assumption that is being made about the covariates, but covariates should not be highly correlated to one another since it may lead to problems with estimation. More covariates require larger sample sizes for estimation procedure. Logistic regression relies heavily on having an adequate number of samples for each combination of independent variables; small sample sizes can lead to widely inaccurate estimates of parameters. The other limitation that when there is non-linear relationship between log odds and covariates one may obtain invalid results. Furthermore, ordinary logistic regression is not an appropriate approach if the data come from sampling designs or complex nature of the survey design, which can lead to invalid statistical inference. In these cases, PROC SURVEYLOGISTIC is appropriate because it considers design (An, 2002). In the next section, we consider the method, which considers the survey design features.

3.15 Survey Logistic regression model

Survey logistic regression model has a similar theory as ordinary logistic regression. However, survey logistic regression accounts for the complexity of the survey design (Dlamini, 2016; Moeti, 2007). 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 can 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; Dlamini, 2016; Lemeshow and Hosmer, 2000). Parameter estimation methods are presented in the following section.

3.15.1 Parameter estimation

In a complex survey design, the assumption of independence does not hold. When clusters are drawn, they might introduce correlation among observation. This correlation might affect the standard error of the estimate. 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 Province). In the second stage, the household was sampled.

Thus, we specify the response variable as $y_{hijk}\big(h = 1,2, \dots\ , H_{kji}; i = 1,2, \dots\ , n_{kj}; j = 1,2, \dots\ , m_k; k = 1,2, \dots\ , K\big)$ which is 1 if the event occurred in h^{th} individual within i^{th} household within j^{th} primary sample units nested within k^{th} stratum and 0 otherwise. The total number of observations is given by $n = \sum_{k=1}^{k}\sum_{j=1}^{mk} n_{kj}$ and sampling design weight for the $kjih^{th}$are given in the dataset which are denoted by w_{kjih}. The weights are based on the sampling probability calculated at each stage. These design weights are 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^{th} individual within i^{th} household within j^{th} primary sample units nested within k^{th} stratum be $\pi_{kjih} = P(y_{hijk} = 1)$ and the probability that the event did not occur in h^{th} individual within i^{th} within j^{th} primary sample units nest within k^{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 to the likelihood (Dlamini, 2016; Lemeshow and Hosmer, 2000). The contributor 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

$$L(\beta;Y)=\prod_{k=1}^{K}\prod_{j=1}^{m_k}\prod_{i=1}^{n_{kj}}\prod_{h=1}^{H_{kji}}\pi_{kjih}^{w_{kjih}y_{kjih}}(1-\pi_{kjih})^{(1-w_{kjih}y_{kjih})} \quad (3.35)$$

The pseudo log-likelihood function is given by

$$l(\beta;Y)=\sum_{k=1}^{K}\sum_{j=1}^{m_k}\sum_{i=1}^{n_{kj}}\sum_{h=1}^{H_{kji}}\left\{w_{kjih}y_{kjih}log\left(\frac{\pi_{kjih}}{1-\pi_{kjih}}\right)-log\left(\frac{1}{1-\pi_{kjih}}\right)\right\} \quad (3.36)$$

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

$$X'W(y-\pi)=0 \quad (3.37)$$

where X is the $n\times(p+1)$ matrix of covariates values, W is as $n\times n$ diagonal matrix containing weights, y is the $n\times 1$ vector of observed outcome values and $\left[\pi_{1111},\dots,\pi_{km_kn_{kj}H_{kji}}\right]'$ is the $n\times 1$ vector of logistic probabilities.

The survey logistic regression model is given by

$$logit(\pi_{kjih})=log\left\{\frac{\pi_{hjih}}{1-\pi_{kjih}}\right\}=X'_{kjih}\beta \quad (3.38)$$

where 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 β is the vector of unknown model coefficients. In the following model, selection and checking procedure are discussed.

3.16 Survey logistic regression model selection and checking

3.16.1 Model selection

The SVY command in STATA provides a way to perform logistic regression with survey data. However, for survey data the estat gof, table group (10) command ordinarily used for estimating the Hosmer–Lemeshow goodness-of-fit test statistic associated with a fitted logistic regression model is not available after SVY estimation (Archer, K.J. and Lemeshow, S., 2006), and the Receiver Operating Characteristic (ROC) curve after fitting SVY prefix is also not appropriate. To address the goodness-of-fit problem, a Stata ado-command, svylogitgof, for estimating the F-adjusted mean residual test after SVY: logit or SVY: logistic estimation has been developed (Archer and Lemeshow, 2006). The same model fitted in Section 3.13 is fitted using SVY estimation.

3.16.2 Testing hypothesis about $\boldsymbol{\beta}$

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, 2007). The estimate of the covariance matrix of the estimator of coefficients is given by

$$\widehat{Var(\hat{\beta})} = (X'DX)^{-1}S(X'DX)^{-1} \tag{3.39}$$

where,$D = WV$ is the $n \times n$ diagonal matrix with general elements

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

The matrix S is a pooled within stratum estimator of the covariance matrix in the left side of equation (3.31). Lest us denote the general element of the vector of the score equation as

$$Z'_{kjih} = w_{kjih}\pi_{kjih}(1-\pi_{kjih})$$

Thus,

$$Z_{kj} = \sum_{i=1}^{n_{kj}} Z_{kjih} \tag{3.40}$$

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

$$S_k = \frac{m_k}{M_k} \sum_{j=1}^{m_k} (z_{kj} - \bar{z}_k)(z_{kj} - \bar{z}_k)' \tag{3.41}$$

The pooled estimator $S = \frac{m_k}{M_k} \sum_{k-1}^{K} (1 - f_k) S_k$. $(1 - f_k)$ is the finite population correlation 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 known one can assume that M_k is large enough so that f_k approaches zero, thus infinite population correction factor will be 1 (Dlamini, 2016; Lemeshow and Hosmer, 2000). The Wald statistics for testing all coefficients in the fitted model are equal to zero is by

$$Wald = \hat{\beta}' \left[\widehat{var(\hat{\beta})}_{p \times p} \right]^{-1} \hat{\beta} \tag{3.42}$$

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

$$p - value = P\left(\chi^2(p)\right) \geqslant wald)$$

Variances of the survey logistic regression parameters and odds ratios are computed using a Taylor series linearization (Siller and Tompkins, 2006). SVY code uses a Taylor series linearization approximation and incorporates the sample design information, including stratification, clustering, and unequal weighting. This also computes variances within each stratum and then pools the variance estimates 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

$$\hat{\beta}_j \pm Z_{1-\frac{\alpha}{2}} \sqrt{V_j} \tag{3.43}$$

where α is the level of significant $Z_{1-\frac{\alpha}{2}}$ is $100\left((1 - \frac{\alpha}{2}\right)$ 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 is on logit scale. The SVY code uses both Taylor series linearization and maximum likelihood. Procedures such as Jackknife Repeated Replication (JRR) and Balanced Repeated Replication (BRR) can be used to estimate variance of

each parameter. This procedure is used in this study to construct logistic regression model that account for the complex nature of the survey design.

3.16.3 Model checking

SVY command in STATA does not produce Hosmer-Lemeshow statistic test. However, Akaike's Information Criterion (AIC) and Schwarz Criterion (SC) are used to compare the goodness of fit (GOF) of the two-nested model (Dlamini, 2016; Moet, 2007). The goodness-of-fit test applied in complex survey data is called the $F_{adjusted}$ mean residual is obtained in the following manner, after the usual logistic regression model is fitted, the residuals $\hat{r}_{ji} = y_{ji} - \hat{\pi}(x_{ji})$ can be obtained. The goodness-of-fit test is based on the residuals since large departures between observed and predicted values that indicates lack of fit (Archer and Lemeshow, 2006). If we use grouping strategy, the observations are sorted into deciles based on their estimated probabilities, and each decile of risk includes approximately equivalent total sampling weights (Archer; Lemeshow and Hosmer, 2007).The survey estimates of the mean residual by decile of risk $\widehat{M} = (\widehat{M}_1, \widehat{M}_2, \widehat{M}_3, \dots, \widehat{M}_{10})$ are obtained such that the $\widehat{M}_g = \sum_j \sum_i w\hat{r}_{ji} / \sum_j \sum_i w_{ji}$ $(g = 1, \dots, 10)$. Here w_{ji} represent the sampling weights associated with the ordered residuals grouped into decile of risk. The association estimated variance-covariance matrix $\hat{V}(\widehat{M})$ is obtained using linearization, which is based on a first order Taylor series approximation. Therefore, the goodness-of-fit test implemented in *svylogitgof* is of the form

$$F_{adjusted} = \frac{f-g+2}{f_g} \widehat{M}^t \hat{V}\left(\widehat{M}\right)^{-1} \widehat{M} \qquad (3.44)$$

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

3.17 Design Effects

The design effect is defined as the ratio of the variance of an estimate under the complex sample design to the variance of the same estimate that would apply with a simple random sample of the same size (Kish, 1965). The sample size and sampling design determine the precision of the parameter estimates. However, due to the practical constraints such as cost and manpower, the national survey would not adopt the simple random sampling (Dlamini, 2016; Shackman, 2001). The complex design would be adopted instead. The problem we face in complex sample design is that sampling errors for survey estimates cannot be easily computed using the formulae found in statistical texts (Shackman, 2001). The design effects of survey estimates can be used as tools for measuring sample efficiency and for survey planning. The

STATA code *estat effects, deff deft* calculates the design effects for regression coefficients. Design effect is given by

$$DEFF = \frac{variance\ under\ the\ compex\ design}{variance\ under\ the\ simple\ random\ sample} \tag{3.45}$$

The denominator of the equation is computed under the assumption that the design is simple random variable with no stratification, clustering and weighting. The variance can be computed under the assumption of simple random sampling. If we consider both sampling weights and sampling rates (population totals) for the analysis, then the sampling rates under the assumption of simple random sampling is given by

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

where n is the sample size and w estimates the population size. If the sum of the weights (population size) is less than the samples size, f_{srs} is set to zero.

The design effect (DEFF) ratio indicates whether the sampling variability of an estimate was increased or decreased by the design used. A design effect of less than one indicates that fewer cases would be needed to obtain the same measurement precision obtained with simple random sampling. A design effect ratio greater than one indicates that more cases would be needed to obtain the same sampling precision found with a simple random sample. When the design effect is greater than one, then the standard errors of estimates from commonly used statistical packages underestimate the true sampling variability (Lee, Forthofer & Lorimer, 1989).

This design effect is used to compare variance when the data were obtained from simple random sampling and the variance when the data were obtained under complex design. Hence, one can use DEFT which is simply the square root of DEFF. The DEFT can be used to reduce variability since DEFT is less variable than DEFF. The DEFT can also be used to estimate confidence interval directly (Dlamini, 2016; Shackman, 2001). 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. The model fitting follows in the next section.

3.18 Fitting the Survey Logistic Regression Model

The model was fitted using SVY command in STATA to estimate, parameter estimates, standard errors and odds ratio. A model similar to the one fitted in section (3.13.2) was fitted and interpreted.

3.18.1 Model checking

The *estat ic* command in STATA does not produce plots and Hosmer-Lemeshow statistics. Therefore, one may use the Akaike's Information Criterion (AIC) and Schwarz Criterion (SC) to check if the model is a good fit or not. The AIC of the full model smaller compared to the AIC of the reduced model and this indicates that the smaller AIC fits the data well. The table 3.13 indicates the model fit that explains the data better.

Table 3. 13: Model fit statistics for Survey logistic regression

Model	Obs	Log-likelihood(null)	Log-likelihood (model)	df	AIC	BIC
.	3.471	-5.49e+08	-4.88e+08	22	9.76+08	9.76e+08
.	3,471	-5.49e+08	-4.74e+08	19	9.89e+08	9.84e+08

3.18.2 Goodness-of-fit test

The *estat gof* command in STATA is fitted in survey logistic to model the relationship between the categorical outcome (child survival status) and the set of predictor variables. However, the estat gof, command ordinarily used for estimating the Hosmer–Lemeshow goodness-of-fit test statistic associated with a fitted logistic regression model is not available after SVY estimation (Archer and Lemeshow, 2006). Therefore, the F-adjusted mean residual goodness-of-fit test was applied, and this indicates that the model fits the data well.

Logistic model for child survival status, goodness-of-fit test

$$F_{adjusted} = 33.07$$
$$P - value = 0.0001$$

3.18.3 Interpretation of Survey Logistic Regression coefficients, standard error and odds ratios

Table 3.14: shows the estimated coefficients, standard errors, p-values and odds ratios for the multivariate models. The variables that were found to be significant if p-values which were less than 0.05.

The effect of the size of the child at birth (very small weight) was found to be positively associated with under-five mortality (p-value=0.001). The corresponding odds ratio was 3.7311 with (95% CI: 0.7599; 1.8735). The odds of death for very small birth size were estimated to be 3.7311 times the odds of death for average weight of a child. The effect of the size of a child at birth, very large weight was also found to be a negatively associated with under-five mortality (p-

value=0.001). The corresponding odds ratio was 3.1161 with (95% CI: 0.5468; 1.7264). The odds of very large birth size were estimated to be 3.1161 times the odds of death for average weight of a child. The effect of not breastfeeding was found to be positively associated (p-value=0.017) with under-five child mortality. The corresponding odds ratio was 2.6993 with (95% CI: 0.1790; 1.8070). The odds of death for a child from a mother who does not breastfeed was estimated to be 2.6993 times the odds of death for a child from a mother who breastfeeds. The effect of ethnicity was found to be negatively associated with under-five child mortality (p-value =0.001). The corresponding odds ratio was 15.8002 times with (95% CI: 1.1307;4.3893). The odds death of a child born from coloured population group was estimated to be 15.8002 times the odds of death for a child born from black African population. The effect of number of children 5 years and under in household that is above two was found to be positively associated (p-value=0.030) with under-five child mortality. The corresponding odds ratio was 1.0036 with (95% CI: 0.0981; 1.9090). The odds of death for a child from a mother with two or more children alive were estimated to be 1.0036 times the odds of death for a child from a mother who has less than two children alive. The effect of a total number of children ever born that is above two was found to be positively associated (p-value=0.043) with under-five child mortality. The corresponding odds ratio was 0.4977 with (95% CI: -1.3953; -0.0029). The odds of death for a child from a mother who gave birth to two or more children alive were estimated to be 0.4970 times the odds of death for a child from a mother who has less than two children alive. The effect of not drinking safe water was found to be positively associated (p-value=0.035) with under-five child mortality. The corresponding ratio was 0.8922 with (95% CI: -0.6355;0.4073). The odds of death for child who doesn't not drink safe water was 0.8922 times the odds death of a child who is drinking safe water.

Table 3. 14: Survey Logistic Regression Model Coefficients, Standard errors, P-values and Odds ratios

Analysis of Maximum Likelihood Estimates and Odds Ratio estimates						
Demographic characteristics						
Parameter	**Estimate**	**Standard Error**	**P-Value**	**Odds ratio**	**95%Confidence interval Lower**	**Upper**
Mother's age (ref <20 years)						
<20 years	-0.4522	0.3084	0.143	0.6362	0.3476	1.1645
>35 years	-0.1746	0.3004	0.561	0.8398	0.4661	1.5133
Size of child at birth (ref. Average)						
Very small	1.3167	0.2841	<0.001	3.7311	0.7599	1.8735
Very large	1.1366	0.3009	<0.001	3.1161	0.5468	1.7264
Currently breastfeeding (ref. Yes)						
No	0.9930	0.4153	0.017	2.6993	0.1790	1.8070
Birth order number (ref. First birth)						
2 – 3 births	0.3048	0.3426	0.374	1.3564	-0.3666	0.9762
More than 3 births	0.1555	0.5008	0.042	0.5850	-0.8260	1.1370
Current marital status (ref. never married)						
Married	0.2352	0.3382	0.487	1.2651	-0.4276	0.8979
Living with partner	-0.0616	0.2809	0.826	0.9403	-0.6122	0.4890
Ethnicity (ref. Black African)						
Coloured	2.7600	0.8313	0.001	15.8002	1.1307	4.3893
Others	0.0218	0.0422	0.605	1.0220	0.6612	1.2010
Socio-Economic characteristics						
Wealth index (ref. Poor)						
Middle	0.5610	0.3169	0.077	1.7525	-0.0600	1.1821
Rich	0.7754	0.3985	0.052	2.1716	-0.0060	1.5565
Number of children 5 and under (ref.< 2 Children)						
2 or more children	1.0036	0.4620	0.030	2.7280	0.0981	1.9090
Total children ever born (ref. less than 2 Children)						
2 or more children	-0.6991	0.3552	0.043	0.4970	-1.3953	-0.0029
Household Environment characteristics						
Source of drinking water (ref. safe water)						
Not safe water	-0.1141	0.2660	0.035	0.8922	-0.6355	0.4073
Province (ref. Western Cape)						
Eastern Cape	0.1317	0.6376	0.836	1.1408	-1.1179	1.3813
Free State	-0.2265	0.6579	0.731	0.7973	-1.5159	1.0630
Gauteng	0.1622	0.6698	0.809	1.1762	-1.1506	1.4751
KwaZulu-Natal	0.5147	0.6366	0.419	1.6730	-0.7331	1.7624
Limpopo	0.8666	0.6699	0.196	2.3788	-0.4465	2.1797
Mpumalanga	-0.3055	0.6317	0.009	0.7368	-1.5436	0.9327
North West	0.1605	0.6740	0.812	1.1762	-1.1606	1.4816
Northern Cape	-0.2481	0.6748	0.713	0.7803	-1.5706	1.0744

The effect of a province was found to be negatively associated (p-value=0.009) with under-five child mortality. The correspondence ratio was 0.7368 with (95% CI: -1.5436;0.9327). The odds of death for a child from a mother who lives in Mpumalanga was estimated to be 0.7368 times the odds death of a child from a mother who lives in the Western Cape.

3.18.4 Similarities of Logistics and Survey Logistics regression model

Tables 3.9 and 3.14: both contains regression coefficients, standard error, p-value, odds ratios and confidence interval for logistic and survey logistic regression respectively. Since the sample was not drawn using simple random sample, the parameter estimates for both models are not the same. One of the assumptions for logistic regression is that the observations are independent, but for complex design this assumption is violated thus a better model may be the one fitted using survey logistic regression since it accounts for the complexity of the design. The models fitted by both methods produce the areas under the curve, which are between 0.77 and 0.76. This suggests that both models had good prediction accuracy.

3.18.5 Interpretation of Design effects

Table 3.15: below shows the DEFF and DEFT, which is calculated as the squared root of DEFF for each estimated coefficient.

The effect of size of child at birth, average weight was found to be negatively associated with under-five mortality has DEFF value of 2.1479 and DEFT value of 1.4656. The standard error and confidence interval are 1.4656 times larger as they would be for simple random sampling. The effect of size of child at birth, very large weight was also found to be negatively associated with under-five mortality has DEFF value of 1.3184 and DEFT value of 1.1482. The standard error and confidence interval are 1.1482 times large as they would be for sample random sampling. The effect of not breastfeeding was found to be positively associated with under-five child mortality has DEFF value of 1.9302 and DEFT value of 1.3893. The standard error and confidence interval are 1.3893 times large as they would be for simple random sampling. The effect of ethnicity was found to be negatively associated with under-five child mortality has DEFF value of 0.3709 and DEFT value of 0.6090. The standard error and confidence interval are 0.6090 times large as they would be for sample random sampling. The effect of number of children 5 year and under in household that is above two was found to be negatively associated with under under-five child mortality has DEFF value of 1.1160 and DEFT value of 1.0564. The standard error and confidence interval are 1.0564 times large as they would be for simple random sampling. The effect of total number of children ever born that is above two was found to be negatively associated with under-five child mortality has DEFF value of 1.1872 and DEFT value of 1.0896. The standard error and confidence interval are 1.0896 times large as they would be for simple random sampling. The effect of a province was to be negatively associated with under-five child mortality has DEFF value of 1.6519 and DEFT value of 1.2853. The standard error and confidence interval are 1.2853 times large as they would be for simple random sampling.

Table 3. 15: Survey Logistic Regression Coefficients, Standard errors, P-values, Odds ratios and Design effects

Analysis of Maximum Likelihood Estimates and Odds Ratio estimates								
Demographic characteristics								
Parameter	**Estimate**	**Standard Error**	**P-Value**	**Odds ratio**	**95%Confidence interval**		**Design Effects**	
					Lower	**Upper**	**DEFF**	**DEFT**
Mother's age (ref 20-35 years)								
<20 years	-0.4522	0.3084	0.143	0.6362	0.3476	1.1645	1.1972	1.0942
>35 years	-0.1746	0.3004	0.561	0.8398	0.4661	1.5133	1.1116	1.0543
Size of child at birth (ref. Average)								
Very small	1.3167	0.2841	<0.001	3.7311	0.7599	1.8735	2.1479	1.4656
Very large	1.1366	0.3009	<0.001	3.1161	0.5468	1.7264	1.3184	1.1482
Currently breastfeeding (ref. Yes)								
No	0.9930	0.4153	0.017	2.6993	0.1790	1.8070	1.9302	1.3893
Birth order number (ref. First birth)								
2 – 3 births	0.3048	0.3426	0.374	1.3564	-0.3666	0.9762	1.2006	1.0957
More than 3 births	0.1555	0.5008	0.042	0.5850	-0.8260	1.1370	1.1981	1.0946
Current marital status (ref. never married)								
Married	0.2352	0.3382	0.487	1.2651	-0.4276	0.8979	1.3665	1.1690
Living with partner	-0.0616	0.2809	0.826	0.9403	-0.6122	0.4890	1.7347	1.3171
Ethnicity (ref. Black African)								
Coloured	2.7600	0.8313	0.001	15.8002	1.1307	4.3893	0.3709	0.6090
Others	0.0218	0.0422	0.605	1.0220	0.6612	1.2010	0.0026	0.0510
Socio-Economic characteristics								
Wealth index (ref. Poor)								
Middle	0.5610	0.3169	0.077	1.7525	-0.0600	1.1821	1.7359	1.3175
Rich	0.7754	0.3985	0.052	2.1716	-0.0060	1.5565	2.1099	1.4526
Number of children 5 and under (ref.< 2 Children)								
2 or more children	1.0036	0.4620	0.030	2.7280	0.0981	1.9090	1.1160	1.0564
Total children ever born (ref. less than 2 Children)								
2 or more children	-0.6991	0.3552	0.043	0.4970	-1.3953	-0.0029	1.1872	1.0896
Household Environment characteristics								
Source of drinking water (ref. safe water)								
Not safe water	-0.1141	0.2660	0.035	0.8922	-0.6355	0.4073	1.3401	1.1576
Province (ref. Western Cape)								
Eastern Cape	0.1317	0.6376	0.836	1.1408	-1.1179	1.3813	1.6448	1.2825
Free State	-0.2265	0.6579	0.731	0.7973	-1.5159	1.0630	1.3376	1.1565
Gauteng	0.1622	0.6698	0.809	1.1762	-1.1506	1.4751	2.1940	1.4812
KwaZulu-Natal	0.5147	0.6366	0.419	1.6730	-0.7331	1.7624	1.7047	1.3057
Limpopo	0.8666	0.6699	0.196	2.3788	-0.4465	2.1797	1.4587	1.2078
Mpumalanga	-0.3055	0.6317	0.009	0.7368	-1.5436	0.9327	1.6519	1.2853
North West	0.1605	0.6740	0.812	1.1762	-1.1606	1.4816	1.4340	1.1975
Northern Cape	-0.2481	0.6748	0.713	0.7803	-1.5706	1.0744	0.6153	0.7844

Generally, we observe that the most design effects values are above one and this suggest that the standard errors assuming simple random sampling are underestimates of the true standard errors. Therefore, the variance was underestimated while using logistic regression model compared to those computed while using complex design. This confirm that standard

errors are larger under survey logistic regression. This shows that there was an under estimation of variance while using logistic regression assuming that data was sampled using simple random sampling. Thus, using survey logistic regression model is good since it takes into consideration survey design features.

3.19 Limitations of Survey Logistic Regression

Even though survey logistic regression 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 the Goodness of fit of the model. The variable selection procedures are not available as a result one has to select variables manually which can be time consuming when there are many variables and possible errors may occur while choosing variables. The model is chosen based on the Akaike Information Criterion and Bayesian information criterion both of which introduce a penalty to the -2Log-likelihood of having many parameters. As they both have -2log-likelihood term in their formulation, they are used only in the case of ungrouped data (Hosmer and Lemeshow, 2000).

Chapter 4

Generalized Linear Mixed Models

Introduction

Generalized linear mixed models (GLMMs) are natural extensions of the generalized linear models (GLMs) discussed in section 3.7 that allow for additional components of variability due to unobservable effects. Typically, the unobserved effects are modelled by the inclusion of random effects in the generalized linear model (Song and Lee, 2006). This inclusion of random effects in the analysis results into Generalized linear mixed models. These models provide all advantages of a logistic regression such an information on a sample size, they can do one analysis with all random effects on it, and they accommodate the binary response variable. Moreover, the advantage of GLMMs is its ability to handle unbalanced data due to missing observations and ability to account for correlated data (Dlamini, 2016; Manning, 2007).

GLMMs are powerful since they combine features of both linear mixed models and generalized linear models, such as fixed effects and random effects. They can handle a wide range of response distributions and data with observations sampled in some group structure instead of completely independent (Dean and Nielsen, 2007). In the following section, the theory of linear mixed models is reviewed.

4.1 Review of Linear Mixed Models

The generalized linear model in section 3.7 is not appropriate for the inclusion of random effect, however, it is necessary to expand the model

$$y = X\beta + \epsilon \tag{4.1}$$

This becomes generalized linear mixed models, which include both the fixed, and random effect is expressed as follows:

$$y = X\beta + Zu + \epsilon \tag{4.2}$$

where y is a $N \times 1$ column vector, the outcome variable;

X is a $N \times p$ matrix of the p predictor variables;

β is a $p \times 1$ column vector for the fixed effects coefficients;

Z is the $N \times q$ design matrix for the q random effects (the random complement to the fixed X);

μ is a $q \times 1$ vector of the random effects (the random complement to the fixed β), and

ϵ is a $N \times 1$ column vector of the residuals, which have multivariate normal distribution with the mean vector 0 and variance covariance matrix R i.e $\epsilon \sim N_n(0, R)$**.** Given nature random effect

hypothesis, U is treat differently from β. Statistical linear mixed models state that observed data consist of two parts that is, random and fixed effects (Littell *et al*., 2000). We defined 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 the covariation between measures on the same unit (Kincaid, 2005 and Littell *et al*., 2000) Representing variance of the model as $V(y)$ shown in the equation (4.3) below is known as modelling covariance structure. It is modelled as function of relatively small number of parameters (Littell *et al*., 2000). The specification of the covariance structure for mixed model is done through G and R as

$$V(y) = ZGZ' + R \tag{4.3}$$

where ZGZ' represents the between subject portion of the covariance structure and 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 the mean 0 and variance-covariance matrix G. Random effect can be predicted and not estimated. The variance components are estimated instead.

The diagonal elements of the matrix G are the variance component for each random effect while off-diagonal elements are covariance that exists 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, G is a $k \times k$ for k random effect. Suppose $k = 3$ random present five different covariance structures in the Table 4.1 below and discuss them. Table 4.1: displays the list of simplest covariance structures that can be modelled in using PROC MIXED procedure.

Table 4. 1: List of simplest covariance structure

Structure	**Description**	**Number of parameters**	**i,jth element**
AR(1)	First Order Autoregressive	2	$\sigma_{ij} = \sigma^2\rho^{\lvert i-j\rvert}$
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	$\sigma_{ij} = \sigma_{\lvert i-j\rvert+1}$
VC	Variance Component	q	$\sigma_{ij} = \sigma_k^2 1(i=j)$

Variance Component (VC)

The variance component structure is the simplest, where the correlation of errors within a subject is presumed to be zero. This structure is the default setting in PROC Mixed but is not a reasonable choice for repeated measure designs.

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

Compound Symmetry (CS)

The only covariance structure that incorporates within-subject correlated errors is compound symmetry (CS). Here we see correlated errors between time points within subjects, and these correlations are assumed to be the same for each set of times, regardless of how distant in time the repeated measures are made.

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

First Order Autoregressive AR (1)

The autoregressive (Lag1) 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. As the measurements get further apart, they are less correlated (Kincaid, 2005; Littell *et al*., 2000). However, this structure is only applicable for evenly spaced time intervals for the repeated measure.

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

Toeplitz (TOEP)

The covariance structure 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 correlation does not necessarily have the same pattern. The AR (1) is basically a special case of Toeplitz itep (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 (UN) is the most complex since it is estimating unique correlations for each pair of time points. It is not uncommon to find out that you are not able to use this structure. SAS will return an error message indicating that there are too many parameters to estimate with the data.

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

The assumptions made for generalized linear models are retained in generalized linear mixed models. Hence, it is possible to have a variable that appears in both X and Z. In this case, the fixed effect is an average across all levels of random effects and the estimate is the amount of variance in the effect between levels. If X contains a single column of ones, then this leads to the random intercept model. If 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 (Dlamini, 2016; McCullagh and Nelder, 1989). Therefore, the linear mixed model can be viewed as a special case of the generalized linear mixed model.

4.2 Model Formulation

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

$$f(Y|\theta) = exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\emptyset)} - c(y_i, \emptyset\right\} \tag{4.4}$$

where $\emptyset$ is the scale parameter. Based on the model the conditional y related to $\emptyset_i$ is given by

$$E(y|\theta) = \frac{\partial b(\theta_i)}{\partial \theta_i}$$

The model with both random and fixed effects is given by

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

where $\eta_i = g(\theta_i), g$ is the link function and U_i is a vector of random effects. In this study, the child survival status is either zero (child not alive) or one (child alive). Hence we use the logistic regression where we consider $g(.)$ as the logit link with X_i and Z_i $(i = 1,2, \dots, n)$ being p-dimension and q dimension a vector of know covariates, while β is a p-dimension vector of unknown fixed effects regression coefficient.

4.3 Maximum Likelihood Estimation

To obtain the maximum likelihood estimates in GLMMs. The marginal likelihood is maximized which is obtained by integrating over the distribution of the q-dimensional random effects. The contribution of the i^{th} cluster to the likelihood is given by.

$$f_i(y_{ij}|\beta, G, \emptyset) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\gamma_i, \beta, \emptyset) f(\gamma_i|G)\, d\gamma_i \tag{4.6}$$

where $f(\gamma_i|G)$ is the distribution of the random effects.

Therefore, the complete likelihood function for β, G and $\emptyset$ is given by

$$L(\beta, G, \emptyset) = \prod_{i=1}^{m} f_i(y_{ij}|\beta, G, \emptyset)$$

$$= \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|\gamma_i, \beta, \emptyset) f(\gamma_i|G) d\gamma_i \quad (4.7)$$

In this case of normality assumptions, the method of maximum likelihood for the estimation of the fixed effects in the GLMM becomes the same as that for the Linear Mixed Model (LMM). However, for numerous cases of the GLMM, the likelihood function typically does not have a close form expression (Roberts, 2014 and Jiang, 2007). Usually, this is due to the likelihood involving high dimensional integrals that cannot be evaluated analytically. Hence, the approximations are required to evaluate the likelihood function. There is a number of proposed methods of approximation (Roberts, 2014 and Hedeker, 2005), though there are three basic approaches namely: Approximation of the integrand, Approximation of the integral itself and Approximation of the data. The listed methods are discussed in the following sections.

4.4 Estimation Techniques for GLMMs

4.4.1 Laplace approximation (LA)

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). This method is based on an approximation of the integrand (Jiang, 2007). Suppose one wishes to approximate an integral in the form,

$$\int e^{Q(x)} dx \quad (4.8)$$

where $Q(x)$ is known and unimodal function and x is a $q \times 1$ vector of variables. If $\hat{x}$ is such that $Q(x)$ is minimized, then the second order Taylor series expansion of $Q(x)$ around $\hat{x}$ is

$$Q(x) \approx Q(\hat{x}) + \frac{1}{2}(x - \hat{x})' Q''(\hat{x})(x - \hat{x}) \quad (4.9)$$

where $Q''(\hat{x})$ is the Hessian of Q evaluated at $\hat{x}$.

This yields approximation to Equation (4.8)

$$\int e^{Q(x)}\,dx \;\approx (2\pi)^{\frac{q}{2}}|\,Q^{"}(\hat{x})|^{-\frac{1}{2}}e^{-Q'(\hat{x})} \tag{4.10}$$

The approximation to this integral uses as many different estimates of $\hat{x}$ as necessary according to the different models of function Q. Since $\gamma \sim N(0, G)$, it can be shown that the integral in the likelihood equation (4.7) is proportional to the integral in equation (4.8), where the function Q is given by

$$Q(\gamma) = \emptyset^{-1}\sum_{j=1}^{n_i}\left[y_{ij}\left(x'_{ij}\beta + Z'_{ij}\gamma\right) - b\left(x'_{ij} + Z'_{ij}\gamma\right)\right] - \frac{1}{2}\gamma' G\gamma \tag{4.11}$$

Thus, Laplace method can be applied. This approximation method tends to be better for larger cluster sizes and can be improved by adding higher order terms to the Taylor series expansion.

4.4.2 Gaussian Quadrature

It is pointed out above that Laplace approximation is based on the linearization method of the integrand. An alternative to this is the Gauss-Hermite quadrature and Adaptive Gauss-Hermite quadrature, which often approximates the integral or numerical integration because of its relation with Gauss densities that give an approximation to an integral in the following form (Roberts, 2014; Lui, and Pierce, 1994).

$$\int h(x)e^{-x^2}dx \tag{4.12}$$

To apply these two methods, the likelihood contribution for the i^{th} cluster in equation (4.6) must be represented in the form of the integral in equation (4.12). This is done by standardizing the random effects such that they have an identity variance-covariance matrix I. Let $\delta_i = G^{-\frac{1}{2}}\gamma_i$. Thus, δ_i has a normal distribution with mean 0 and variance-covariance matrix I. The linear predictor, therefore, becomes $\theta_{ij} = x'_{ij}\beta + z'_{ij}G^{-\frac{1}{2}}\delta_i$, which now contains the variance component in G. Thus, the likelihood contribution for the i^{th} cluster is given by

$$f_i\left(y_{ij}|\beta, G, \emptyset\right) = \int\prod_{j=1}^{n_i} f_{ij}\left(y_{ij}|\gamma_i, \beta, \emptyset\right)f(\gamma_i|G)d\gamma_i \tag{4.13}$$

$$= \int \prod_{j=1}^{n_i} f_{ij}\left(y_{ij}\middle|\delta_i, \beta, G, \emptyset\right) f(\delta_i) d\delta_i \tag{4.14}$$

Therefore, this equation is now in the form of equation (4.12) and can be approximated using the Gauss-Hermite quadrature or adaptive Gauss-Hermite quadrature. In Gauss-Hermite quadrature the integral in equation (4.12) is approximated by

$$\int h(x) e^{-x^2} dx \approx \sum_{i=1}^{K} w_i h(x_i) \tag{4.15}$$

where the nodes x_i are the solutions of the K^{th} order to the Hermite polynomial and the $w_i$are suitable corresponding weights. The values of x_i and w_i for $i = 1,2,3, \dots, 20$ are found in tables given by (Abramowitz & Stegun, 1972). Increasing K improves the approximation. However, in this case, when the sum is taken from 1 to K, the Gauss-Hermite quadrature gives accurate solutions for all polynomials of degree $2K - 1$ (Roberts, 2014; McCulloch & Searle, 2001). The only disadvantage to this method of approximation is the quadrature points x_i chosen independently of the function, $h(x)$, hence may result in x_i not lying in the region of interest (Roberts, 2014; Pinheiro & Bates, 1995). This method of random effects in the model is increased (Hedeker, 2005).

In order to overcome the difficulties with Gauss-Hermite quadrature discussed above the quadrature points are rescaled and shifted such the integrand in equation (4.12) is sampled in a suitable range (Roberts, 2014; Lui & Pierce, 1994). This method is referred to as the adaptive Guass-Hermite quadrature, is based on centering the quadrature points with respect to the mode of the function being integrated and scales them according to the estimated curvature at the mode (Roberts, 2014; Hartzel *et al.,* 2001). This method requires significantly fewer quadrature. However, this adaptive Gauss-Hermite quadrature is much more time consuming to compute as the mode and curvature is calculated for each cluster in the dataset (Roberts, 2014; Hartzel *et al*., 2001). The adaptive Gauss-Hermite quadrature reduces to the Laplace Approximation when $K = 1$.

Newton Raphson and Fisher Scoring iterative producers can be utilized to maximize the likelihood after applying these numerical approximations. These methods work relatively well in the case of a single random effects or even when there are two or three nested random effects in the model. However, for structures that are more complicated these methods fail (Roberts, 2014; McCulloch & Searle, 2001).

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 $Var(Y_{ij}|U_i)$. This method used Taylor expansion around estimates $\hat{\beta}$ and $\widehat{U}$ of fixed and random effects respectively (Dlamini, 2016; Bolker *et al*., 2009; Moeti, 2007). Thus

$$\begin{aligned} Y_{ij} &= \mu_{ij} + \epsilon_{ij} = h\left(X'_{ij}\beta + Z'_{ij}U\right) + \epsilon_{ij} \\ &\approx h\left(X'_{ij}\hat{\beta} + Z'_{ij}\widehat{U}\right) + h\left(X'_{ij}\hat{\beta} + Z'_{ij}\widehat{U}\right)X'_{ij}\left(\beta - \hat{\beta}\right) + h\left(X'_{ij}\hat{\beta} + Z'_{ij}\widehat{U}\right)Z'_{ij}\left(U - \widehat{U}\right) + \epsilon_{ij} \\ &= \hat{\mu}_{ij}V\left(\hat{\mu}_{ij}\right)X'_{ij}\left(\beta - \hat{\beta}\right) + V\left(\hat{\mu}_{ij}\right)Z'_{ij}\left(U - \widehat{U}\right) + \epsilon_{ij} \end{aligned} \quad (4.16)$$

and

$$Y_{ij} = \hat{\mu}_i + \hat{V}_i\hat{X}_i\left(\beta - \hat{\beta}\right) + \hat{V}_i Z_i\left((U) - \widehat{U}\right) + \epsilon_i$$

where $\hat{\mu}_i$ contains values of $\hat{\mu}_{ij} = h(X'_{ij}\hat{\beta} + X'_{ij}\widehat{U})$, V_i is the diagonal matrix with elements $V\left(\hat{\mu}_{ij}\right) = h\left(X'_{ij}\hat{\beta} + Z'_{ij}\widehat{U}\right)$, X_i and Z_i contain the X'_{ij} and Z'_{ij} respectively.

Rearranging the above expression and multiply by V_i^{-1} we obtain

$$\begin{aligned} Y_i^* &= \hat{V}_i^{-1}(V_i - \hat{\mu}_i) + X_i\hat{\beta} + Z_i\widehat{U} \\ &\approx V_i\hat{\beta} + Z_i\widehat{U} + \epsilon_i^* \end{aligned} \quad (4.17)$$

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

Algorithm

Step 1: Given starting value for parameter β, $\emptyset$ and G. In the marginal likelihood, empirical Bayes estimates are calculated for U_i and pseudo data Y_i^* are computed. Step 2: Approximate linear mixed model is fitted, which gives updated estimates for β, $\emptyset$ and G. The updated estimates are used to update the pseudo data. This entire scheme is iterated until convergence is reached, and the 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 (Dlamini, 2016; Molenberghs and Verbeke, 2006).

4.5 Generalized Linear Mixed Models in SAS

The PROC GLIMMIX procedure allows GLMM to be fitted to the data. The random statement specifies the random effect to be incorporated into the model. To account for the heterogeneity between clusters in the under-five child mortality data. An intercept term that varied at cluster level was included in the model, thus resulting in a random intercept model. Once again, the logit link function with binary distribution was specified. The model was fitted using the Laplace approximation method, as this is likelihood based, hence it allows comparison of models using model selection criteria such AIC and BIC. The necessity for a random intercept was assessed by testing the responding covariance parameter if it is equaled to zero. This was done using the COVTEST statement in SAS, which produces likelihood ratio tests for covariance parameters. Since the parameter under the null hypothesis fell on the boundary of the parameter. The p-value for the test that was determined using a linear combination of central Chi-square probabilities. Table 4.2: below shows the null hypothesis of the covariance parameter equal to zero was rejected, thus suggestion random cluster was significant in the model.

Table 4. 2: Test of covariance parameters based on the likelihood

Label	DF	-2Log Likelihood	χ^2	P-value
No G-side effects	1	1016.30	3117.33	0.0034

Table 4.3: below gives the type 3 tests of fixed effects the model fitted using laplace method in GLMMs. The denominator degree of freedom (Den DF) was calculated as 2861. 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. The size pf child at birth, currently breastfeeding, birth order number, wealth index, number of children 5 and under, total number of children ever born, and source of drinking water were all significant. The Pearson chi-square statistics over its degrees of freedom was 0.88, which is close to one and this indicates the data was properly modeled.

Table 4. 3: Type III Tests of Fixed Effects

Effects	Numerator DF	F-Value	P-Value
Mother's age	2	0.39	0.6782
Size of child at Birth	2	14.86	<0.0001
Currently breastfeeding	1	14.22	0.0002
Birth order number	2	1.76	0.0318
Marital status	2	2.53	0.0729
Ethnicity	2	2.35	0.0955
Wealth index	2	2.11	0.0091
Number of children 5 and under	1	6.58	0.0104
Total children ever born	1	4.20	0.0405
Source of drinking water	1	5.70	0.0170
Province	8	1.52	0.1432

The variance component for the random effect was estimated as 0.1204 with a standard error of 0.4945 using Laplace. This estimate is relatively far from zero, hence confirming the need for the random effect in the model as shown in the table 4.4: below.

Table 4. 4: Random effect and model information

Random effects			
	Laplace Estimate (SE)	**Gauss-Hermite Quadrature Estimate (SE)**	**Penalized Quasi-Likelihood Estimates (SE)**
Variance (Intercept)	0.0637(0.6313)	0.1517 (0.3212)	0.0097 (0.2713)
Model Information			
-2 Log likelihood AIC	1016.29 1068.29	9.7275E8	514.7135

Table 4.5, Table 4.6 and Table 4.7: shows the solution for the fixed effects. Parameter estimates, standard errors, P-value, odds ratio and 95% confidence intervals. The estimated parameters for the model are fitted in GLMMs using three different estimation methods, namely Laplace, Gauss-Hermite Quadrature and Penalized Quasi-Likelihood. The models fitted are random intercept models and it can be observed that standard errors are slightly larger than those in the fitted model in section 3.13.1. The parameters estimated are found to be different in all three methods; however, the parameters estimated by Laplace and Penalized Quasi-Likelihood are found to be more significant than those estimated by the Gauss-Hermite Quadrature method. 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 3.13.1 and this is because this model accounts for the random effects. The odds ratio obtained using the procedure PROC GLIMMIX are lightly different from those obtained using PROC LOGISTIC in GLMs and the variables that were found to be significant in logistic regression may not found to be significant in GLMMs. Table 4.6 and 4.7 can be obtained in Appendix B.

The effect of the size of a child at birth (average weight) was found to be negatively associated with under-five mortality (p-value=0.0003). The corresponding odds ratio was 0.796 with (95% CI: 0.503; 1.261). The odds of death for average birth size were estimated to be 0.796 times the odds of death for very small weight of a child. The effect of not breastfeeding was found to be positively associated (p-value=0.0002) with under-five child mortality. The corresponding odds ratio was 3.015 with (95% CI: 1.699; 5.352). The odds of death for a child from a mother who does not breastfeed were estimated to be 3.015 times the odds of death for a child from a mother who breastfeed. The effect of childbirth order number that is between two to three births was found to be negatively associated with under five child mortality with (p-value=0.0001). The corresponding odds ratio was 0.775 times with (95% CI: 0.458;1.311). The odds death of a childbirth order number that is between two to three births was estimated to be 0.775 times the odds of death of the first birth. The effect of childbirth order number of more than three births was found to be positively associated with under five child mortality with (p-value=0.0001). The corresponding odds ratio was 1.206 times with (95% CI: 0.574;2.534). The odds death of a childbirth order number that is more than three births was estimated to be 1.206 times the odds of death of the first birth. The effect of number of children 5 year and under in household that is above two was found to be positively associated (p-value=0.0104) with under-five child mortality. The corresponding odds ratio was 0.328 with (95% CI: 0.140; 0.769). The odds of death for a child from a mother with two or more children alive were estimated to be 0.328 times the odds of death for a child from a mother who has less than two children alive. The effect of total number of children ever born that is above two was found to be positively associated (p-value=0.0405) with under-five child mortality. The corresponding odds ratio was 1.773 with (95% CI: 1.025;3.068). The odds of death for a child from a mother who gave birth to two or more children alive were estimated to be 1.773 times the odds of death for a child from a mother who has less than two children alive. The effect of source of drinking water was found to be positively associated with under five child mortality with (p-value = 0.0170). The corresponding ratio was 1.681 times with (95%CI: 1.097;2.574). The odds of death for a child from a mother who does not drink safe water was estimated to be 1.681 times the odds of death for a child from a mother drinks safe water.

Table 4. 5: Laplace, estimated coefficients, odds ratios, standard errors, p-values and confidence interval.

Analysis of Maximum Likelihood Estimates and Odds Ratio estimates						
Demographic characteristics						
Parameter	**Estimate**	**Standard Error**	**P-Value**	**Odds ratio**	**95%Confidence interval Lower**	**Upper**
Intercept	-17.1932	345.60	0.9603			
Mother's age (ref >35 years)						
<20 years	-0.02694	0.2730	0.4633	0.973	0.570	1.663
20 – 35 years	0.1713	0.2336	0.9214	0.187	0.751	1.876
Size of child at birth (ref. Very large)						
Very small	0.9339	0.2548	0.3306	2.544	1.544	4.193
Average	-0.2281	0.2344	0.0003	0.796	0.503	1.261
Currently breastfeeding (ref. Yes)						
No	1.1037	0.2926	0.0002	3.015	1.699	5.352
Birth order number (ref. First birth)						
2 – 3 births	-0.2550	0.2681	0.0001	0.775	0.458	1.311
More than 3 births	0.1876	0.3785	0.0001	1.206	0.574	2.534
Current marital status (ref. never married)						
Married	-0.2846	0.2245	0.1503	1.381	0.889	2.145
Living with partner	0.3230	0.2676	0.2877	0.752	0.445	1.271
Ethnicity (ref. Others)						
Black African	1.2099	345.63	0.9718	3.353	0.850	1.436
Coloured	1.5310	345.63	0.9757	4.623	1.263	4.671
Socio-Economic Characteristics						
Wealth index (ref. Rich)						
Middle	0.2120	0.3171	0.5038	1.236	0.664	2.302
Poor	0.5427	0.2875	0.0592	1.721	0.979	3.023
Number of children 5 and under (ref.< 2 Children)						
2 or more children	-1.1137	0.4341	0.0104	0.328	0.140	0.769
Total children ever born (ref. less than 2 Children)						
2 or more children	0.5729	0.2795	0.0405	1.773	1.025	3.068
Household Environment Characteristics						
Source of drinking water (ref. safe water)						
Not safe water	0.5192	0.2175	0.0170	1.681	1.097	2.574
Province (ref. Western Cape)						
Eastern Cape	0.0217	0.6049	0.9715	1.022	0.312	3.346
Free State	0.4862	0.6257	0.4371	1.626	0.477	5.546
Gauteng	-0.0042	0.6260	0.9946	0.996	0.292	3.398
KwaZulu-Natal	-0.2719	0.6167	0.6593	0.762	0.227	2.553
Limpopo	-0.5988	0.6276	0.3401	0.549	0.161	1.881
Mpumalanga	0.4656	0.5900	0.4301	1.593	0.501	5.066
North West	0.0051	0.6197	0.9935	1.005	0.298	3.388
Northern Cape	0.2952	0.6522	0.6509	1.343	0.374	4.826

4.6 Summary of Generalized Linear Mixed Models

As discussed earlier on in the Chapter that GLMMS are natural extension of the GLM. 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 LMMs. GLMMs could be used to include correlations in the model and identify sensitive subjects. For GLMMS the modeling is straight forward, first one has to identify the distribution of data, understand what needs to be modeled and 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, Gauss-Hermite Quadrature and Penalized Quasi-Likelihood. The results obtained using PROC GLIMMIX procedure and the Laplace method was the preferred method for the results since it is more accurate than Penalized Quasi-Likelihood and faster than Gauss-Hermite Quadrature; however, the Penalized Quasi-Likelihood had more significant results. The alternative is to use general additive models.

Chapter 5

Generalized Additive Models

5.1 Introduction

The statistical models that have been discussed above assume linearity parametric form for the explanatory variables. However, 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 explanatory variables. However, these parametric models are not flexible for modeling a complicated relationship between response set of explanatory variables. The limitation of the parametric modeling is that it is restrictive in many cases. The section describes the flexible statistical non-parametric models that can be used to model complicated relationship between the response and set of explanatory variables. These models are known as generalized additive models (GAMs) proposed by (Hastie and Tibshirani, 1986). These models assume that the mean of the dependent variable depends on an additive predictor through a non-linear link function. GAMs can handle non-linear, linear and non-monotonic relationships between response and predictor variables. They can be used in settings that include standard continuous response regression, count, binary response survival data and time series data. GAMs are suitable for exploring the data set, visualizing the relationship between the dependent variable and the set of explanatory variables (Liu, 2008). The GAMs generalize the generalized linear model by replacing the linear form $\beta_0 + \sum_{j=1}^{p} x_j \beta_j$ with the additive form $f_0 + \sum_{j=1}^{p} f_j\left(x_j\right)$, where f_j is unspecified (non-parametric) function. To determine the appropriate smooth function f, the steps in GLM are replaced by the non-parametric regression steps. Therefore, the GAM using the notation of (Wood, 2006) can be presented as:

$$g(\mu_i) = X_i^* \theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}) + \cdots \tag{5.1}$$

where $\mu_i \equiv E(Y_i)$ and Y_i has a distribution that follows exponential family distribution, X_i^* is the design matrix, θ is the corresponding parameter vector and $f_j(.)$ are smooth functions of covariates. Model (5.1) is simply an additive model if g is the identity link and the response is normally distributed (Faraway, 2006). This function $f_i\left(x_j\right)$can be estimated in a flexible manner using cubic spline smoother, in an iterative method called back fitting algorithm (Lui, 2008; Hastie and Tibshirani, 1990). The name cubic is from the piecewise polynomial fit, with the order $k = 3$ (Lui, 2008 and Dlamini, 2016). We define smoother as the tool for summarizing the trend of a dependent variable as function of one or more independent variables. The smoother produces an estimate known as smooth (Lui, 2008). The main property of the 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 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. 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 smooth is the tool for summarizing the trend of response variable Y as function of one or more independent variables $X_1, \ldots . X_p$ (Lui, 2008). We first model the simplest smooth function where the model contains one smooth function of one independent variable.

$$y_i = f(x_i) + \epsilon_i \tag{5.2}$$

where y_i is the response variable, x_i is the covariate, $f(.)$ 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 $(y - X\beta)'(y - X\beta)$ will not yield the smooth curve at all (Wood, 2006). However, the cubic spline smoother does forces smoothness on $f(x)$. Then the model is fitted by minimizing the following penalized least square function.

$$\sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 + \lambda \int_a^b \big[f^{"}(x)\big]^2 dx \tag{5.3}$$

where λ is fixed constant and $a \leq x_1 \leq \cdots \leq x_n \leq b$. We assume (a, b) includes all possible range. The function f can be approximated by linear combination of basic functions $f_j(x)$ as $f(x) = \sum_{j=1}^{q} f_j(x)\, \beta_j$ and $\int \big[f^{"}(x)\big]^2$ measure the "Wiggliness" of the function $f(x)$. If the $\int \big[f^{"}(x)\big]^2 = 0$ (indicate the straight line) then we have a function f that is a linear function. However, a non-linear function of f will produce values $\int \big[f^{"}(x)\big]^2 > 0$ (smoother f 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 $\big(y_i - f(x_i)\big)^2$ and the model smoothness (Dlamini, 2016; Hastie and Tibshirani, 1990). The larger the value λ the smoother f becomes and the penalty term becomes more important. Moreover, the small values of λ yield a wiggly curves and penalty become unimportant (Lui, 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 f is the linear combination of the parameters and one can show that the penalty from penalized least square is a quadratic form of β. This is given below.

$$\int \left[f^{"}(x)\right]^2 dx = \beta' H \beta \tag{5.4}$$

Suppose now the model has two smoothers as follows

$$Y_i = f_1(x_i) + f_2(x_i) + \epsilon_i \tag{5.5}$$

The smoothers has the form $f_1(x) = \sum_{j=1}^{q1} b_j(x_i)\beta_j$ and $f_2(x) = \sum_{j=1}^{q2} b_{2j}(z_i)\gamma_j$,where x and y 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 spine basic functions of f_1 and f_2 respectively. When two smoothers are now used in place of one smoother then this the definition of Y as a function of q,X and β. However, the general form does not (Wood, 2012). The optimization becomes

$$\sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 + \lambda_1 \beta' H_1 \beta + \lambda_2 \beta^1 H_2 \beta \tag{5.6}$$

where 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, 2012). Here $H = \int d(x)d(x)'$dx is the penalty matrix, which consists of known coefficients and $d(x)$ is given by

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

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

$$(y - X\beta)'(y - X\beta) + \lambda \beta' H \beta \tag{5.7}$$

This can also be written as

$$y'y - y'X\beta - X'\beta'y + \beta'(X'X + \lambda H)\beta$$

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

$$\hat{\beta} = (X'X + \lambda H)^{-1}X'y \tag{5.8}$$

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

$$A = X(X'X + \lambda H)^{-1}X' \tag{5.9}$$

We first require some method for choosing λ.

5.4 Selection of Smoothing Parameters $\boldsymbol{\lambda}$

To minimize cubic 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 (Wood, 2006).

5.4.1 Average Mean Square and predictive Square Error

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

$$AMSE(\lambda) = \frac{1}{n}\sum_{i=1}^{n} E\left[\hat{f}_\lambda(x_i) - f(x_i)\right]^2 \tag{5.10}$$

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

$$PSE(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i^* - \hat{f}_\lambda(x_i)\right]^2 \tag{5.11}$$

AMSE and PSE differ by constant δ where Y_i^* is the new observation at $x_i, Y_i^* = f(x_i) + \epsilon_i^*$ and ϵ_i^* is independent of ϵ's. There are other methods for estimating and selecting λ, for example Cross Validation (CV) and Generalized Cross Validation (GCV).

5.4.2 Cross Validation (CV)

The Statistical approach for partitioning sample into two subsets is known as Cross Validation (Lui, 2008 and Wood, 2006). This technique is sufficient when the sample is large. The data recycled by switching the role of test samples and training in CV. Cross validation could be used in selecting $\boldsymbol{\lambda}$ by minimizing,

$$CV(\boldsymbol{\lambda}) = \sum_{i=0}^{n}\left[y_i - \hat{f}_\lambda^{-i}(x_i)\right]^2 \tag{5.12}$$

where $\hat{f}_\lambda^{-i}(x_i)$ 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 (GCV)

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

$$V_g = \frac{n\|y_i - X\hat{\beta}_\lambda\|}{[n - tr(F_\lambda)]^2} \tag{5.13}$$

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

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

5.4.4 Degrees of Freedom of a Smoother

Degrees of freedom is the other way of expressing the required smoothness of the function in terms of λ. In SAS, the procedure PROC GAM can select the value of a smoothing parameter simply by specifying the degree of freedom for the smoother and this is sometimes called effective number of parameters. The effective number of parameters indicates the amount of smoothing. Suppose there is a linear smoother say $F_\lambda$then the degrees of freedom are 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 (Lui, 2008).

5.5 Back fitting and Generalized Local Scoring Algorithm

The basic idea behind generalized additive models is to plot the value of the response variable together with independent variable 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. The main focus is to explore the data set and visualize the relationship between response and set of independent variables (Lui, 2008; Marx and Eilers, 1998). However, the GLMs focus specifically in estimation and inference. The data is divided into number of sections 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 algorithm 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 (Dlamini, 2016; 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

Back fitting is known as the simple iterative procedure used to fit a generalized additive model. It defined as the partial residual

$$R_j = Y - f_0 - \sum_{k \neq j} f_k(x_k)$$

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

Step1 Initialize:

$$f_0 = E(Y), f_1^1 = \cdots = f_p^1, m = 0.$$

Step2. Iterate: $m = m + 1$ for $j = 1$ to p do:

$$R_j = Y - f_0 - \sum_{k=1}^{j-1} f_k^m(x_k) - \sum_{k=j+1}^{p} f_k^{m-1}(x_k)$$

$$f_j^m = E(R_j|X_j)$$

Step 3: Calculate

$$RSS = AVG\left(Y - f_0 - \sum_{j=1}^{p} f_j^m(x_j)\right)^2$$

until fails to decrease. $f_j^m(.)$ Denotes the estimate of $f_j(.)$ 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

Step 1: Initialize,

$$f_0 = E(Y), f_1^1 = \cdots = f_p^1, m = 0$$

Step2: Iterate $m = m + 1$, from the adjusted dependent variable

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

$$\eta^{m-1} = f_0 + \sum_{j=1}^{p} f_j^{m-1}(x_{ij}),$$

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

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

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 $f_i^m(.)$ 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 (Lui, 2008).

$$\eta^{est}(x) = \eta^{given} + \delta \tag{5.14}$$

Here,

$$\delta = \frac{Score\ fuction}{Expected\ Information\ matrix}$$

$$\delta = \frac{\frac{\partial l}{\partial \eta}}{E\left(-\frac{\partial^2 l}{\partial \eta^2}|x\right)} \tag{5.15}$$

$$= E\left(\eta(x) - \frac{\frac{\partial l}{\partial n}}{E\left(\frac{-\partial^2 l}{\partial \eta^2}|x\right)}|x\right)$$

Using chain rule, we have that

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

$$\frac{\partial l}{\partial \mu_i} = \frac{1}{\mu_i} - (1 - y_i)\frac{1}{(1 - \mu_i)} \tag{5.16}$$

$$= \frac{y_i - \mu_i}{(1-\mu_i)\mu_i}$$

We know that $Var(Y_i) = E\left(Y_i^2\right) - \left(E(Y_i)\right)^2$

$$\begin{aligned} Var(Y_i) &= E\left(Y_i^2\right) - \left(E(Y_i)\right)^2 \\ &= 1^2\mu_i + 0^2(1-\mu_i)(-\mu_i) \\ &= \mu_i(1-\mu_i) \end{aligned} \tag{5.17}$$

and

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

Thus

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

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

Therefore

$$E\left(\frac{\partial^2 l}{\partial \eta^2}|x\right) = -\left(\frac{\partial \mu}{\partial \eta}\right)^2 V^{-1}$$

$$\eta^{est}(x) = E\left[\eta(x) + (Y-\mu)\frac{\partial \eta}{\partial \mu}|x\right] \tag{5.18}$$

Replacing the conditional estimation with smoothers, we have the improved estimates

$$\eta^{est}(x) = smoother\left[\eta(x) + (Y-\mu)\frac{\partial \eta}{\partial \mu}|x\right] \tag{5.19}$$

5.6 Estimation of the Parameter Estimation $\boldsymbol{\beta}$

If the data is non-normal, the framework of the GLM can be applied. The linear predictor is modeled as the sum of the B-spline and iterative method is used. The number of B-spline and the value of the coefficients or amplitudes will influence the smoothness of the curve. If these are almost equal, then the curve will be flat. The curve will show many of wiggles if the amplitude varies widely.

5.6.1 Splines

The spline curve is a piecewise polynomial curve that joins two or more polynomial curves. The locations of the joins are known as knots. In addition, there are boundary knots, which can be located at or beyond the limits of data.

5.6.1.1 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. Even though non-parametric provides rich exploratory flexibility it is not possible 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: Firstly, one the range of the covariate. Secondly, on the number and position of knots (equally spaced), and the degree of the B-spline.

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

$$S = \sum_{i=1}^{N} \left(y_i - \sum_{t-1}^{n} b_{it} a_t \right)^2 \tag{5.20}$$

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 B and B known as B-spline matrix of dimension $N \times n_i$.

5.6.1.2 P-splines

There is another way representing the cubic splines using 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

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

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

5.6.2 Penalized Likelihood and Estimation

The penalized likelihood is another way to find regression coefficients for categorical variables. The likelihood is maximized by using iterative methods such as the Newton Raphson algorithm and scoring method. Newton Raphson method is a technique used to find the zeros of a function taking real values (Wood, 2006; Marx and Eilers, 1998).

5.6.2.1 Penalized Likelihood

The disadvantage of 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

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

where $S = \sum_{j=1}^{p}\lambda_j H_j$, L denotes the 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 $\hat{\beta}_j's$.

5.6.2.2 Estimation

The penalized log-likelihood in Eq. (5.23) can be maximized through iterative Re-weighted Least-Squares. Here we assume that λ_j is known. To maximize this equation, we need to take its derivative with respect to $\beta_j$and equate to zero, thus we have,

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

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

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

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. To find Least Square, we take a derivative with respect to β_j and equating to zero. The iterative of equations will be as in Eq. (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).

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

The pseudo data is defined as

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

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

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

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

Step 1: 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 $\hat{\beta}^{[k+1]}$, so that $\eta^{[k+1]} = X\beta^{[k+1]}$ and increase the value of k by one unit. Then, the converged $\hat{\beta}$ solves equation (5.24).

5.7 The Generalized Additive Logistic Model

The most popular and widely used approach for binary is Logistic regression. In this model the outcome is coded the same as in Chapter 3, with zero indicating the child is not alive and one the child is alive.

$$Y_i = \begin{cases} 1, & Child\ is\ alive & \pi_i(x) \\ 0, & Child\ is\ not\ alive & 1-\pi_i(x) \end{cases}$$

where $X = (X_1, \dots, X_p)$ is a vector of covariates.

$$logit\big(\pi(x)\big) = log\frac{\pi(x)}{1-\pi(x)} = \beta_0 + \sum_{j=1}^{P}\beta_j\left(x_{ij}\right)$$

and

$$\pi(x) = \frac{Exp\left(\beta_0 + \sum_{j=1}^{P}\beta_j(x_{ij})\right)}{1+Exp\left(\beta_o+\sum_{j=1}^{P}\beta_j(x_{ij})\right)} \qquad (5.29)$$

In logistic GAM. The basic Idea is to replace the linear predictor with an additive one. However, the logistic regression assumptions still apply except for linearity assumption.

$$logit\big(\pi(x)\big) = log\frac{\pi(x)}{1-\pi(x)} = f_0 + \sum_{j=1}^{P} f_j\big(x_{ij}\big)$$

and

$$\pi(x) = \frac{Exp\big(f_0 + \sum_{j=1}^{P} f_j\big(x_{ij}\big)\big)}{1 + Exp\big(f_0 + \sum_{j=1}^{P} f_j\big(x_{ij}\big)\big)} \qquad (5.30)$$

The functions $f_1, f_2, \dots, f_p$ are estimated by the algorithm described above, back fitting algorithm. This happens when the model consists of parametric and non-parametric terms.

Generally, let $E(Y|X) = \mu$, $(\mu_i = 1 * \pi_i + 0 * (1 - \pi_i) = \pi_i)$

So that,

$$\eta(x) = g(\mu) = log \frac{\pi(x)}{1 - \pi(x)} \tag{5.31}$$

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

$$Z_i = \eta_i + (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right) \tag{5.32}$$

Instead of fitting an additive model Y, we fit an additive model to the Z_i 's where it is treated as the response variable Y in $\mu = f_0 + \sum_{j=1}^{P} f_0\left(x_{ij}\right)$. This algorithm is to fit the smoothing functions and is analogous to the algorithm described above (Lui, 2008).

5.7.1 Fitting Generalized Additive Models Logistic using GAM procedure

The Generalized additive models are useful in finding predictor- response relationships in many kinds of data without using a specific. GAMs combine the ability to explore many nonparametric relationships simultaneously with distributional flexibility of generalized linear models. The SAS procedure PROC GAM is a powerful tool for the nonparametric regression model, and it provides great flexibility in predictor-response relationships. Carrying out exploratory modelling with PROC GAM could inspire parsimonious parametric models. Thus, using PROC GAM, under model option, some variables are included in the keyword spline; in this case, non-linearity assumption is made for them. In the present section, we assume some of the covariates have a linear relationship with the log odds and some have non-linearity, this yields the semi parametric model.

5.7.2 Fitting the Logistic Additive Model

We consider the first part of the output that is obtained using PROC GAM procedure. Table 5.1: depicts a summary for the back fitting and local scoring algorithms. The deviance for the final estimate is also provided in the table 5.1 and it can be used in computing the AIC as presented below

$$\begin{aligned} AIC &= Deviance + 2pf, \\ &= 863.08 + 2 \times 13 \times 1, \\ &= 889.06 \end{aligned} \tag{5.33}$$

where p is the model degrees of freedom and f is the scale parameter ($f = 1$ for Binomial and Poisson). The model degrees of freedom are $1 + 12 = 13$. This AIC value can be used to compare models fitted by PRO GAM. In PROC GENMOD the AIC value is calculated as

$$AIC = -2LL + 2p,$$

where LL is the log likelihood of the fitted model.

Table 5. 1: Summary of algorithms used in fitting the model

Iteration Summary and Fit Statistics	
Number of local scoring iterations	43
Local scoring convergence criterion	8.67E-17
Final Number of Back fitting Iterations	1
Final Back fitting Criterion	8.42E-17
The Deviance of the Final Estimate	863.08

Table 5.2: shows the linear portion and parameter estimates for parametric part of the model, standard errors, t-values and p-values. The effect of size of child at birth, small weight was found to be positively associated with under-five child mortality (p-value= 0.0001). The breastfeeding was found to be positively associated with child mortality (p-value = 0.0007). The linear predictors mother's age, number of children 5 and under, total children ever born, and birth order number were found to be significantly associated with under-five child mortality, with their p-values less than 0.05 significance level. Table 5.3: shows the summary of smoothing components of the nonparametric part of the model. This table presents the smoothing parameter, degrees of freedom, value of GCV for each component and a number of unique observations. The degrees of freedom are an indication of the amount of smoothing. The more the smoothing means less degrees of freedom or higher span. Each smoothing component has approximately 4 degrees of freedom. For univariate spline component, one degree of freedom is taken up by parametric linear portion of the model, so the smoothing was almost equal to one and the corresponding degree of freedom is 3.

Table 5. 2: Analysis of Model

Regression Model Analysis Parameter Estimates				
Parameter	**Parameter Estimate**	**Standard Error**	**t-value**	**p-value**
Intercept	-15.081	133.964	-0.13	0.8947
Size of child at birth (ref. Very large)				
Average	-0.1523	0.2462	-0.62	0.5363
Small	1.0475	0.2600	4.03	<.0001
Currently breastfeeding (ref. yes)				
No	1.0792	0.3198	3.37	0.0007
Current marital status (ref. Never married)				
Married	-0.3016	0.2834	-1.06	0.2874
Living with partner	0.0668	0.2436	0.27	0.7840
Ethnicity (ref. Others)				
Black African	11.728	113.9610	0.10	0.9180
Coloured	9.8770	113.9636	0.09	0.9309
Wealth index (ref. Rich)				
Middle	-0.1087	0.3296	-0.33	0.7417
Poor	0.3668	0.2930	1.25	0.2107
Source of drinking water (ref. safe water)				
Not safe water	0.2896	0.2156	1.34	0.1793
Province (ref. Western Cape)				
Eastern Cape	0.1992	0.6309	0.32	0.7522
Free State	0.8304	0.6420	1.29	0.1959
Gauteng	-0.0315	0.6565	-0.05	0.9618
KwaZulu-Natal	-0.2114	0.6326	-0.33	0.7383
Limpopo	-0.2441	0.6576	-0.37	0.7105
Mpumalanga	0.5926	0.6090	0.97	0.3306
North West	0.1226	0.6300	0.19	0.8457
Northern Cape	0.5927	0.6749	0.88	0.3799
Mother's Age				
Linear (Mother_Age)	-0.0524	0.0206	-2.54	0.0111
Number of children 5 and under				
Linear (NoOfChildrenU5)	-1.1145	0.1201	-9.28	<0.0001
Total children ever born				
Linear (ChildrenEverBorn)	1.7876	0.1964	9.10	<0.0001
Birth order number				
Linear (Bord)	-1.3622	0.1992	-6.84	<0.0001

Table 5. 3: Smoothing Model Analysis

Smoothing Model Analysis Fit Summary for Smoothing Components				
Component	**Smoothing Parameter**	**DF**	**GCV**	**Num Unique Obs**
Spline (Mother Age)	0.998820	3	2.0601	35
Spline (NoOfChildrenU5)	0.984050	3	27.362	8
Spline (ChildrenEverBorn)	0.995178	3	25.062	11
Spline (Bord)	0.999157	3	421.32	12

Table 5.4: below shows the most important part of PROC GAM results, Analysis of Deviance. This table provides a Chi-Square (χ^2) test comparing the deviance between full model and the one without non-parametric component variable for each smoothing effect in the model. The analysis of deviance results shows that non-parametric effects of three continuous predictors are significant since their p-value is less 5% significance level.

Table 5. 4: Analysis of deviance

Smoothing Model Analysis Analysis of Deviance				
Source	**DF**	**Sum of Squares**	**Chi-Square**	**P-value**
Spline (Mother Age)	3	14.095830	14.0958	0.0043
Spline (NoOfChildrenU5)	3	16.062196	16.0622	0.0009
Spline (ChildrenEverBorn)	3	2.05271	2.0525	0.5616
Spline (Bord)	3	15.278154	15.2782	0.0033

Figure 5.1: shows plots of the partial prediction for each of the continuous predictor considered. These plots can be used to investigate as to why PROC GAM and PROC GENMOD provide different result. 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. The plots show that the partial predictions corresponding to total children ever born have a quadratic pattern. This suggests that under-five mortality was associated with a quadratic pattern for total children ever born. The mother's age, birth order number and number of children 5 and under have 95% confidence limits containing zero axes and the line was straight, this means that mother's age, birth order number and number of children 5 and under in household had no quadratic effect on the child survival status.

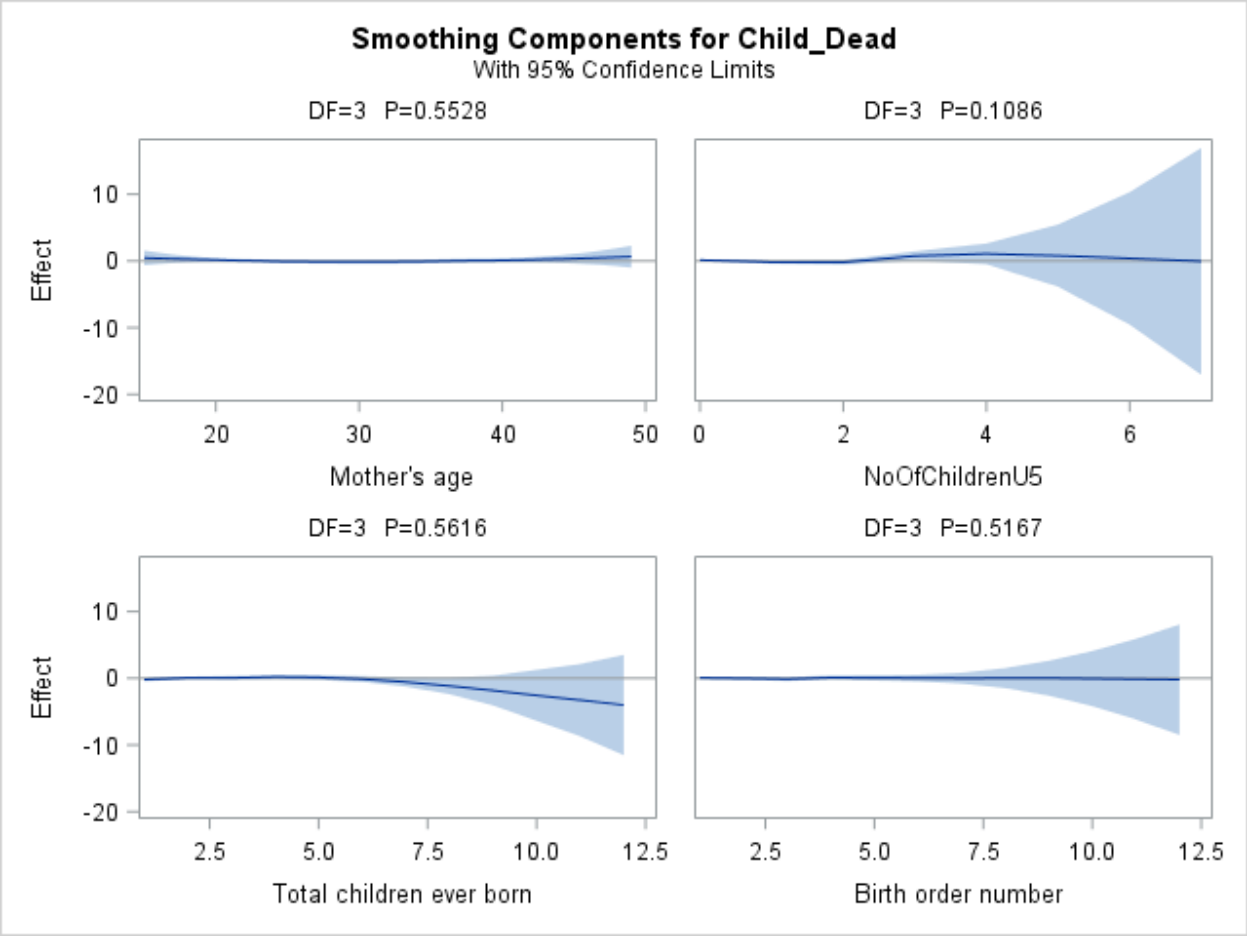

Figure 5. 1: Partial prediction for each predictor

5.8 Summary of Generalized Additive Model

Generalized Additive Models are an alternative method to Logistic regression, since the assumption about the linearity link function (logit) and predictors need to be made. This assumption may not hold in Logistic regression thus we introduce GAMs. 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 term. Secondly, select the 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 hypothesis can be tested using GLM methods on un-penalized GAM, and confidence intervals can be obtained using Bayesian smoothing model (Wood, 2006). With the use of PROC GAM, we found that under-five child mortality was significantly associated with the size of child at birth and breastfeeding. The under-five child mortality was also associated with linear pattern of mother's age, number of children 5 and under, total children ever born and Birth order number. Lastly, under-five child mortality was associated with quadratic pattern of total children ever born.

Chapter 6

Discussion

The main objective of the study was to determine the factors associated with the under-five child mortality in South Africa. The determined factors can be used to help the South African Government, non-governmental organizations and other partners in the health sector to know and understand the important areas they need to focus on. Therefore, this will help to develop more policies, programmes and evaluate progress made towards achieving the Millennium Development Goals 4. In order to achieve these objectives statistical models such as generalized linear models, survey logistic regression, generalized linear mixed models and generalized additive models were used to identify the risk of factors affecting under-five child mortality. The data used in this study was obtained from South Africa Demographic and Health Survey 2016, which was conducted by Statistics South Africa (Stats SA) in partnership with the South African Medical Research Council (SAMRC) at the request of the National Department of Health (NDoH). The response variable was child survival status indicating whether the child is dead or alive, coded as zero for dead and 1 for alive.

The associated demographic, socio-economic and environmental factors used in the study were: age of the mother, sex of child, birth order number, breastfeeding, marital status, ethnicity, education, wealth index, employment, number of children 5 years and under in household, total children ever born, residential area, main source of water and main floor material. The generalized linear models known as logistic regression assumes that the survey data was obtained using simple random sampling. Due to large number of variables, stepwise selection procedure was adopted to eliminate non-significant variables. After fitting logistic regression, the size of child at birth, breastfeeding, birth order number, ethnicity, number of children 5 and under in a household, total children ever born, source of drinking water and province were significantly associated with under-five child mortality. However, mother's age, marital status and wealth index were not significantly associated with under-five child mortality. The model checking and goodness-of-fit test was performed using Hosmer-Lemeshow. The test failed to reject the model selection.

The model was refitted using survey logistic regression and generalized linear mixed models since they account for the complexity of the survey design. However, the conclusion reached by survey logistic regression was similar to the one reached using generalized linear mixed models, despite the differences in the models. From the results of logistic regression and survey logistic regression models presented in Chapter 3. We observe that the standard errors of logistic regression model are smaller than the standard errors of survey logistic regression model for each parameter estimate and that suggests under estimation of the variance (Heeringa *et al*., 2017 and Yirga *et al*., 2019). This shows that the assumption made in order to use logistic regression resulted in a wrong conclusion. We obtained the appropriate estimates considering the survey design features. The parameter estimates and odds ratios are almost the same for

both models. However, the confidence intervals for logistic regression are narrow and this has resulted in underestimation of the variance (Dlamini, 2016 and Yirga *et al*., 2019). 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 variable is binary. However, the assumption about linearity between log odds and independent variables need to be made. If the assumption does not hold the generalized additive models could be used as an alternative. Using generalized additive models, the under-five child mortality was found to be significantly associated with, size of a child at birth and breastfeeding. Under-five child mortality was also found to be significantly associated with linear predictors mother's age, number of children 5 and under in household, total children ever born and birth order number. Lastly, the under-five child mortality was also found to be significantly associated with the quadratic pattern for total children ever born and has no quadratic pattern with mother's age, number of children 5 and under in household, and birth order number.

From the results, we observed that the incidents of child death for a mother who does not breastfeed was higher compared to incidents of child death from a mother who breastfeeds. Breastfeeding, initiated within the first hour of birth, provided exclusively for six months, and continued up to two years or beyond with the provision of safe and appropriate complementary foods, is one of the most powerful practices for promoting child survival and wellbeing (UNICEF, 2018). Generally, the breastfed children are less vulnerable to the risk of under-five child death than are artificially fed children. Similar results were reported by Akwara (1994); Barros and Victora (1990) that the absence of breastfeeding exposes children to various diseases that facilitate under-five child mortality. This could be attributed to, malnutrition, lack of vitamins and calcium. So, the survival of breastfed children is as a result of nutrient nourishment by breast milk (UNICEF, 2018).

This study has found the incidents of a child death to be more for a mother whose childbirth order number was more than three births and it was higher compared to incidents of a mother whose birth order number was less than two. The maternal mother might have less parental care for the children after the first born and that could lead under-five child mortality because children are not looked after properly and be exposed to hazardous environment. Elliott (1992) also reported that first born and early born children will spend early years having exclusive attention of parents while later born will have to compete with siblings over resources. Younger siblings are likely to be introduced to developmentally inappropriate activities by older siblings. Thus, many siblings increase the likelihood of communicable diseases being introduced into the family, and younger siblings may be more susceptible to these diseases (Elliott, 1992).

The incidents of a child death of a very small weight of child at birth was found to be higher compared to the incidents of an average weight size of the child at birth. This is a major determinant in developing countries which is caused by poor maternal nutrition status at conception, low gestational weight gains due to inadequate dietary, and short maternal stature due to the mother's own childhood under nutrition (Pojda and Kelly, 2000). Similar results were

reported by Suparmi *et al*. (2016) that low birth weight is closely associated with under-five child mortality and affects child development and future risk of chronic disease.

The results also showed the total number of children ever born and the number of children 5 year and under in household have influence in under-five child mortality. The risk of child death for a mother with more than two children was found to be higher as compared to the incidents of child death for a mother with less than two children. These results are consistent with the results reported by Woldenmicael (2001); Mambugu (2014) that, where many children live together, there is high chance of spreading germs and poor hygiene. Therefore, this may lead to health problems. Many children in a household increase the likelihood of having disease like infections because of crowding and competition for the available resources. An environment of such nature has 60% chance of experiencing diarrhea if there are six or more children living in the household than if the number is less than three (Woldemicael, 2001 and Wambugu, 2014).

The differences among population groups has been found to be linked with under-five child mortality. For example, the 1998 South Africa Demographic and Health Survey estimated that in 1996 the under-five child mortality for Africans was 47%, for Coloured people 19% and 11% for White people. This could be caused by high inequality among population groups in accessing proper health services in the country. These results are consistent with the results reported by Heaton and Amoateng (2007); Yach (1994) where similar pattern to this were estimated, with 51% for Africans, 38% for Coloured people, 8% for Asians and 7% for White people.

The incidents of child death for a mother who does not drink safe water was found to be higher compared to incidents of child death from a mother who drinks safe water. The risk of potentially fatal diarrheal diseases is high among households with no clean drinking water or with no safe sanitation which increases the likelihood of under-five child mortality. Mahmood (2002) also reported that families living in households with piped water connected in their houses have a significantly lower post neonatal mortality than those families that depend on wells for drinking water.

In our study, provinces have been found to be closely associated with under-five child mortality. South Africa is divided into regions called province which are also divided into residential areas called urban and rural areas. This could be the difference between urban and rural health facilities since, South Africa is still a developing country. Rural mothers and children are often disadvantaged in term of access to basic health services that can lead to under-five child mortality. Similar work has been reported by Kabir *et al*. (2001); Kembo and Ginnken (2009) that risk of death of children is lower in the urban than that of rural areas. This is general expectation considering the level of development is more advanced for urban than for rural areas. One of the DHS studies from Brazil had shown urban areas had low child mortality. The differences were not clarified through urban life advantage, but community variables such as ecological setting, political economy and health system, played an important role through socioeconomic characteristics (Fotso, 2006).

Conclusion

The results of the study have demonstrated that there is a decline in under-five child mortality in South Africa. Similar results were reported by Rademer (2017) that the under-five child mortality has significantly declined in South Africa. Even though there is notable decline of under-five mortality, it should be noted that the pace of reduction sheds light on improvement of child health services in South Africa. We cannot only isolate the astounding decline in under-five child mortality to unfocused priorities but also there is equally a need to address demographic, socio-economic and environmental factors in order to reduce under-five child mortality. The South African government has adopted the Sustainable Development Goals (SDGs) in 2015 that are extension and expansion of the work done under Millennium Development Goals and established ambitious targets for improving child survival by 2030. The country also has its own National Development Program (NDP) that includes many ambitious targets that go beyond what SDGs hope to achieve by 2030. The NDP aims at alleviating poverty and inequality by improving health inequalities between the poor and non-poor by 2030. Therefore, these results can be used to project information on reduction of under-five child mortality to achieve the country's (NDP). The identified factors will assist policy makers to understand the areas they need to improve on in order to enhance the planning and evaluation of health policies to prevent under-five child mortality in South Africa.

Recommendations

Given the findings, a number of recommendations were made. Mothers should breastfeed to reduce more death of under-five child mortality. If the government strengthen the progress on reducing child mortality by tackling numerous policies and programme adjustments that have expanded coverage of child health interventions. Fighting poverty and inequality by improving the health of the poor and reducing health disparities between the poor and non-poor. Drinking unsafe water should be eliminated because source of water is amongst detrimental factors associated with under-five child mortality.

References

Abramowitz, MILTON., 85. l. A. Stegun, 1972: Handbook of Mathematical Functions. *National Bureau of Standards Applied Mathematics Series*, *55*, pp.589-626.

Akaike, H., 1974. A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike* (pp. 215-222). Springer, New York, NY.

Akwara, P.A., 1994. Breastfeeding and infant and child mortality, in Amagoro Division of Busia District, Kenya. *African Population Studies*, *9*(1).

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

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

Anderson, B.A., Romani, J.H., Phillips, H.E. and Van Zyl, J.A., 2002. Environment, access to health care, and other factors affecting infant and child survival among the African and coloured populations of South Africa, 1989–94. *Population and Environment*, *23*(4), pp.349-364.

Archer, K.J. and Lemeshow, S., 2006. Goodness-of-fit test for a logistic regression model fitted using survey sample data. *The Stata Journal*, *6*(1), pp.97-105.

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), pp.4450-4464.

Arokiasamy, P., 2002. Gender preference, contraceptive use and fertility in India: Regional and development influences. *International Journal of Population Geography*, *8*(1), pp.49-67.

Barros, F.C. and Victora, C.G., 1990. *Breastfeeding and diarrhea in Brazilian children* (No. 3). New York, NY: Population Council.

Bawah, A.A. and Zuberi, T., 2005. Socioeconomic status and child survival in southern Africa. *Genus*, pp.55-83.

Bello, R.A. and Joseph, A.I., 2014. Determinants of child mortality in Oyo State, Nigeria. *African Research Review*, *8*(1), pp.252-272.

Bhuiya, A. and Streatfield, K., 1991. Mothers' education and survival of female children in a rural area of Bangladesh. *Population studies*, *45*(2), pp.253-264.

Bhutta, Z.A., Chopra, M., Axelson, H., Berman, P., Boerma, T., Bryce, J., Bustreo, F., Cavagnero, E., Cometto, G., Daelmans, B. and de Francisco, A., 2010. Countdown to 2015 decade report (2000–10): taking stock of maternal, newborn, and child survival. *The lancet*, *375*(9730), pp.2032-2044.

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), pp.127-135.

Bradshaw, D. and Nannan, N., 2006. Mortality and morbidity among women and children: maternal, child and women's health. *South African health review*, *2006*(1), pp.127-150.

Bradshaw, D., Groenewald, P., Laubscher, R., Nannan, N., Nojilana, B., Norman, R., Pieterse, D., Schneider, M., Bourne, D.E., Timæus, I.M. and Dorrington, R., 2003. Initial burden of disease estimates for South Africa, 2000. *South African Medical Journal*, *93*(9), pp.682-688.

Buchanan, R., 1975. Breast-feeding-aid to infant health and fertility control. *Population Reports. Series J: Family Planning Programs*, (4), p.19.

Chen, L.C., Huq, E. and d'Souza, S., 1981. Sex bias in the family allocation of food and health care in rural Bangladesh. *Population and development review*, pp.55-70.

Clark, S. and Hamplová, D., 2013. Single motherhood and child mortality in sub-Saharan Africa: A life course perspective. *Demography*, *50*(5), pp.1521-1549.

Cleland, J., 1989. *Maternal education and child survival: further evidence and explanations*. Australian National University.

Cooperation, S., 2017. Stata 15. *Stata Cooperation, College Station, TX*

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

Dean, C.B. and Nielsen, J.D., 2007. Generalized linear mixed models: a review and some extensions. *Lifetime data analysis*, *13*(4), pp.497-512.

Demographic, S.A., Health Survey (SADHS), 2016. *Key indicators report 2016*.

Department of Health, 1998. South Africa demographic and health survey 1998.

Dlamini, W.J., 2016. *Statistical models to understand factors associated with under-five child mortality in Tanzania* (Masters dissertation).

Dobson, A.J. and Barnett, A., 2008. *An introduction to generalized linear models*. Chapman and Hall/CRC.

Elliott, B.A., 1992. Birth order and health: Major issues. *Social science & medicine*, *35*(4), pp.443-452.

Eswaran, M., 2002. The empowerment of women, fertility, and child mortality: Towards a theoretical analysis. *Journal of Population Economics*, *15*(3), pp.433-454.

Ezra, M. and Gurum, E., 2002. Breastfeeding, birth intervals and child survival: analysis of the 1997 community and family survey data in southern Ethiopia. *Ethiopian Journal of Health Development*, *16*(1), pp.41-51.

Fotso, J.C., 2006. Child health inequities in developing countries: differences across urban and rural areas. *International journal for equity in health*, *5*(1), p.9.

Goro, M., 2007. The stalling child mortality: the case of three northern regions. In *The 5th conference of union for Africa population, Tanzania*.

Harrell Jr, F.E., 2015. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer.

Hartzel, J., Agresti, A. and Caffo, B., 2001. Multinomial logit random effects models. *Statistical Modelling*, *1*(2), pp.81-102.

Hastie, T.J. and Tibshirani, R.J., 1986. Generalized additive models (with discussion), Statistical Science.

Hastie, T.J. and Tibshirani, R.J., 1990. Generalized additive models, volume 43. *Monographs on statistics and applied probability*, p.15.

Heaton, T.B. and Amoateng, A.Y., 2007. The family context for racial differences in child mortality in South Africa. *Families and households in post-apartheid South Africa: Socio-demographic perspectives*.

Hedeker, D., 2005. Generalized linear mixed models. *Encyclopedia of statistics in behavioral science*.

Heeringa, S.G., West, B.T. and Berglund, P.A., 2017. *Applied survey data analysis*. Chapman and Hall/CRC.

Hesketh, T. and Xing, Z.W., 2006. Abnormal sex ratios in human populations: causes and consequences. *Proceedings of the National Academy of Sciences*, *103*(36), pp.13271-13275.

Hobcraft, J., 1993. Women's education, child welfare and child survival: a review of the evidence.

Hosmer Jr, D.W., Lemeshow, S. and Sturdivant, R.X., 2013. *Applied logistic regression* the city of publication. John Wiley & Sons.

Hosmer, D.W. and Lemeshow, S., 2000. Applied Logistic Regression the city of publication. John Wiley & Sons. *New York*.

Jiang, J., 2007. *Linear and generalized linear mixed models and their applications*. Springer Science & Business Media.

Jolly, M., Sebire, N., Harris, J., Robinson, S. and Regan, L., 2000. The risks associated with pregnancy in women aged 35 years or older. *Human reproduction*, *15*(11), pp.2433-2437.

Kabir, A., Islam, M.S., Ahmed, M.S. and Barbhuiya, K., 2001. Factors influencing infant and child mortality in Bangladesh. *J Med Sci*, *5*, pp.292-5.

Kembo, J. and Van Ginneken, J.K., 2009. Determinants of infant and child mortality in Zimbabwe: Results of multivariate hazard analysis. *Demographic Research*, *21*, pp.367-384.

Kerber, K.J., Lawn, J.E., Johnson, L.F., Mahy, M., Dorrington, R.E., Phillips, H., Bradshaw, D., Nannan, N., Msemburi, W., Oestergaard, M.Z. and Walker, N.P., 2013. South African child deaths 1990–2011: have HIV services reversed the trend enough to meet Millennium Development Goal 4?. *AIDS (London, England)*, *27*(16), p.2637.

Khodaee, G.H., Khademi, G. and Saeidi, M., 2015. Under-five Mortality in the World (1900-2015). *International Journal of Pediatrics*, *3*(6.1), pp.1093-1095.

Kincaid, C., 2005, April. Guidelines for selecting the covariance structure in mixed model analysis. In *Proceedings of the thirtieth annual SAS users group international conference* (Vol.30, pp. 198-130). SAS Institute Inc Cary NC.

Kish, L., 1965. Sampling organizations and groups of unequal sizes. *American sociological review*, pp.564-572.

Kishor, S., 1993. " May God Give Sons to All": Gender and Child Mortality in India. *American Sociological Review*, pp.247-265.

Kleinbaum, D.G., Dietz, K., Gail, M., Klein, M. and Klein, M., 2002. *Logistic regression*. New York: Springer-Verlag.

Kleinbaum, D.G., Kupper, L., Nizam, A. and Muller, K.E., Applied regression analysis and other multivariable methods. 2008. *Australia: Thomson Brooks/Cole*, *4*.

Kleinman, R. L. (ed.) (1984). "Breastfeeding: Fertility and Contraception." IPPF Medical Publication, England.

Kumar, P., Gemechis, 2010.". *Infant and child mortality in Ethiopia: As statistical analysis approach*.

Kyei, K.A., 2011. Socio-economic factors affecting under five mortality in South Africa-an investigative study. *Journal of Emerging Trends in Economics and Management Sciences*, *2*(2), pp.104-110.

Lagerdien, K., 2005. Reviewing child deaths in South Africa-a rights perspective.

Lee, E.S., Forthofer, R.N. and Lorimer, R.J., 1989. Analyzing complex survey data. Series no. 07-071. *Quantitative applications in the social sciences.*

Littell, R.C., Pendergast, J. and Natarajan, R., 2000. Modelling covariance structure in the analysis of repeated measures data. *Statistics in medicine*, *19*(13), pp.1793-1819.

Liu, H., 2008. Generalized additive model. *Department of Mathematics and Statistics University of Minnesota Duluth: Duluth, MN, USA*.

Liu, Q. and Pierce, D.A., 1994. A note on Gauss—Hermite quadrature. *Biometrika*, *81*(3), pp.624-629.

Mahmood, M.A., 2002. Determinants of neonatal and post-neonatal mortality in Pakistan. *The Pakistan Development Review*, pp.723-744.

Mahy, M., 2003. *Childhood mortality in the developing world: a review of evidence from the Demographic and Health Surveys* (Vol.4). MEASURE DHS+, ORC Macro.

Maluleke, T. and Chola, L., 2015. Millennium Development Goal 4: Reduce child mortality 2015.

MANDA, S.O.M., 1998. Unobserved family and community effects on infant mortality in Malawi. *Genus*, pp.143-164.

Manning, C., 2007. Generalized Linear Mixed Models (illustrated with R on Bresnan et al.'s datives data).

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

Mathers, C., Stevens, G. and Mascarenhas, M., 2009. *Global health risks: mortality and burden of disease attributable to selected major risks*. World Health Organization.

Mathews, S., Martin, L.J., Coetzee, D., Scott, C. and Brijmohun, Y., 2016. Child deaths in South Africa: Lessons from the child death review pilot. *South African Medical Journal*, *106*(9), pp.851-852.

McCullagh, P. and Nelder, J.A., 1989. Binary data. In *Generalized linear models* (pp. 98-148). Springer US.

McCulloch, C., &Searle, S. (2001). Generalized, linear, and mixed models the city of publication. John Wiley & Sons, Inc.

Miller, J. and Haden, P., 2006. Statistical analysis with the general linear model. *Creative Commons Attribution*.

Moeti, A., 2007. *Factors affecting the health status of the people of Lesotho* (Masters dissertation).

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

Mosley, W.H. and Chen, L.C., 1984. An analytical framework for the study of child survival in developing countries. *Population and development review*, *10*(0), pp.25-45.

Motshwaedi, L.M.F., 2011. *An Analysis of Causes of Child Mortality in South Africa: 1997-2006* (Doctoral dissertation, North-West University (Mafikeng Campus)).

Mturi, A.J. and Curtis, S.L., 1995. The determinants of infant and child mortality in Tanzania. *Health policy and planning*, *10*(4), pp.384-394.

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), pp.1934-722.

Nannan N, Dorrington RE, Laubscher R, Zinyakatira N, Prinsloo M, Dirakwa TB, Matzopoulos R, Bradshaw D. Under-5 mortality statistics in South Africa: Shedding some light on the trends and causes 1997-2007. Cape Town: South African Medical Research Council, 2012.

National Department of Health (NDoH), Statistics South Africa (Stats SA), South African Medical Research Council (SAMRC), and ICF, 2017. South Africa Demographic and Health Survey 2016: key indicators. https://www.dhsprogram.com/data/dataset_admin

Nations, 2000. http://www.un.org/millennium/ (last accessed 28 July 2018)

Nelder, J.A. and Wedderburn, R.W., 1972. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135*(3), pp.370-384.

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

Pande, R.P., 2003. Selective gender differences in childhood nutrition and immunization in rural India: the role of siblings. *Demography*, *40*(3), pp.395-418.

Pinheiro, J.C. and Bates, D.M., 1995. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of computational and Graphical Statistics*, *4*(1), pp.12-35.

Pojda, J. and Kelley, L., 2000. Low Birth weight-Nutrition policy discussion paper No. 18. *United Nations Administrative Committee on Coordination*

Rademeyer, S., 2017. *Provincial differentials in under-five mortality in South Africa* (Doctoral dissertation).

Raji, O.O., 2010. *Examining mother's related socioeconomic and demographic determinants of infant and child mortality in the Eastern Cape, South Africa* (Masters of Philosophy).

Roberts, D., 2014. *Prevalence and risk factors of malaria in children under the age of five years old in Uganda* (Masters dissertation).

Ronsmans, C., Graham, W.J. and Lancet Maternal Survival Series steering group, 2006. Maternal mortality: who, when, where, and why. *The lancet*, *368*(9542), pp.1189-1200.

Sahn, D.E. and Stifel, D.C., 2003. Urban–rural inequality in living standards in Africa. *Journal of African Economies*, *12*(4), pp.564-597.

SAS Institute, 2015. *Base SAS 9.4 procedure guide*. SAS Institute.

Sastry, N., 1996. Community characteristics, individual and household attributes, and child survival in Brazil. *Demography*, *33*(2), pp.211-229.

Schwarz, G., 1978. Estimating the dimension of a model. *The annals of statistics*, *6*(2), pp.461-464.

Setel, P.W., Macfarlane, S.B., Szreter, S., Mikkelsen, L., Jha, P., Stout, S., AbouZahr, C. and Monitoring of Vital Events (MoVE) writing group, 2007. A scandal of invisibility: making everyone count by counting everyone. *The Lancet*, *370*(9598), pp.1569-1577.

Shackman, G., 2001. Sample size and design effect. presented at the Albany Chapter of the American Statistical Association. *Albany, NY: New York State Department of Health*.

Shehzad, S., 2006. The determinants of child health in Pakistan: an economic analysis. *Social indicators research*, *78*(3), pp.531-556.

Siller, A.B. and Tompkins, L., 2006, March. 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* (pp. 26-29). SAS Institute Inc.

Šimundić, A.M., 2008. Measures of diagnostic accuracy: basic definitions. *Medical and biological sciences*, *22*(4), pp.61-65.

Song, X.Y. and Lee, S.Y., 2006. Model comparison of generalized linear mixed models. *Statistics in medicine*, *25*(10), pp.1685-1698.

SPSS, I., 2012. Statistical package for the social sciences. *Data analysis software packages*. *Version*, 21.

Stone, M., 1979. Comments on model selection criteria of Akaike and Schwarz. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp.276-278.

Suparmi, S., Chiera, B. and Pradono, J., 2016. Low birth weights and risk of neonatal mortality in Indonesia. *Health Science Journal of Indonesia*, *7*(2), pp.113-117.

Twum-Baah, K.A., Nyarko, P.E., Quashie, S.E., Caiquo, I.B. and Amuah, E., 1994. Infant, child and maternal mortality study in Ghana.

UNICEF, 2007. Young lives: Statistical data on the status of children aged 0–4 in South Africa.

UNICEF, 2011. WHO, The World Bank, the United Nations Population Division. 2011. *Levels and trends in child mortality, report*.

UNICEF, 2015. UN inter-agency group for child mortality estimation. *Levels and trends in child mortality*.

UNICEF, W., 2018. Capture the Moment-Early initiation of breastfeeding: The best start for every newborn. *New York: UNICEF.*

UNICEF., 2009. *State of the world's children: Celebrating 20 years of the convention on the rights of the child*. Unicef.

United Nations, 2003. Indicators for monitoring the millennium development goals. United Nations, New York.

Van de Poel, E., O'Donnell, O. and Van Doorslaer, E., 2007. Are urban children really healthier? Evidence from 47 developing countries. *Social science & medicine*, *65*(10), pp.1986-2003.

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.

Wagstaff, A., 2000. Socioeconomic inequalities in child mortality: comparisons across nine developing countries. *Bulletin of the World Health Organization*, *78*, pp.19-29.

Wambugu, M.R., Determinants of under-5 mortality in Kenya during upsurge and declining trends period.

Woldemicael, G., 2001. Diarrhoeal morbidity among young children in Eritrea: environmental and socioeconomic determinants. *Journal of health, population and nutrition*, pp.83-90.

Wolfinger, R. and O'connell, M., 1993. Generalized linear mixed models a pseudo-likelihood approach. *Journal of statistical Computation and Simulation*, *48*(3-4), pp.233-243.

Wood, S.N., 2006. On confidence intervals for generalized additive models based on penalized regression splines. *Australian & New Zealand Journal of Statistics*, *48*(4), pp.445-464.

World Health Organization and UNICEF, 2015. Trends in maternal mortality: 1990-2015: estimates from WHO, UNICEF, UNFPA, World Bank Group and the United Nations Population Division.

World Health Organization, 2016. *World health statistics 2016: monitoring health for the SDGs sustainable development goals*. World Health Organization.

Wray, J.D., 1978. Maternal nutrition, breast-feeding and infant survival. In *Nutrition and human reproduction* (pp. 197-229). Springer, Boston, MA.

Yach, D., 1994. Health status and its determinants in South Africa. *Africa health*, (Spec No), pp.5-8.

Yee, T.W. and Mitchell, N.D., 1991. Generalized additive models in plant ecology. *Journal of vegetation science*, *2*(5), pp.587-602.

Yirga, A.A., Melesse, S.F., Ayele, D.G. and Mwambi, H., 2019. The use of Complex Survey Design Models to Identify Determinants of Malnutrition in Ethiopia. *J Hum Ecol*, *66*(1-3), pp.1-11.

Zhang, H., Lu, N., Feng, C., Thurston, S.W., Xia, Y., Zhu, L. and Tu, X.M., 2011. On fitting generalized linear mixed-effects models for binary responses using different statistical packages. *Statistics in Medicine*, *30*(20), pp.2562-2572

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

Appendix A

SAS and STATA codes

A.1 Logistic regression SAS code

PROC LOGISTIC fits the linear logistic regression for dichotomous response variable assuming that it was obtained from a simple random sample using maximum likelihood.

```
proc logistic data = Cm plots=all;
class Mother_Age(ref="20-35 years") Birth_Size (ref="Average") NoOfChildrenU5 (ref="Less than 2 children") ChildrenEverBorn (ref="Less than 2 children") Breastfeeding (ref="No") Bord (ref "2-3 births") MaritalStatus (ref="Never Married") Ethnicity (ref="Black African") Wealth_Index (ref="Poor") WaterSouce (ref="Safe water") Province (ref="KwaZulu-Natal") / param = ref;
model Child_Dead (event='Yes') = Mother_Age Birth_Size NoOfChildrenU5 ChildrenEverBorn Breastfeeding Bord MaritalStatus Ethnicity Wealth_Index WaterSouce Province / lackfit;
run;
```

A.2 Survey Logistic Regression using STATA Code

SVY command in Stata fits logistic regression model for binary response survey data using maximum likelihood method. This command incorporates survey design such as clustering, stratification and sample weight.

```
svyset ClusterNo [pweight=SampleWeight], strata (Stratification)

logit Child_Dead i.Birth_Size i.NoOfChildrenU5 i.ChildrenEverBorn i.Breastfeeding i.Bord i.MaritalStatus i.Ethnicity i.Wealth_Index i.WaterSource i.Province [pweight = SampleWeight],robust

logistic Child_Dead i.Birth_Size i.NoOfChildrenU5 i.ChildrenEverBorn i.Breastfeeding i.Bord i.MaritalStatus i.Ethnicity i.Wealth_Index i.WaterSource i.Province [pweight = SampleWeight],robust

estat ic

estat effects, deff deft
```

A.3 Generalized Linear Mixed Model SAS Code

PROC GLIMMIX fits the generalized linear mixed models to the data. This procedure allows random statement to specify the random effect to be incorporated in the model.

```
proc glimmix data = Cm method laplace;
class Mother_Age Birth_Size NoOfChildrenU5 ChildrenEverBorn Breastfeeding Bord MaritalStatus Ethnicity Wealth_Index WaterSouce Province ClusterNo;
model Child_Dead (event='Yes') = Mother_Age Birth_Size NoOfChildrenU5 ChildrenEverBorn Breastfeeding Bord MaritalStatus Ethnicity Wealth_Index WaterSouce Province / link=logit DIST= binary ODDSRATIO Solution;
LSMEANS Mother_Age Birth_Size NoOfChildrenU5 ChildrenEverBorn Breastfeeding Bord MaritalStatus Ethnicity Wealth_Index WaterSouce Province/PLOT=DIFFPLOT ADJUST=TUKEY ALPHA=0.05;
random intercept / intercept / subject=ClusterNo;
covtest zerog;
run;
```

```
proc glimmix data = Cm method = Quard empirical=classical;
class ClusterNo Mother_Age Birth_Size NoOfChilidrenU5 ChildrenEverBorn Breastfeeding Bord MaritalStatus Ethnicity Wealth_Index WaterSource Province;
model Child_Dead (event='Yes') Mother_Age Birth_Size NoOfChilidrenU5 ChildrenEverBorn Breastfeeding Bord MaritalStatus Ethnicity Wealth_Index WaterSource Province/link=logit DIST=binary ODDSRATIOS Solution;
LSMEANS Mother_Age Birth_Size NoOfChilidrenU5 ChildrenEverBorn Breastfeeding Bord MaritalStatus Ethnicity Wealth_Index WaterSource Province/PLOT=DIFFPLOT ADJUST=TUKEY ALPHA=0.05;
RANDOM INT/SUBJECT = ClusterNo weight=SampleWeight;
run;
```

A.4 Generalized Additive Model SAS Code

PROC GAM fits a logistic additive model with binary response variable child survival status and covariates.

```
proc gam data = Cm plots=components (clm commonaxes);
class Birth_Size Breastfeeding MaritalStatus Ethnicity Wealth-Index WaterSource Province;
model Child_Dead(event='Yes') = spline(Mother_Age) spline(NoOfChildrenU5) spline ChildrenEverBorn) spline(Bord) param(Birth_Size) param(Breastfeeding) param(MaritalStatus) param(Ethnicity) param(Wealth_Index) param(WaterSource) param(Province)/DIST=binomial;
score data=Cm out=Cm;
run;
```

Where:

Cm= Child mortality data

Child_dead = Child survival status

Mother_Age = Mother's age

Birth size = Size of child at birth

Breastfeeding = Currently breastfeeding

NoOfChildrenU5 = Number of children 5 and under

ChildrenEverBorn = Total number of children ever born

Bord = Birth order number

MaritalStatus = Current marital status

Ethnicity = Ethnicity

Wealth_index = Wealth index

WaterSource = Source of drinking water

Province = Province

ClusterNo = clustering

stratification = Stratification

SampleWeight = Sample weight.

Appendix B

Additional Results

B.1 Generalized Linear Mixed Model Results

Table 4.6: Gauss-Hermite Quadrature, estimated coefficients, odds ratios, standard errors, p-values and confidence interval.

Analysis of Maximum Likelihood Estimates and Odds Ratio estimates						
Demographic characteristics						
Parameter	**Estimate**	**Standard Error**	**P-Value**	**Odds ratio**	**95%Confidence interval Lower**	**Upper**
Intercept	-21.0488	688.74	0.9756			
Mother's age (ref >35 years)						
<20 years	-0.1806	0.5199	0.7284	0.835	0.301	2.313
20 – 35 years	0.2758	1.2570	0.8264	1.318	0.112	15.480
Size of child at birth (ref. Very large)						
Very small	1.1616	0.6733	0.0845	3.195	0.854	11.957
Average	-0.1838	0.4753	0.6990	0.832	0.328	2.112
Currently breastfeeding (ref. Yes)						
No	1.0083	0.4538	0.0263	2.741	1.126	6.672
Birth order number (ref. First birth)						
2 – 3 births	-0.2914	0.5205	0.5755	0.747	0.269	2.072
More than 3 births	-0.0286	0.4630	0.9508	0.972	0.362	2.408
Current marital status (ref. never married)						
Married	-0.2183	0.3691	0.5542	0.804	0.390	1.657
Living with partner	0.0509	0.2800	0.8558	1.052	0.608	1.822
Ethnicity (ref. Others)						
Black African	1.2465	0.688	0.9812	3.478	0.641	1.522
Coloured	1.4522	0.697	0.9844	4.273	0.724	1.731
Socio-Economic characteristics						
Wealth index (ref. Rich)	.					
Middle	0.1800	0.6164	0.7703	1.197	0.358	4.007
Poor	0.7355	0.5519	0.1826	2.087	0.707	6.155
Number of children 5 and under (ref.< 2 Children)						
2 or more children	-1.0539	0.4868	0.0304	0.349	0.134	0.905
Total children ever born (ref. less than 2 Children)						
2 or more children	0.7636	0.5065	0.1317	2.146	0.795	5.792
Household Environment characteristics						
Source of drinking water (ref. safe water)						
Not safe water	0.1281	0.0969	0.1862	1.137	0.940	1.374
Province (ref. Western Cape)						
Eastern Cape	-0.1276	0.6706	0.8491	0.880	0.236	3.276
Free State	0.2520	0.7029	0.7200	1.287	0.324	5.102
Gauteng	-0.1773	0.7891	0.8223	0.838	0.178	3.933
KwaZulu-Natal	-0.5060	0.6593	0.4428	0.603	0.166	2.195
Limpopo	0.8416	0.6910	0.2233	0.431	0.111	1.670
Mpumalanga	0.3094	0.6146	0.6146	1.363	0.409	4.545
North West	-0.1347	0.6164	0.8270	0.874	0.261	2.925
Northern Cape	0.2838	0.7048	0.6871	1.328	0.334	5.287

Table 4.7: Penalized Quasi-Likelihood, estimated coefficients, odds ratios, standard errors, p-values and confidence interval.

Analysis of Maximum Likelihood Estimates and Odds Ratio estimates						
Demographic characteristics						
Parameter	**Estimate**	**Standard Error**	**P-Value**	**Odds ratio**	**95%Confidence interval Lower**	**Upper**
Intercept	3.2896	0.62	<0.0001	26.832	2.069e+00	4.511e+00
Mother's age (ref <20 years)						
20 – 35 years	-0.1425	0.17	0.4123	0.867	-4.819e-01	1.970e-00
>35 years	-0.0092	0.18	0.9603	0.991	-3.700e-01	3.516e-01
Size of child at birth (ref. Very small)						
Average	1.2380	0.14	<0.0001	3.449	9.681e-01	1.508e+00
Very large	1.1228	0.16	<0.0001	3.073	8.047e-01	1.441e+00
Currently breastfeeding (ref. Yes)						
No	1.1030	0.19	<0.0001	3.013	7.394e-01	1.467e+00
Birth order number (ref. First birth)						
2 – 3 births	0.2598	0.16	0.1145	1.297	-6.172e-02	5.814e-01
More than 3 births	-0.2134	0.24	0.3680	0.808	-6.766e-01	2..497e-01
Current marital status (ref. never married)						
Married	0.4539	0.18	0.0114	1.574	1.035e-01	8.043e-01
Living with partner	-0.3666	0.16	0.0193	0.693	-6.727e-01	-6.053e-02
Ethnicity (ref. Black African)						
Coloured	1.5798	0.59	0.0070	4.854	4.353e-01	2.724e+00
Others	23.488	33759	0.9994	1.587e+10	-6.594e+0.4	6.599e+04
Socio-Economic characteristics						
Wealth index (ref. Poor)						
Middle	0.2310	0.17	0.1824	1.260	-1.074e-01	5.694e-01
Rich	0.3765	0.22	0.0875	1.457	-5.393e-02	8.069e-01
Number of children 5 and under (ref.< 2 Children)						
2 or more children	1.2199	0.29	<0.0001	3.387	6.600e-01	1.780e+00
Total children ever born (ref. less than 2 Children)						
2 or more children	-0.5686	0.18	0.0017	0.566	-9.232e-01	-2.141e-01
Household Environment characteristics						
Source of drinking water (ref. safe water)						
Not safe water	-0.6088	0.16	0.0002	0.544	-9.273e-01	-2.903e-01
Province (ref. Eastern Cape)						
Free State	-0.1677	0.66	0.7983	0.846	-1.451e+00	1.116e+00
Gauteng	-0.4995	0.69	0.4714	0.607	-1.856e+00	8.568e-01
KwaZulu-Natal	-0.7809	0.66	0.2378	0.458	-2.074e+00	5.122e-09
Limpopo	-0.1787	0.65	0.7835	0.836	-1.451e+00	1.092e+00
Mpumalanga	-0.4504	0.65	0.4874	0.637	-1.719e+00	8.180e-01
North West	-0.0384	0.67	0.9546	0.962	-1.357e+00	1.280e+00
Northern Cape	-0.8037	0.63	0.2056	0.448	-2.045e+00	4.375e-01
Western Cape	0.5076	0.67	0.4507	1.661	-8.085e-01	1.824e+00

Appendix C

Derivation of some properties of the Exponential Family

B.1 Properties of the Exponential Family

Here we get the general expression for the mean and the variance of the exponential distribution in term of a,b,∅.

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

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

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

Differentiating both sides with respect to θ we get

$$\frac{\partial}{\partial\theta}\left[\int exp\left\{\frac{y\theta - b(\theta)}{a(\emptyset)} + C(y,\emptyset)\right\}dy\right] = 0$$

$$\int \frac{\partial}{\partial\theta} exp\left\{\frac{y\theta - b(\theta)}{a(\emptyset)} + C(y,\emptyset)\right\}dy = 0,$$

$$\int \left[\frac{y - b'(\theta)}{a(\emptyset)}\right] f(y,\theta,\emptyset)dy = 0,$$

$$\int \frac{yf(y,\theta,\emptyset)}{a(\emptyset)}dy - \int \frac{b'(\theta)f(y,\theta,\emptyset)}{a(\emptyset)} = 0,$$

$$\int \frac{yf(y,\theta,\emptyset)}{a(\emptyset)} = \int \frac{b'(\theta)f(y,\theta,\emptyset)}{a(\emptyset)}dy,$$

$$\int yf(y,\theta,\emptyset)dy = b'(\theta)\int f(y,\theta,\emptyset)dy,$$

$$E(y) = b'(\theta) \times 1, since \int f(y, \theta, \emptyset)dy = 1,$$

$$E(y) = b'(\theta)\ is\ the\ mean\ of\ y.$$

Taking the second derivative with respect to θ we get,

$$\int \left[\frac{y - b'(\theta)}{a(\emptyset)}\right] f(y, \theta, \emptyset)dy = 0,$$

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

$$\int \left[\frac{y - b'(\theta)}{a(\emptyset)}\right]^2 f(y, \theta, \emptyset)dy = \frac{b''(\theta)}{a(\emptyset)} \int f(y, \theta, \emptyset)dy,$$

$$\frac{1}{a(\emptyset)^2} \int [y - b'(\theta)]^2 f(y, \theta, \emptyset)dy = \frac{b''(\theta)}{a(\emptyset)},$$

$$\frac{Var(y)}{a(\emptyset)^2} = \frac{b''(\theta)}{a(\emptyset)},$$

$$Var(y) = a(\emptyset)b''(\theta).$$

C.2 Sampling distribution of the Maximum Likelihood Estimator (MLE)

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

$$f(x) = f(a) + (x - a)f'(a) + \frac{1}{2}(x - a)^2 f''(a) + \frac{1}{3}(x - a)^3 f'''(a) + \cdots$$

$$\approx f(a) + (x - a)f'(a) \qquad (C.1)$$

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

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

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

$$I^{-1}U(\beta) \approx (\hat{\beta} - \beta) \tag{C. 2}$$

Taking the expected value in Eq. (C.2) we get

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

This implies that $E(\beta) = \beta$, so β is the consistent estimator of β. The variance is therefore given by

$$\begin{aligned} Var(\beta) &= E\left[(\hat{\beta} - \beta)'\right] \\ &= E\left[(I^{-1}U(\beta))(I^{-1}U(\beta))'\right] \\ &= I^{-1}E[(U(\beta))(U(\beta))']I^{-1} \\ &= I^{-1}Var(U(\beta))I^{-1} \\ &= I^{-1}II^{-1} \\ &= I^{-1} \end{aligned} \tag{C. 3}$$

Therefore $\beta \sim MVN(\beta, I^{-1})$ then we have this

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

which is known as the Wald statistics.