Flexible statistical modelling of the determinants of childhood anaemia in Tanzania and Angola

UNIVERSITY OF
KWAZULU - NATAL

INYUVESI
YAKWAZULU-NATALI

Qondeni Ndlangamandla (214520530)

, 2020

Flexible statistical modelling of the determinants of childhood anaemia in Tanzania and Angola

by

Qondeni Ndlangamandla (214520530)

A thesis submitted to the
University of KwaZulu-Natal
in fulfilment of the requirements for the degree
of
MASTER OF SCIENCE
in
STATISTICS

Thesis Supervisor: Prof Shaun Ramroop

Thesis Co-supervisor: Prof Henry Mwambi

UNIVERSITY OF KWAZULU-NATAL
SCHOOL OF MATHEMATICS, STATISTICS AND COMPUTER SCIENCE
PIETERMARITZBURG CAMPUS, SOUTH AFRICA

Declaration - Plagiarism

I, Qondeni Ndlangamandla (214520530) , declare that

1. The research reported in this thesis, except where otherwise indicated, is my original research.
2. This thesis has not been submitted for any degree or examination at any other university.
3. This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowlegded as being sourced from other persons.
4. This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then
   (a) their words have been re-written but the general information attributed to them has been referenced, or
   (b) where their exact words have been used, then their writing has been placed in italics and referenced.
5. 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 reference sections.

08 November 2020

Qondeni Ndlangamandla (214520530) (Student) Date

9 November 2020

Prof Shaun Ramroop (Supervisor) Date

9 November 2020

Prof Henry Mwambi (Co-supervisor) Date

Disclaimer

This document describes work undertaken as a Masters programme of study at the University of KwaZulu-Natal (UKZN). All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the institution.

Abstract

Anaemia is one of the major causes of morbidity and mortality in children aged five or less in Africa, affecting 25% of the world's population. In developing countries, it accounts for more than 89% of the disease burden. Although anaemia affects all population groups, the more vulnerable groups are children under five years of age and women of reproductive age (15–49 years) compared to any other age group. According to the World Health Organization's 2008 report, 50% of anaemia cases in Africa were associated with insufficient consumption of iron (iron deficiency anaemia). This study aims to determine the factors associated with childhood anaemia in Tanzania and Angola.

For us to serve our aim, the Tanzania Demographics and Health Survey (TDHS) and the Angola Demographics and Health Survey (ADHS) data sets were fitted to several statistical models that could robustly model the response variable, anaemia, which is binary. Survey Logistic Regression (SLR), which is under the class of Generalized Linear Models (GLM), fits because of its robustness, not only in modelling dichotomous responses, but also in it ability to deal with data that assumes complex survey designs. The SLR model was extended by a Generalized additive mixed model (GAMM), which was fitted to relax the assumption of normality and to fit other terms non-parametrically. Furthermore, to cater for the effect of spatial effect and spatial variability, a Spatial Generalized linear mixed model (SGLMM) was fitted to the two data sets to help in the investigation of factors that are spatially related to childhood anaemia. The SLR and SGLMM models were fitted using the SAS software (PROC SURVEYLOGISTIC and PROC GLIMMIX, respectively), while the GAMM model was fitted using the statistical-R software. Moreover, smooth maps were produced for the outcome variable using ARCGIS software for the purpose of identifying the hot spots of childhood anaemia in the country.

Our aim for this study was successfully achieved. After the three models were fitted into the two data sets, they revealed that the factors that were highly associated with childhood anaemia in both countries are: the highest level of education of caretak-

ers (mothers), child gender, age of the child and stunting status. The models also revealed that the standard of living in Tanzania has a significant effect in childhood anaemia

keywords:Childhood anaemia, survey logistic regression (SLR), generalized additive mixed models (GAMM), spatial generalized linear mixed models (SGLMM), smoothing, demographic and health survey (DHS), adjusted odds ratios (aOR)

Acknowledgements

This research included in this thesis could not have been possible without the assistance of many individuals.

First and foremost, I would like to extend my deepest gratitude to my supervisor, Prof. Shaun Ramroop, for his inspired guidance, valuable suggestions, and constant encouragement and support throughout the entire period of my research study. My sincere and deepest gratitude goes to my co-supervisor, Prof Henry Mwambi, for helping me to get funding for my masters, for his unending support, valuable suggestions and great patience. I would like to extend my heartfelt acknowledgement to Dr N. Yende, head of the statistics department at CAPRISA, and the rest of the team at the stats department at CAPRISA for mentorship and for the opportunity to attend the SUSAN conference held in Cape Town in September 2019. I convey a special thanks to Dr Faustin Habyarimanamy for always being there for me if ever I needed help. I would further like to extend my deepest gratitude to my fellow friends Sakhile Mnguni and Mohanad Mohammed and to my siblings Mthembeni, Khethiwe, Cebisile, Bongekile, Menziwa and Mcebiseni for all their support and encouragement throughout the period of this thesis.

I express my deep sense of love to my mother, Zombi Gladys Dlamini, and father, Mandlenkosi Albert Ndlangamandla, for the moral and unending support, for believing in me, for their love and constant encouragement. Without you and your blessings, this work wouldn't have been possible.

Finally, All thanks to God the Almighty for giving this opportunity of a lifetime.

Note

A poster from this thesis was presented at the IBS-SUSAN-SSACAB conference held in Cape Town, South Africa, from 8 September 2019 to 12 September 2019. The presentation was specifically based on survey logistic results, titled: ***'Modelling determinants of childhood anaemia in Tanzania and Angola'***. The presented poster was nominated as the best among all those presented at the conference.

Contents

List of Figures

List of Tables

Abbreviations

Hb	Hemoglobin
TDHS	Tanzania demographics and health survey
ADHS	Angola demographics and health survey
WHO	World health organisation
OR	Crude or unadjusted odds ratios
aOR	Adjusted odds ratio
LM	Linear models
OLS	Ordinary least squares
GLM	Generalized linear models
LR	Logistic regression
SLR	Survey logistic regression
GAMM	Generalized linear mixed models
GAM	Generalized additive models
GAMM	Generalized additive mixed models
ML	Maximum likelihood
REML	Restricted maximum likelihood
PQL	Penalized quasi likelihood
D-PQL	Double penalized quasi likelihood
CI	Confidence interval
SPSS	Statistical package for social science
SGLMM	Spatially generalized linear mixed models

Chapter 1

Introduction

1.1 Background

Anaemia is one of the major causes of morbidity and mortality in children in Africa aged five or less, affecting 25% of the world's population (McLean et al., 2009). It accounts for more than 89% of the disease burden in developing countries (Kassebaum, 2016). Although the serious health problem of childhood anaemia affects both developing and well-developed countries, it is more prevalent in developing countries (Ewusie et al., 2014; Balarajan et al., 2011a). In general, anaemia is defined as the condition where the body does not have a sufficient haemoglobin (Hb) level to provide enough oxygen into the body tissues (McLean et al., 2009). In pregnant women shortage of Hb in the body is highly associated with increased risk of maternal and perinatal mortality and small size or weight of the child at birth (Young, 2018; Stevens et al., 2013).

Maternal anaemia is amongst one of the common causes of morbidity and mortality in both the mother and her baby; it also increases the rate of miscarriages, stillbirths, prematurity, and low birth weight (Young, 2018; Peña-Rosas & García-Casal, 2014). The World Health Organization's 2011 report highlighted that maternal and neonatal death ranges from 2.5 million to 3.4 million per year (Stevens et al., 2013). Childhood anaemia can adversely affect cognitive, and motor development and results in fatigue and low productivity (Stevens et al., 2013; Crawley, 2004), which may persist even after treatment (Ewusie et al., 2014). According to health specialists anaemia is caused by the imbalance production of erythrocytes and the removal of erythrocyte (Allali et al., 2017). Furthermore, anaemia is the indicator of both poor nutrition and poor health (Peña-Rosas & García-Casal, 2014).

Many Sub-Saharan African countries and Southeast Asian countries are at the mercy

of malnutrition and, consequentially, childhood anaemia is more prevalent in these regions than in any other parts in the world (Peña-Rosas & García-Casal, 2014). Anaemia affects all population groups but women of productive age and children under five years of age (children aged 0-59 months) are identified as the most vulnerable among other population group in many studies (Kassebaum, 2016; Crawley, 2004; Sharmanov, 1998; Osungbade & Oladunjoye, 2012; McLean et al., 2009).
The prevalence of childhood anaemia decreases with age; children aged 6-24 months are more likely to suffer from anaemia (Ngwira & Kazembe, 2015; Sanou & Ngnie-Teta, 2012). The highest prevalence of childhood anaemia is in sub-Saharan African countries where 67% of children under five years suffer from the illness and in South East Asia, where 65.5% of young children have anaemia (Habyarimana et al., 2017). Also, the 2011 global report showed that about 9.6 million children were found to be severely anaemic worldwide, with a 95% confidence interval of (6.9 million to 14.1 million) (Organization et al., 2015).

A total of 4.9 million of children under five years of age (0 months to 59 months) were from the African region. The study shows that only 0.18 million children were from the American zone, 2.7 million were from South East Asia, 0.2 million were from the European regions, 1.5 million were from the Eastern Mediterranean region, and only 0.2 million out of 9.6 million children were from the Western Pacific regions (Organization et al., 2015). Hence, more than half of the children who were found to be severely anaemic were from the African region followed by the East Asian regions. Thus this report serves as a proof that anaemia is more prevalent in these two regions compared to other regions in the world. In both regions, under-nutrition is one of the common health problems that contributes to mortality and morbidity of children under five, and thus to childhood anaemia. In addition, about 90% of the people making up the population of Africa are black, and several studies have shown that the distribution of population haemoglobin is lower in blacks than in whites (Johnson-Spear & Yip, 1994; Dallman et al., 1978; Balarajan et al., 2011b), contributing to the high prevalence of anaemia in Africa.

The causes of anaemia are multifactorial by nature and they are interrelated in a complex way (Peña-Rosas & García-Casal, 2014). The most common cause in young children is low consumption and absorption of iron-rich foods (i.e, meat and meat products). Iron-deficiency accounts for approximately 50% of the cases of anaemia globally (Habyarimana et al., 2017; Ngwira & Kazembe, 2015; Peña-Rosas & García-Casal, 2014). In most of these cases, young children and pregnant and post-partum women are the most commonly and severely affected because of the high iron demands for infant growth and for pregnancy; young children also need iron for growth

and development (Kassebaum, 2016; Milman, 2011). In every human body, various nutritional deficiencies and different infections may play a role, but iron is the most vitally important mineral and an important component of metalloproteins required during oxygen transportation and during the process of metabolism (Milman, 2011). A well-nourished human body contains about three to four grams of iron and the haemoglobin protein contains oxygen of approximately two-thirds of the protein (Milman, 2011; Balarajan et al., 2011b). According to WHO, iron deficiency anaemia is considered as a public health problem when the prevalence of Hb concentration is more than 5% of the population. Iron deficiency is most common in low-income and middle-income countries (Milman, 2011), where the prevalence of anaemia is thus high. Parasitic infections are considered to be the second most common cause of anaemia. Studies in East Africa have shown that Plasmodium falciparum malaria and iron deficiency account for much to childhood anaemia (Newton et al., 1997).

According to global records between one million and three million deaths, every year in Africa are the result of malaria (Stevens et al., 2013). The species Plasmodium falciparum is the most pathogenic and can lead to severe anaemia and cognitive heart failure (Balarajan et al., 2011b). The mechanism of malaria-related anaemia has evolved substantially and can be broadly characterized by a both decrease and increase of erythrocyte destruction. Malaria infection during pregnancy can result in reduced birth weights and contributes to maternal infant mortality and poor foetal development (Murphy & Breman, 2001). Many other causes of anaemia have been studied that include genetic haemoglobin disorder, chronic thalassemia's, haemoglobin variants in ovalocytes, among others. Nutritional factors include deficiency in iron, folic acid , vitamin A , Vitamin B12, and protein-energy malnutrition.

Folic acid is needed during the synthesis and maturation of erythrocytes. A deficiency often results in megaloblastic anaemia, a condition characterized by cells with large and ill-shaped nuclei resulting from impaired DNA synthesis. During pregnancy, folate demands increase, and women entering pregnancy with insufficient folate usually develop megaloblastic anaemia (Kapil & Kapil). Vitamin A takes part in the creation of red blood cells improves haemoglobin concentration and it improves the efficacy of iron supplementation moreover vitamin A reduces susceptibility to infections. Vitamin A deficiency is more common in South East Asian region, affecting approximately 21% of children under five years of age and 6% of pregnant women (Kapil & Kapil).
Infectious diseases that are associated with anaemia include soil-transmitted helminth infestation, malaria, tuberculosis, human immunodefincieny virus, and AIDS. Furthermore, anaemia in young children results not only from childhood events (Vita-

min A,B12 deficiency; protein energy malnutrition; etc), but also from maternal iron deficiency and maternal anaemia, which are associated with impaired fetal development and iron-deficient and anaemic infants (de Savigny et al., 2003).

Anaemia is a common known risk factor that causes death at both the mild and moderate levels (Stoltzfus et al., 2004), but the severe anaemic level is the most common cause of morbidity and mortality in African children (Calis et al., 2016; Koram et al., 2000). According to WHO disease classification, severe anaemia is considered as a health problem in the population when its prevalence exceeds 40% (Peña-Rosas & García-Casal, 2014; Hall et al., 2001). Furthermore, the Severe anaemic stage causes fragility and distortion of bones in young children (Vogiatzi et al., 2009). According to the established level of anaemia by the WHO, the severe anaemic stage is when the haemoglobin concentration is less than seven grams per decilitre (Hb$\leq$7g/dl) in children (Stevens et al., 2013). Although the whole world has been playing a major role in reducing anaemia by iron supplementation, fortification and diversification of the diet, anaemia remains a major health problem. In Tanzania, the prevalence of the condition is high but the country has made tremendous progress in fighting and reducing it.

Based on the most recent statistics, from the 2015-16 Tanzania Demographic and Health Surveys (TDHS) only 58% of children under five years of age were found to be anaemic (haemoglobin less than 11 g/dl). Twenty-seven percent were found to be mildly anaemic, 30% of children were moderately anaemic and only 2% were found to be severely anaemic. The prevalence of childhood anaemia in Tanzania declined substantially between 2004 and 2005 and in 2012 it went from 72% to 59%. In sharp contrast, there was a small decrease in anaemia among children between 2010 and 2015-2016 from 59% to 58% (Ministry of Health et al., 2016). The prevalence of childhood anaemia in Angola was reported at 71.70% in 1990, then the record decreased to 51,30% in 2015, the percentagee was updated annually and it was recorded as low in 2016 at 50.90%, averaging to 63% as the data was updated every year from 1990 to 2016 (Worldbank, 2016). In Rwanda, the prevalence of childhood anaemia reduced from 52% to 38% between the years 2005 and 2010, and there was a negligible decrease in that age group, 38% to 37% between the year 2010 and 2014-2015 (Habyarimana et al., 2017). Thus anaemia prevalence is still high in some parts of the African continent.

According to the WHO, the established anaemia levels vary according to individual age (Stevens et al., 2013; McLean et al., 2009). Childhood anaemia is defined as

haemoglobin less than 11g/dl (McLean et al., 2009). For example, children that are 6-59 months of age are considered to be non-anaemic if their haemoglobin concentration is greater than 11 grams per decilitre(Hb$\geq$ 11 g/dl); mild anaemic if their haemoglobin concentration lies in the interval of 10.0 g/dl $\leq$Hb$\leq$ 10.9 g/dl, moderately anaemic if 7.0 g/dl $\leq$Hb$\leq$ 9.9 g/dl and severely anaemic if their haemoglobin concentration is less than 7.0 g/dl, while children 5 years and 11.99 years are said to be anaemic if their haemoglobin concentration is less than 115 g/l which is the same as 11.5 g/dl. Children between 12 and 14.99 years and non-pregnant women aged 15 or more are both considered to be anaemic if their haemoglobin threshold is less than 120g/l or 12.0 g/dl. Pregnant women are said to be anaemic when their haemoglobin concentration is less than 110 grams per decilitre (110 g/l or 11.0 g/dl) and men aged 15 or more are considered to be anaemic when their haemoglobin threshold is less than 130g/l (13.0 g/dl) according to the WHO established cut off levels (De Benoist et al., 2008; McLean et al., 2009).

1.2 Study objectives

1.2.1 General objectives

The main aim of this study is to investigate the socioeconomic and socio-geographic factors that are statistically associated with childhood anaemia in children aged five years or less from Tanzania and Angola. Current literature and research shows that the determinants that are significantly associated with childhood anaemia include: the sex of child, the child's age, mother's anaemia status, wealth index, type of place of residence, mother's or guardian's education level, consumption of iron-rich food (vegetables, meat, and fruits), taking milk prior to survey, stunting, wasting, child had fever in the past two weeks, child had cough in the past two weeks before survey, weight at birth, recent diarrhoea, have a television, have a radio, husband/partner's educational level attainment, vitamin A in the six weeks before the survey, whether children under five slept under mosquito bed nets during the last night before the survey, place of birth delivery, size of child at birth and whether drugs for intestinal parasites has been taken in the last six months (Habyarimana et al., 2017; Kejo et al., 2018; Schellenberg et al., 2003; De Pee et al., 2002; Leal et al., 2011).

1.2.2 Specific objectives

The specific objectives are:

- To identify the determinants of anaemia by fitting the appropriate statistical

models.

- To cater for the complexity of the survey design of DHS data sets by fitting survey logistic regression, for us to draw valid statistical inferences.
- To account for DHS data where there is a high possibility of correlation between observations (members from the same household or cluster) by fitting the Generalized additive mixed models.
- To determine the common determinants of childhood anaemia (determining which variables are said to be the common cause of anaemia in children).
- To fit spatial generalized linear mixed models to account for the possible effect of spatial heterogeneity since DHS data sets are spatial in nature.
- To make recommendations to the current health policy makers by suggesting what factors they should place strong focus on to reduce the prevalence of anaemia in children.

1.2.3 Importance of the study

Anaemia in the population of sub-Saharan countries is the most common cause of morbidity and mortality in children under five years of age. For instance, in its attempts to reduce the prevalence of childhood anaemia over 26 years (1990-2016) Tanzania achieved a reduction from its highest value of 78.20% in 1990 to its lowest value of 55.20% in 2016 for children under five years of age ($\leq$ 59 months). It was 71.8% in 2004-2005 (indexmodi, 2016b). For women of reproductive age, its highest value was 50.90% in 1990 and its lowest value was 37.20% in 2015. The maximum value in pregnant women over 26 years (1990-2016) was 55.50% and its lowest value was 48.0% in 2016 (indexmodi, 2016b). 2016b). In Angola the prevalence of anaemia among children aged 5 or less was 50.90% in 2016 (indexmodi, 2016a). Its highest value over 26 years (1990–2016) was 71.70% in 1990, while its lowest value was 50.90% in 2016 among women of reproductive age (15-49 years). In pregnant women it highest value over 26 years (1990-2016) was 58.00% in 1990, while its lowest value was 46.70% in 2012, and in 2016 it was 47.20% (indexmodi, 2016a; Worldbank, 2016).

Thus, anaemia is a health problem indicating that there is still much to be done by health experts to fight the disease. Furthermore, it remains important for researchers to continue studying the possible determinants of anaemia to help policymakers identify the factors they should mostly focus on to reduce its prevalence in Africa and in the world as whole.

1.3 Literature review

A series of studies have been conducted in the field that shows childhood anaemia as the commonest and most intractable cause of both morbidity and mortality in pre-school children (under five years of age) in sub-Saharan countries (Organization et al., 2008; Osungbade & Oladunjoye, 2012). Although the fight to improve the treatment of this disease has been global, it remains common, and iron deficiency is considered the commonest cause of anaemia in both children and adults. There are several studies of childhood anaemia, focusing mainly on the prevalence and the factors that are positively associated with it.

Discussed here is a brief review of earlier published work a few years ago and the most recent studies on anaemia (childhood anaemia) in Africa as a whole and related studies in countries on other continents. Several studies in notable published work on anaemia show that the most common determinants of childhood anaemia are age of the child, wealth index (standard of living), type of place of residence, level of education, iron supplementation, maternal anaemia, parasite infections (malaria) and child nutritional status (stunting and/or wasting, underweight) (Sanou & Ngnie-Teta, 2012; Ngesa & Mwambi, 2014; Habyarimana et al., 2017; Balarajan et al., 2011b; Ngwira & Kazembe, 2015; Foote et al., 2013; Getaneh et al., 2017; Magalhães et al., 2013). There are many other existing factors (socioeconomic and demographic) that contribute to childhood anaemia.

School-based (or community-based or regional-based) and/or country-based studies on the prevalence and factors associated with childhood anaemia have been conducted and they are indeed showing commonalities about the disease in children under five years old. Schellenberg et al. (2003) conducted a community-based childhood anaemia study in southeastern Tanzania with the purpose of investigating its prevalence. The researchers (2003) found that anaemia prevalence was high for children aged 6-11 months, that children who were from poor homes were highly associated with severe anaemia and that children who had a history of recent illness were more likely to be anaemic compared to those with no reported illness.

A regional study by (Kejo et al., 2018) in the Arusha district in Tanzania used a multivariable logistic regression model to investigate the factors that were significantly associated with childhood anaemia (children under five years of age). The study revealed that childhood anaemia was significantly associated with iron deficiency (due to cultural behaviours), maternal factors such as employment and breastfeeding and the age of the child. Children less than 24 months were more likely to be

anaemic compared to older children, similar to the findings of the study by Schellenberg et al. (2003) regarding the prevalence of anaemia in children less than 24 months of age. A similar study was later conducted in Tanzania (Simbauranga et al., 2015) in the Mwanza district or region. This was a cross-sectional hospital-based study of children under five years of age that found childhood anaemia was highly associated with low iron intake, unemployment of caretakers and low level of caretaker education. The prevalence of anaemia was also high for children living in malaria-endemic areas in the region and for malnourished children.

Semedo et al. (2014) assessed the factors and the prevalence of anaemia on nine islands of Cape Verde in West Africa using the multi-variable logistic regression model. Socioeconomic, demographic and environmental factors were assessed, that included the level of education of caretakers, sex of the child, child age, anaemia status of the mother, household conditions, recent episodes of diarrhoea, duration of exclusive breastfeeding, vaccination and nutritional status. It was found that childhood anaemia was significantly associated with child age (more prevalent in children less than two years), household conditions (poor or rich) and the recent episodes of diarrhoea. Anaemia was more prevalent for children residing in poor household conditions compared to children from rich households, and it was found that children with recent episodes of diarrhoea were at a higher risk.
A study by Foote et al. (2013) in western Kenya (Nyando district) about the prevalence and factors associated with childhood anaemia. The study found that the variables (iron deficiency, stunting, wasting, malaria infections and vitamin A deficiency) that were thought to be highly associated with childhood anaemia were wider spread (Foote et al., 2013). This study used a multi-variable PR (prevalence ratios) regression model to determine factors associated with childhood anaemia and PROC SURVEYFREG was used to account for survey design. Anaemia was significantly and highly associated with malaria infection, child age, stunting and iron deficiency.

Getaneh et al. (2017) conducted a study among pre-school children in Gondar town, Ethiopia, using bivariate and multivariate binary logistic re-gression to assess the effects of socioeconomics, sociodemographics and malaria-associated infectious diseases. It was observed that there was no significant difference in the prevalence of anaemia according to the gender of the child, which is very similar to many other studies on childhood anaemia (Kejo et al., 2018; Semedo et al., 2014; Foote et al., 2013). The study outcome variable of interest was binary and used the logistic regression model to assess the predictors, finding that anaemia was highly associated with stunting, and mother's (caretaker) level of education and nutrition status. An-

other study on anaemia prevalence was conducted with Indonesian infants aged 3-5 months (De Pee et al., 2002). This study also used logistic regression to assess the effect of anaemia predictors, revealing that childhood anaemia was a result of maternal anaemia status, iron deficiency, stunting, maternal education level, and low birth weight. Another similar study was conducted in a rural area in India and had very similar results regarding the factors associated with childhood anaemia (Pasricha et al., 2010).

Ncogo et al. (2017) assessed the prevalence and factors of childhood anaemia in rural and urban settings of the Bata district of Equatorial Guinea. Assessed factors included the sex of the child, type of place of residence, wealth index, the age of the child and many common factors used in studies by many researchers. After the poison regression model was fitted, the model revealed that the type of place of residence was significantly associated with childhood anaemia in children from this district (p-value <0.05). Anaemia was prevalent in children from urban areas who were suffering from concomitant malaria infection, the child's age had a significant bearing; anaemia prevalence was high for children aged between 2 and 24 months and also in children who were from rural areas.
Ngwira & Kazembe (2015) conducted a study using the Bayesian random effect modelling to study the determinants of childhood anaemia in Malawi. Similar to many findings in the literature, this study highlighted that child anaemia decreases with child age and is positively associated with the wealth index, showing that children from wealthy households are less likely to suffer from anaemia compared to those from low-income families. Furthermore, there was a U shaped relationship between maternal age and the possibility of a child being anaemic at birth (the risk of childhood anaemia was high for mothers aged 40 and above).

Habyarimana et al. (2017) assessing the determinants of childhood anaemia in Rwanda, carried out their analysis using the structured spatial additive quantitative regression model. They considered all the levels of anaemia as defined by both the WHO and UNICEF. A number of variables were assessed with the model, revealing that the sex of the child, the mother's literature, wealth index, anaemia status, vitamin A supplementation of the child, the duration of breastfeeding, and the nutritional status of the child (underweight, wasting and stunting) were all significant to childhood anaemia (p-value < 0.05). Furthermore, the structured spatial location effects were also found to have a significant effect on the presence of anaemia.

Almost all of these studies have common findings: child nutrition status, child age (specifically children aged < 24 months), iron deficiency, wealth index and type of

place of residence are highly associated with childhood anaemia in many African countries. For instance, a study of risk factors associated with anaemia in preschool children in Sub-Saharan African countries conducted by Sanou & Ngnie-Teta (2012) found that about 50% of anaemia cases were iron deficiency-related. These results are very similar to the 2008 World Health Organization report on the global prevalence of anaemia (McLean et al., 2009). Infectious diseases were also found to be the second most important cause of anaemia in sub-Saharan Africa. These include malaria, hookworm infestations, schistosomiasis, among others which are all highly prevalent in African countries (McLean et al., 2009; Balarajan et al., 2011a).According to the WHO report, the highest prevalence of childhood anaemia is found in malaria-endemic areas. Reference can be made to the work of Balarajan et al. (2011a); De Benoist et al. (2008); McLean et al. (2009); Sanou & Ngnie-Teta (2012); Magalhaes & Clements (2011); Magalhães et al. (2013), among others.

Chapter 2

Exploratory data analysis

2.1 Introduction

This chapter is based on the data sets as they are used in this study. The data sets used are the 2015-2016 Tanzania demographic and health survey (TDHS) and 2015-2016 Angola demographic and health survey (ADHS) which are found on the DHS programme. The SPSS statistical software was used to give descriptive information about the data sets, and the test of independence (association) between the response variable and each predictor variable was done using a Pearson Chi-square test.

2.1.1 The data set

The two demographic and health surveys utilized produce several different data sets ; these differ in individual surveys which are organized into three distribution categories: survey data, HIV test results, and geographic data (Ministry of Health et al., 2016). This study utilizes the 2015-16 ADHS and TDHS data sets. The two DHS data sets were collected for several reasons: to provide up-to-date information on fertility and childhood mortality, and to assess attitudes toward HIV/AIDS and the use of family planning.

The Tanzania Demographic and Health Survey

The 2015 Tanzania Demographic and Health Survey (2015 TDHS) is the fifth of its kind and follows those implemented in 1991/92 (TDHS), 1996 (TDHS), 2004/05 (TDHS), and 2010 (TDHS). A nationally representative sample of about 13 400 households was selected in Tanzania. All women and men aged 15-49 who were usual residents of the selected households or who slept in the households the night before the survey were eligible to take part. The survey resulted in about 13 000 interviews of women aged 15-49 and 3 500 interviews of men. Similar to previous studies, the

2015 TDHS survey's main objectives were to provide up-to-date information on fertility and childhood mortality levels; fertility preferences; awareness, approval, and use of family planning methods; maternal and child health; and knowledge and attitudes toward HIV/AIDS and other sexually transmitted infections (STIs) (Ministry of Health et al., 2016).

The 2012 Tanzania Population and Housing Census (TPHC) was used for the 2015-2016 sampling frame. The sampling frame is a complete list of enumeration areas (EAs) covering the country provided by the National Bureau of Statistics (NBS) of Tanzania, the implementing agency for the 2015 TDHS. The 2015-16 TDHS sample was achieved by stratifying the 2012 census frame, samples were selected independently in each sampling stratum by a two-stage selection. Stratification was achieved by separating each region into rural and urban areas and it resulted in 59 sampling strata. In the first stage, 608 EAs (also called clusters or primary sampling units or enumeration areas) were selected with probability proportional to the EA size (Ministry of Health et al., 2016). Among the 608 EAs, 108 were from urban areas and the rest were from rural areas. At the second stage, a fixed number of 22 households was selected from each EA.

The overall household response rate was 90% and 93.6% for urban and rural areas, respectively (Ministry of Health et al., 2016). On average, the number of women 15-49 per household was 1.14 and 1.01 for urban and rural areas, respectively. The women's individual response rate was 96%, while the average number of men aged 15-49 per household was 0.94 and 0.87 for the urban and rural areas, respectively; the men's individual response rate was 88% and 91% for urban and rural areas, respectively. For further reading on the sample design and calculations, readers can refer to the (Ministry of Health et al., 2016).

The Angola Demographic and Health Surveys

In Angola, the first surveys conducted were the Demographic and Health Survey and the fourth was a Multiple Indicator Cluster Survey, which is same to DHS. The 2015-16 Multiple Indicator and Health Survey was conducted from October 2015 to March 2016. It was designed to provide data for monitoring the population and health situation in Angola. The main objective was to provide the current information regarding the demographic and health situation of women, men, and children, including fertility levels, marriage, sexual activity, fertility preferences, family planning methods, childhood and maternal mortality, maternal and child health, breastfeeding practices, nutrition, malaria, HIV/AIDS, domestic violence, and child

well-being (Angola, 2016). From the national representative sample total of 20 063; 14,379 were women aged 15-49 in 16,109 households and 5,684 were men aged 15-54 in half of the selected households who were interviewed in the 2015-16 IIMS. This represents a response rate of 96% and 94% for women and for men, respectively. The sample design provided estimates at the national and provincial levels for both urban and rural areas (Angola, 2016).

In both data sets, all eligible women in all sampled households were weighed and measured for anthropometric indicators and were asked to provide a few drops of blood from a finger-prick for on-the-spot anaemia testing. In addition, parents or guardians of all children aged 6-59 months living in the interviewed households were asked for permission to test the children for anaemia and administer a rapid test for malaria. These children were also weighed and measured for anthropometric indicators (Angola, 2016; Ministry of Health et al., 2016).

2.2 Study Variables

Dependent variable for the study

In this study, our outcome (response) variable of interest is haemoglobin concentration level (Hb$<$11g/dl or Hb $\geq$ 11g/dl) in the blood measured in grams per decilitre (g/dl), or simply status of anaemia in a child (anaemic or not-anaemic). As stated earlier, according to the WHO, anaemia has four categories, but for this study anaemia is categorized into two, anaemic and non-anaemic. The response variable is labelled anaemic = 1 and non-anaemic = 0.

Independent/explanatory/predictor variables for the study

A set of explanatory variables (also called predictor variables or covariates) examined in this study was studied in several others (Foote et al., 2013; Habyarimana et al., 2017; Ewusie et al., 2014; Semedo et al., 2014; Simbauranga et al., 2015). For this study the predictor variables were: current age of the child, sex of the child, type of place of residence, wealth index, mother's or guardian's highest education level, child's nutritional status (stunting, wasting, underweight), whether children under five slept under mosquito bed nets on the night before the survey, access to the Internet, whether the child was given baby formula, region, whether child had been given meat, child was caughing in the past two weeks before the survey, size of child at birth, use of vitamin A supplementation, type of bed nets used, sex of the household head, age of the household head and whether the household has a television.

Explaining the explanatory (independent) variables used in this study

The current age of the child as subjectively reported by the mother was categorized as: less than one years (0 months $\leq$ aged $<$ 12 months), one years (12 months $\leq$ aged $<$ 24 months), two years (24 months $\leq$ aged $<$ 36 months), three years (36 months $\leq$ aged $<$ 48 months) and be four years (48 months $\leq$ aged $<$ 60 months). The stunting status was categorized as nourished, moderate and severe stunting; the wealth index (poor, middle and rich); the highest level of education (higher, secondary, primary and no education); internet (no access and have access to internet); sex of the child and household head (female and male); size of the child at birth (large, average and small), the age of the household head was treated as a continuous variable, and the rest of variables were categorized as yes or no.

2.3 Descriptive Statistics

2.3.1 Cross-tabulation

The IBM Statistical Package for the Social Sciences (SPSS) latest version was used to give a summary of descriptive statistics information (cross-tabulation) and in the identification of factors that are significantly associated with anaemia in children (Hb $<$ 11g/dl) at 5% level of significance in both data sets (TDHS and ADHS). The IBM SPSS uses the chi-square to test the independence technique to identify factors that have a significant effect on the response variable.

According to WHO childhood haemoglobin concentration is grouped into several categories, namely: Severely anaemic (Hb$<$7.0 g/dl); Moderately anaemic (7.0 g/dl$\leq$Hb$\leq$9.9 g/dl); Mildly anaemic (10.0 g/dl$\leq$Hb$\leq$10.9g/dl) and children with haemoglobin concentration greater than or equal to eleven grams per decilitre (Hb$\geq$11.0 g/dl) are considered to be free from the diseases (as non-anaemic). In this study, the childhood haemoglobin concentration (outcome variable) is grouped into two categories anaemic (Hb $<$11.0 g/dl) and non-anaemic (Hb$\geq$11.0 g/dl), for the purpose of fitting the survey logistic regression model, whereby the response variable is dichotomous (Organization et al., 2008). Furthermore, during the process of cross-tabutions, the system uses the chi-squared statistic to test for any association or dependence between the response variable and the explanatory variables.

The Pearson Chi-Square test of independence

The chi-square test of independence also known as the Pearson Chi-square has two majors components,

- Goodness of fit of the test; and
- Test for association or independence

A chi-square test can be used to test for independence or association between two categorical variables. For instance, in this study it is used to test wealth index (categorized as either poor, middle and rich), to assess if there is any dependence between this variable and childhood anaemia (categorized as non-anaemic and anaemic) in Tanzania and Angola. The Pearson Chi-square test is only applicable provided the number of expected frequencies in all cells on the contingency table is more than five; otherwise, it fails, and a method called Fisher's scoring is used instead (McHugh, 2013).
The null and the alternative hypothesis are given as follows:

H_0: There is no association between the rows and the columns
versus
H_1: There is association between the rows and the columns.

Similarly,

H_0: There is no association between the predictor and the response variable.
versus
H_1: There is an association between the predictor and the response variable.

Equivalently, it can be stated as the rows are independent of the columns versus the rows are dependent on the columns.

The test statistic is given by McHugh (2013):

$$\chi^2 = \sum_{j=1}^{h}\sum_{i=1}^{2} \frac{\left(Y_{ij} - \frac{Y_{i.}Y_{.j}}{n}\right)^2}{\frac{Y_{i.}Y_{.j}}{n}} = \sum_{j=1}^{h}\sum_{i=1}^{2} \left(\frac{Y_{ij} - E_{ij}}{E_{ij}}\right)^2 \tag{2.1}$$

where, $E_{ij} = \frac{Y_{i.}Y_{.j}}{n}$ are the expected frequencies, n represents the total sample size, i and j are the i^{th} raw and j^{th} column. $Y_{i.}$ represents the row marginal total, $Y_{.}j$ represents the column marginal total and Y_{ij} are the observed frequencies

Basically, the Pearson Chi-squared is used to test for independence between the rows and the columns at 5% level of significance, with degrees of freedom of (row-

1)X(column-1) and the test statistic (2.1) could be simply written as:

$$\chi^2 = \sum_{j=1}^{h} \sum_{i=1}^{2} \frac{(Observed - Expected)^2}{Expected} \tag{2.2}$$

2.3.2 Interpretation of the cross-tabulations

The cross-tabulation for each predictor variable crossed with the dependent variable consists of the total number of children who were eligible for that specific question, missing responses or data and the frequency for each variable at its specific level. The Pearson Chi-squared value and the p-value provided on the tables below (cross-tabulations results) are for the overall significance of the predictor variable and not for a specific category.

Looking at the results of the cross-tabulations on the TDHS and the ADHS data sets (Table 2.1 and Table 2.3) we can see that childhood anaemia was highly associated (p-values < 0.0001) with the age of the child, type of place of residence (only in Tanzania), the highest education level of the household head, currently breast-feeding, stunting and wasting. Furthermore, looking at the factors associated with anaemia in children by country, in Tanzania (Table 2.1) we can observe that sex of the child (p-value = 0.002), wealth index (p-value = 0.033), access to internet (p-value = 0.010), whether children under five slept under mosquito bed net (p-value = 0.003) and whether the child had a cough the last two weeks before the study (p-value = 0.015) were also positively associated with childhood anaemia. However in Angola, the additional factors which were positively associated with childhood anaemia include: child gender (p-value = 0.009), wealth index (p-value = 0.005), availability of television in a household (p-value = 0.009) and the size of the child at birth (p-value = 0.002).

In total, the number of children who were eligible for the study in Tanzania and Angola was respectively, n_T=10233 and n_A=14322. Out of the sample of n=10233 children in Tanzania, n=5153 were males and n=5083 were females. In Angola, there were n=7143 males and n=7179 female children eligle for the study. Now looking at the number of potential children for the survey by country. Starting with Tanzania, there were n=1601 non-anaemic and n=2419 anaemic male children, n=1729 non-anaemic female children and n=2265 anaemic female children (Table 2.1). Moreover, out of all the eligible children in Tanzania only n=8014 participated in the study and in Angola only a sample of n=5635 participated on the study.

Table 2.1: Cross tabulation analysis TDHS data sets

Predictor variable	Non-anaemic	Anaemic	χ^2-value	P-value
Child gender				
Male	1601	2419	9.89	0.002
Female	1729	2265		
Child age (in months)				
0-11	213	794	670.46	< 0.0001
12-23	562	1525		
24-35	749	1001		
36-47	882	706		
48-59	924	658		
Type of place of residence				
Urban	816	980	14.36	< 0.0001
Rural	2514	3704		
Gender household head				
Male	2788	3942	0.27	0.601
Female	542	742		
Wealth index				
Poor	701	1132	6.80	0.033
Middle	665	1012		
Rich	1351	1876		
Mother highest education level				
No-education	620	1151	44.51	< 0.0001
Primary	2099	2713		
Secondary	574	789		
Higher	37	31		
Stunting				
Severe	333	618	23.13	< 0.0001
Moderate	777	1132		
Nourished	2207	2899		
Wasting				
Severe	63	158	21.38	< 0.0001
Moderate	361	579		
Nourished	2901	3940		

[a]

[a] **The P-values presented in the tables are not for a specific category but an overall significance of the predictor variable**

Table 2.2: TDHS cross tabulation results continues from the previous page

Predictor variable	Anaemic	Non-anaemic	χ^2-value	P-value
Underweight				
Severe	19	56	15.97$<$ 0.0001	
Moderate	88	175		
Nourished	3204	4416		
Access to internet				
No access	3134	4468	6.71	0.010
Have access	196	215		
Gave child meat (Pork, lamb, beef, etc)				
No	185	332	0.17	0.679
Yes	1599	2987		
Children under 5 slept under mosquito bed net				
No	1164	1496	8.54	0.003
Yes	2151	3181		
Currently breastfeeding				
No	1751	1793	161.43	$<$ 0.0001
Yes	1579	2891		
Iron supplementation				
No	413	698	0.46	0.759
Yes	1654	2709		
Vitamin A supplementation				
No	1913	2764	1.73	0.189
Yes	1400	1904		
Child had cough in the last two weeks				
No	2805	3851	5.92	0.015
Yes	521	830		
Household has a television				
No	2636	3869	24.68	$<$ 0.0001
Yes	584	635		
Size of child at birth				
Small	2598	3651	1.92	0.381
Average	445	621		
Large	180	219		

[a]

[a] **The P-values presented in this tables are not for a specific category but an overall significance of the predictor variable**

Table 2.3: Cross tabulation analysis ADHS data sets

Predictor variable	Non-anaemic	Anaemic	χ^2-value	P-value
Child gender				
Male	960	1909	6.91	0.009
Female	1018	1748		
Child age (in Months)				
0-11	120	541	272.01	< 0.0001
12-23	345	1034		
24-35	440	788		
36-47	526	743		
48-59	547	551		
Type of place of residence				
Urban	1102	1982	1.19	0.275
Rural	876	1675		
Gender household head				
Male	1320	2402	0.63	0.426
Female	658	1255		
Wealth Index				
Poor	1031	1917	10.43	0.005
Middle	455	949		
Rich	492	791		
Mother highest education level				
No-education	677	1308	20.27	< 0.0001
Primary	718	1433		
Secondary	530	865		
Higher	53	51		
Stunting				
Severe	237	638	36.12	< 0.0001
Moderate	447	870		
Nourished	1247	2050		
Wasting				
Severe	69	212	23.38	< 0.0001
Moderate	255	566		
Nourished	1617	2814		

[a]

[a] **The P-values presented in this tables are not for a specific category but an overall significance of the predictor variable**

Table 2.4: ADHS cross tabulation table continues from the previous page

Predictor variable	Non-anaemic	Anaemic	χ^2-value	P-value
Underweight				
Severe	21	39	2.62	0.270
Moderate	64	150		
Nourished	1856	3403		
Access to internet				
No access	1819	3390	0.99	0.318
Have access	159	267		
Gave child meat				
No	207	462	0.21	0.646
Yes	1044	2234		
Children under 5 slept under mosquito bed net				
No	1253	2412	3.54	0.060
Yes	716	1235		
Access to internet				
No access	1819	3390	0.99	0.318
Have access	159	267		
Gave child meat				
No	207	462	0.21	0.646
Yes	1044	2234		
Currently breastfeeding				
No	958	1455	39.12	< 0.0001
Yes	1020	2202		
Iron supplementation				
No	291	700	2.43	0.296
Yes	712	1560		
Child had cough in the last two weeks				
No	1764	3263	23.79	0.067
Yes	1764	3263		
Household has a television				
No	1032	2005	9.47	0.009
Yes	937	1615		
Size of child at birth				
Small	170	331	19.23	0.002
Average	1132	2045		
Large	605	1069		

[a]

[a] **The P-values presented in this tables are not for a specific category but an overall significance**

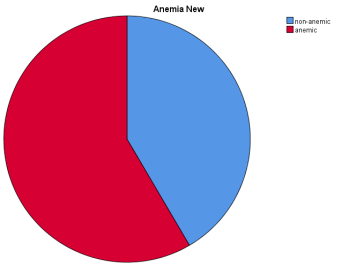
The cross tabulation results for both Tanzania (Table 2.1) and Angola (Table 2.3) show a decreasing trend in the number of anaemic children as the age of the child (in months) increases. Looking at the cross tabulations for Tanzania only, (Table 2.1 and/or Table 2.2). Among children aged 0-11 months, n=213 were non-anaemic and n=794 were anaemic, while of those who were aged 48-59 months, n=924 were non-anaemic and n=658 were anaemic. Many of the children who contributed to the study (TDHS) were from rural areas, and n=3704 and n=2514 of these children were respectively anaemic and non-anaemic, while of children from urban areas, there were n=980 anaemic and n=816 non-anaemic. The cross tabulations show that most of the under five children who were born from rich families, n=1351, were non-anaemic, and n=1876 were anaemic. From poor households, there were n=701 non-anaemic and n=1132 anaemic children. Very few of the under five children were anaemic born to highly educated mothers n=31, while n=1151 children were anaemic if born to uneducated mothers. There were n=2099 non-anaemic and n=2713 anaemic children whose mothers had primary education. Of all the eligible children in Tanzania, a total of n=333 non-anaemic and n=618 anaemic children were severely stunting, and n=2901 and n=3940 were respectively non-anaemic and anaemic. Many children were residing in households that had no internet access, of which n=3134 were non-anaemic and n=4468 were anaemic.

In Angola (Table 2.3 and Table 2.4) anaemic male children were higher in number than female children, n=1909 and n=1748 respectively. Children aged 12-23 months were higher in number compared to all the other groups, n=345 were non-anaemic and n=1034 were anaemic. The cross tabulations show that of the many eligible children who were from urban areas, n=1102 were non-anaemic and n=1982 were anaemic and of those from rich families, n=1031 were non-anaemic and n=1917 were anaemic. A similar trend observed in the TDHS cross tabulations (Table 2.1) is observed in the ADHS cross tabulations (Table 2.3) regarding the number of anaemic children and the level of education of the mother or household head. There were very few children with mothers who were highly educated, and of these, n=53 were non-anaemic and n=51 were anaemic. Many anaemic children had uneducated mothers, with n=1308 who were anaemic and n=677 who were not anaemic. Furthermore, we observed a similar trend in Angola, regarding the number of children who participated in the study, as in Tanzania. The number of anaemic children who were severely stunted was high, with n=237 non-anaemic children and n=638 anaemic children, in households with no television, n=1032 were non-anaemic and n=2005 were anaemic, and in households with no access to internet, n=1819 were non-anaemic and n=3390 were anaemic. In addition, many of the children who participated in the study were not suffering (had normal status) from underweight (n=1856

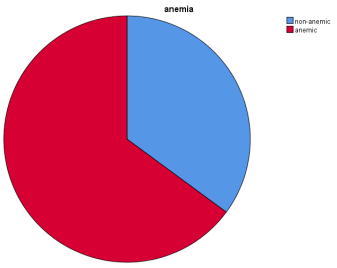
non-anaemic and n=3403 anaemic), wasting (n=1617 non-anaemic and n=2814 anaemic) and stunting (n=1247 non- anaemic and n=2050 anaemic). Moreover, the two data sets were represented graphically (specifically, pie charts presentations) in the next subsection.

Graphical presentation of the two data sets

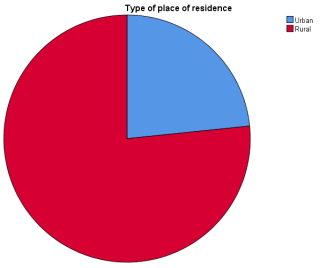
Pie charts were used to show the proportion of children per categories (not the number of children per category) on each variable used to predict anaemia in children. Specifically, the pie charts are only for child age , child gender, wealth index, mother's education level and type of place of residence. The pie charts showing the proportions of anaemic and non-anaemic children are provided for each country.



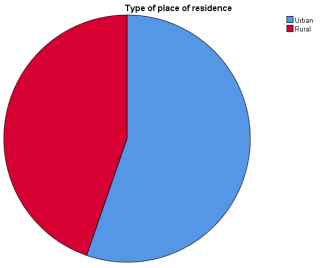
(a) Anemia Status (Tanzania)



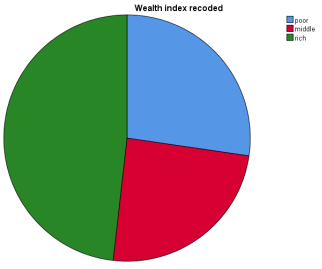
(b) Anemia Status (Angola)



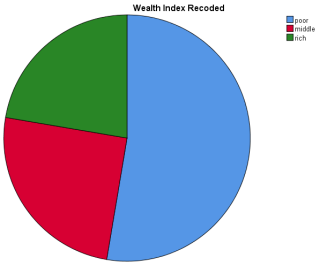
(a) Residence type (Tanzania)



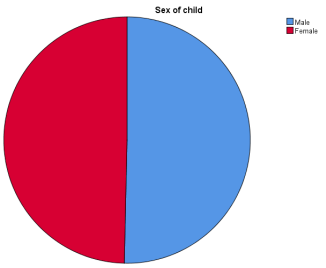
(b) Residence type (Angola)



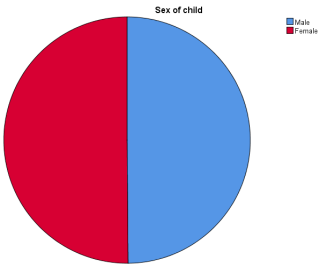
(a) Wealth Index (Tanzania)



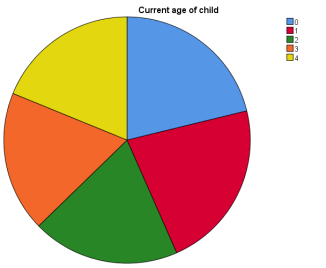
(b) Wealth Index (Angola)



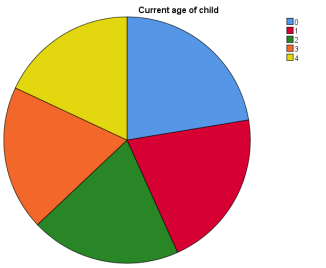
(a) Child gender (Tanzania)



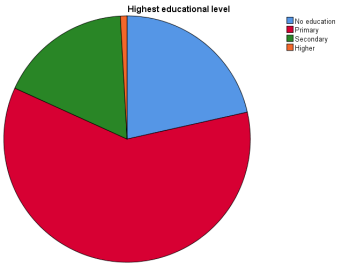
(b) Child gender (Angola)



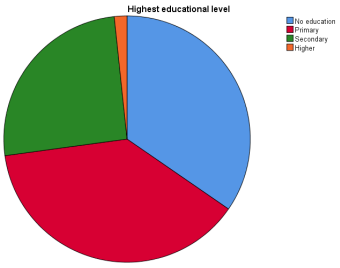
(a) Child age (Tanzania)



(b) child age (Angola)



(a) Education Level (Tanzania)



(b) Education Level (Angola)

Figure 2.4a and Figure 2.4b, show that in both countries the number of male and female children who participated in the study were almost equal. Many children were found to suffer from anaemia in both countries (Figure2.1a and Figure 2.1b) in comparison to being non-anaemic. The pie chart Figure 2.5b and Figure 2.5a show almost equal proportions of children participated in the study; regarding age as a predictor variable. Although there was high number of children from urban areas who participated in the TDHS study (Figure 2.2a), however in Angola (ADHS) its vice versa (Figure 2.2b). Furthermore, the proportions of children who had highly educated mothers look similar in both countries (Figure 2.6a and Figure 2.6b), showing that there were few children in both countries in the study with educated mothers. Although a higher number children had mothers with primary education. All the conclusions drawn here, based on the size of the slice in a pie chart, show that a similar trend can be followed to make inferences about the rest of the predictor variables used in the study of childhood anaemia.

Chapter 3

Generalized Linear Models

3.1 General Linear Regression

Introduction

Regression is the study of dependence (Weisburg, 2005). Regression analysis is the statistical methodology that is used to investigate the functional relationship between the dependent variable, Y, and the set of independent variables, $x_1, x_2, x_3 \ldots, x_p$, (Montgomery et al., 2012). This relationship may be expressed in the form of an equation or a model coupling the response (dependent) variable and the predictor variables (Mildenberger, 2012; Khuri, 2009).

Model overview and estimation of regression coefficients

The response variable is denoted by Y and the set of predictor variables are denoted by $x_1, x_2, x_3 \ldots, x_p$, where p is the total number of predictor variables. Mildenberger (2012), Depending on the nature of the explanatory variable if they appear linearly then true relationship between the response and the predictor (explanatory) variables can be approximated by the linear model

$$Y_i = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \varepsilon_i, \tag{3.1}$$

Where $i = 1, 2, \ldots, n$. $x_1, x_2, x_3 \ldots, x_p$ are the explanatory variables that represents the levels of associated factors, and $\beta_1, \beta_2, \ldots, \beta_p$ are the unknown regression parameters, ε is the random error term representing the deviation in the approximation. The error is there to account for the failure of the model to fit the data exactly and is assumed to be identically, independently and normally distributed with a mean of zero and a variance of σ^2 , $\varepsilon \sim N(0, \sigma^2)$ under the homogeneous variance assumption.

$x_1, x_2, x_3 \ldots, x_p$ are commonly referred as control, input, regressor, predictor, independent or explanatory variables. A more general expression of the linear regression is one of the form (Khuri, 2009).,

$$Y_i = f(x_{i1}, x_{i2}, x_{i3} \ldots, x_{ip}) + \varepsilon_i. \tag{3.2}$$

The function $f(x_1, x_2, x_3 \ldots, x_p)$ describes the relationship between the response variable Y and the predictor variables, $x_1, x_2, x_3, \ldots, x_p$, $i = 1, 2, \ldots, n$.

In general, depending on the study of interest, if someone is interested in studying the effect of a single predictor variable on the output and if the nature of the relationship between the output and the predictor variable is linear, simple linear regression would be fitted to study such relationship. For a study that involves two or more predictor variables that study the effect of the predictor on the dependent variable, it is referred to as the multiple linear regression model (Mildenberger, 2012).

Simple Linear regression

The simple linear regression model involves only one explanatory variable, x, and the response variable, Y. The equation is defined as follows

$$Y = \beta_0 + \beta_1 x + \varepsilon, \tag{3.3}$$

Where, β_0 is the model intercept, β_1 the unknown model parameter and ε is the random error term to cater for the failure of the model to fit data exactly (Montgomery et al., 2012).

The mean and the variance of the response variable, Y (assuming that the response variable, X is fixed) can be given as

$$E(Y|X) = \mu_{xy} = E(\beta_0 + \beta_1 x_1 + \varepsilon), \tag{3.4}$$

and the

$$var(Y|X) = \sigma^2{}_{xy} = var(\beta_0 + \beta_1 x_1 + \varepsilon) = \sigma^2. \tag{3.5}$$

Estimation of the model parameters

The unknown model parameters for general linear models could be estimated by using the well-known method of Ordinary Least Squares(OLS). In this study we do not really focus on Linear models, thus for interested readers they can refer to the

texts by (Montgomery et al., 2012; Mildenberger, 2012; Bingham & Fry, 2010) and the text by Weisburg (2005) for more reading about the full theory of general linear models.

There are other several types of regression analysis, and the table below shows the type of regression and their conditions, specifying when each model can be applied

Type of regression

The table (Table 3.1) below summarizes some of the types of regression and gives conditions on when they can be applied.

Type of regression	**Condition**
Univariate regression	Only one quantitative response variable
Multivariate regression	Two or more quantitative response variable
Simple regression	Applicable when there is only one predictor(scalar) variable
Multiple regression	Applicable when there are two or more predictor variables
Linear regression	Applicable when all the parameters enter the equation linearly
Nonlinear regression	When the relationship between some of the predictor and response variable is nonlinear or when some of the parameters are nonlinear
Analysis of variance regression	Applicable when ther are mixed predictor variables, both qualitative and quantitative
Logistic regression	The type of predictive model that can be used when the target(response) variable is a categorical variable with two (binary) categories (for example the response could be: affected / not affected, male or a female, yes/no etc.)
Ordinal regression model	The type of predictive model that can be used when the target(response) variable is a categorical variable with more than two categories (for example the response could be: the disease is either Severe, Moderate and mild)

Furthermore, general linear models are appropriate to model the relationship between the response and the explanatory variables only if the response variable is continuous, otherwise it fails. General linear models do not hold if:

- The range of Y is restricted (e.g. binary, count); and
- If the variance depends on the mean $(E(Y_i) = \mu_i)$.

3.2 Generalized Linear Models

3.2.1 Introduction

The term generalized linear models (GLMs) refers to a large class of models popularized by Nelder & Wedderburn (1972), and fully developed by McCullagh (1989). Generalized linear models are a generalization of ordinary linear models. They are applicable when the response variable Y_i is assumed to follow an exponential family distribution with a mean μ_i. The Exponential family distribution includes distributions such as normal distribution, poisson distribution, binomial distribution and the gamma distribution McCullagh (1989). Generalized linear models are useful for non-normal data, such as binary data.

These types of models accommodate response variables that violate general linear models assumptions through two mechanisms: a link function and a variance function; it links the mean of the dependent variable Y_i, which is $E(Y_i) = \mu_i$, to the linear term $x_i^T \beta_i$, for $i = 1, 2, 3..., n$, in such a way that the range of the non-linearly transformed mean $g(\mu_i)$ ranges from -∞ to ∞ (Nelder & Wedderburn, 1972). Thus, a linear equation $g(\mu_i) = x_i^T \beta_i$ will result and an iteratively re-weighted least squares method for maximum likelihood can be used to estimate the model parameters. Furthermore, the asymptotic normality and constancy of variance are no longer required (Nelder & Wedderburn, 1972). There is not much difference between generalized linear models and general linear models or ordinary linear models in terms of the process of model specification, except that generalized linear models use a link function to account for the non-continuity and possibly bounded response variable.

3.2.2 Model structure

General linear models have a set of restrictive assumptions, one of which is that the dependent variable Y is normally distributed conditioned on the value of predictors with a constant variance regardless of the predicted response value (Nelder & Wedderburn, 1972; Olsson, 2002). The advantages of linear models include: easy to compute, an interpretable model form and the ability to compute certain diagnostic information about the quality of the fit. Generalized linear models relax these restrictions which are often violated in practice. For example, a binomial distribution has a binary (yes/no or 0/1) responses, the variance is not the same across classes. Furthermore, the sum of terms in a linear model can typically have large ranges encompassing very negative and very positive values, (McCullagh, 1989; Turner, 2008).

A generalized linear model (Dobson & Barnett, 2008; Dunteman & Ho, 2005; Turner,

2008) is made up of a linear predictor and two more functions, namely: link function and variance function. GLM's are expressed in the following equation:

$$g(\mu_i) = \eta_i = \beta_0 + \sum_{i=j}^{p} X_{ij}\beta_{ij}. \tag{3.6}$$

Where the link function $g(\mu_i)$ describes how the mean $E(Y_i) = \mu_i$ depends on the linear predictor and the variance function, $var(\mu_i)$, describes how the variance depends on the mean through the equation.

$$var(\mu_i) = \phi V(\mu)$$

, where the constant ϕ is called a dispersion parameter. $g(\cdot)$ is a pre-specified function called the link function, β_0 is the intercept or constant term and $\beta_1, \beta_2, \ldots, \beta_p$ are the set of regression parameters.

Model assumptions

Generalized linear models are the extension of linear models with the purpose of relaxing the assumption of normality. The following assumptions about GLMs are based on Kutner et al. (2005); McCullagh (1989)

- The response variables $Y_1, Y_2, \ldots Y_n$ are independently distributed.
- The response variable does not necessarily have to follow normal distribution, but typically assumes an exponential family distribution (Poison, binomial, etc).
- The homogeneity of variance does not need to be satisfied: $var(y_i) = \phi V(\mu_i)$.
- The expectation of the error terms $E(\varepsilon) = 0$, hence the expectation of the response variable Y_i could be expressed as $E(Y) = E(X\beta) + E(\varepsilon) = X\beta$.

These assumptions are somewhat strict for data which assume no normality. Therefore, GLMs can be used to model such data whose distributions are from the exponential family of distributions

3.2.3 Exponential family

The exponential family is a general class of distributions that is made of well-known distributions known as the special cases (Dobson, 2002). It comprises both the discrete and continuous random variables that includes a normal distribution (Olsson, 2002). Binomial, poison, gamma, multinomial and weibull distributions are also

special cases of exponential family. It can be shown that a distribution is a class of exponential family provided the probability distribution function of an observation Y_i for $i = 1, 2, 3, \ldots, n$ is known and can be shown to be in the form

$$f(y_i, \theta_i, \phi) = exp\left(\frac{(y_i\theta_i - b(\theta_i))}{a_i(\phi)} + (\phi + c(y_i, \phi))\right), \tag{3.7}$$

where $a_i(\phi)$ are $b(\theta_i)$ are known functions and $c(y_i, \phi)$ is a function of y_i and ϕ. The parameter θ_i is known as the cononical parameter, and ϕ is the dispersion parameter.

For all distributions that belong to the class of exponential family their general expression of the mean and variance are respectively given by, $\mu = E(X) = b'(\theta_i)$ and the variance is given by $var(x) = a(\phi)b''(\theta_i)$. The derivation of the mean and variance are shown below: Generally, the area under a curve add up to one, if a function $f(y, \theta, \phi)$ is known. Therefore the area under f can be expressed as (Dobson & Barnett, 2008)

$$\int f(y, \theta, \phi)dy = 1, \tag{3.8}$$

where the integration is over all the possible values of y, for a discrete random variable y the integration sign is substituted by the summation sign

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

Differentiating 3.8 on both sides with respect to θ, results in

$$\frac{d}{d\theta}\int f(y, \theta, \phi)dy = 0. \tag{3.9}$$

Reversing the order of integration and the differentiation, results in

$$\int \frac{d}{d\theta} f(y, \theta, \phi)dy = 0. \tag{3.10}$$

Similarly, if we differentiate 3.8 twice with respect to θ and again reverse the order of integration and differentiation it gives,

$$\int \frac{d^2}{d\theta^2} f(y, \theta, \phi)dy = 0. \tag{3.11}$$

These results can be applied for any distribution that belongs to the class of exponential family. From 3.10 substituting for $f(y, \theta, \phi)$

$$\int \frac{d}{d\theta} f(y,\theta,\phi)dy = 0,$$
$$\int \frac{d}{d\theta} exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right) dy = 0,$$
$$\int \left(\frac{y\theta - b'(\theta)}{a(\phi)}\right) f(y,\theta,\phi)dy = 0,$$
$$\int \left(\frac{yf(y,\theta,\phi)}{a(\phi)}\right) dy - \int \left(\frac{b'(\theta)f(y,\theta,\phi)}{a(\phi)}\right) dy = 0,$$
$$\int \left(\frac{yf(y,\theta,\phi)}{a(\phi)}\right) dy = \int \left(\frac{b'(\theta)f(y,\theta,\phi)}{a(\phi)}\right) dy.$$

Therefore,

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

But $\int yf(y,\theta,\phi)dy = E(y)$, thus

$$\begin{aligned} E(y) &= b'(\theta)\int f(y,\theta,\phi)dy \\ &= b'.(\theta) \end{aligned} \tag{3.12}$$

And from substituting for $f(y,\theta,\phi)$ in Equation 3.11

$$\begin{aligned} \int \frac{d^2}{d\theta^2} f(y,\theta,\phi)dy = 0, \\ \int \frac{d^2}{d\theta^2} exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right) dy = 0. \end{aligned} \tag{3.13}$$

Following the same steps as we did in deriving for the $E(y)$, we can see that 3.13 can be solved to its simplest form as follows

$$\begin{aligned} \int \left(\left(\frac{y - b'(\theta)}{a(\phi)}\right) f(y,\theta,\phi) - \frac{b''(\theta)}{a(\phi)} f(y,\theta,\phi)\right) dy = 0. \\ \int \left(\frac{y - b'(\theta)}{a(\phi)}\right)^2 f(y,\theta,\phi)dy = \int \frac{b''(\theta)}{a(\phi)} f(y,\theta,\phi)dy. \end{aligned} \tag{3.14}$$

Thus,

$$\frac{1}{a(\phi)^2}\int \big(y - b'(\theta)\big)^2 f(y,\theta,\phi)dy = \frac{b''(\theta)}{a(\phi)}. \tag{3.15}$$

In general, a variance is defined as $var(y) = \int (y - E(y))^2 f(y)$. Therefore, considering only the left side of Equation 3.15

$$\frac{1}{a(\phi)^2}\int \big(y - b'(\theta)\big)^2 f(y,\theta,\phi)dy = \frac{1}{a(\phi)^2}var(y). \tag{3.16}$$

Therefore,

$$\frac{1}{a(\phi)^2}var(y) = \frac{b''(\theta)}{a(\phi)}. \tag{3.17}$$

Hence,

$$var(y) = b''(\theta)a(\phi). \tag{3.18}$$

The following section focuses on how to estimate the unknown parameters $(\beta_1, \ldots, \beta_p)$ of the generalized linear model.

3.2.4 Parameter estimation

For generalized linear models a well-known method of maximum likelihood is theoretically used for parameter estimation (Dobson, 2002). Assuming that the outcomes are random and independent variables, the maximum likelihood function is defined as the product of the joint probability distribution (Allison, 2012). The likelihood function can be expressed as follows, for $i = 1, 2, \ldots, n$

$$\begin{aligned} L(y_i, \theta_i) &= \prod_{i=1}^{n} f(y_i, \theta_i, \phi) \\ &= \prod_{i=1}^{n} exp\left(\frac{y\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi)\right). \end{aligned} \tag{3.19}$$

Taking a natural log on both side of equation 3.19

$$\begin{aligned} l(y_i, \theta_i) &= \sum_{i=1}^{n}\left(\frac{y\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right). \\ &= \sum_{i=1}^{n} a_i(\phi)^{-1}\left((y_i\theta_i - b(\theta_i)\right) + c(y_i, \phi). \end{aligned} \tag{3.20}$$

For a single observation i the log-likelihood function is given by

$$l_i(\beta_i) = a_i(\phi)^{-1}\left((y_i\theta_i - b(\theta_i)\right) + c(y_i, \phi)$$

partially differentiating the log-likelihood function for observation i, with respect to the regression coefficient β_j for $j = 0, 1, 2, \ldots, p$. We obtain a score vector function

given by, $\left(\bigcup_{\beta_0}, \bigcup_{\beta_1}, \ldots, \bigcup_{\beta_p}\right)'$. Where $\bigcup_{\beta_j}$

$$\begin{aligned} \bigcup_{\beta_j} &= \frac{\partial l(\beta_j)}{\partial \beta_j}, \\ &= \frac{\partial l_i}{\partial \beta_j}. \end{aligned} \tag{3.21}$$

Applying chain rule to 3.21

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_i}. \tag{3.22}$$

but,

$$\frac{\partial l_i}{\partial \theta_i} = \frac{y_i - b'(\theta)}{a_i(\phi)}. \tag{3.23}$$

For $E(y) = \mu_i = b'(\theta)$ and $var(y) = a(\phi)b''(\theta)$, therefore 3.23 becomes

$$\frac{\partial l_i}{\partial \theta_i} = \frac{y_i - \mu_i}{a_i(\phi)}. \tag{3.24}$$

and

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta) = \frac{var(y_i)}{a_i(\phi)}. \tag{3.25}$$

Furthermore, the link function is defined as (η),

$$\eta_i = \sum \beta_j X_{ij}.$$

Therefore,

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$$

where x_{ij} is the j^{th} element of the covariates vector x_i for the i^{th} observation. Substituting back to the score function 3.21, the score function becomes,

$$\frac{\partial l}{\partial \beta_j} = \sum \frac{y_i - \mu_i}{a_i(\phi)}\frac{a_i(\phi)}{var(y_i)}x_{ij}\frac{\partial \mu_i}{\partial \eta_i} = \sum \frac{y_i - \mu_i}{var(y_i)}x_{ij}\frac{\partial \mu_i}{\partial \eta_i}. \tag{3.26}$$

$\frac{\partial \mu_i}{\partial \eta_i}$ depends on the link function for each distribution. The estimating function can also be used to determine the asymptotic covariance matrix of $\hat{\beta}$, the information

matrix $I(\beta)$ is given by

$$\begin{aligned} I(B) = -E\left(\frac{\partial^2 l}{\partial\beta_i\beta_j}\right) &= E\left[\left(\frac{\partial l}{\partial\beta_i}\right)\left(\frac{\partial l}{\partial\beta_i}\right)\right], \\ &= E\left[\frac{y_i-\mu_i}{var(y_i)}x_{ih}\frac{y_i-\mu_i}{var(y_i)}x_{ij}\frac{\partial\mu_i}{\partial\eta_i}\right]. \end{aligned} \tag{3.27}$$

Which can be simplified to,

$$I(B) = \sum_i n\frac{x_{ih}x_{ij}}{var(y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2. \tag{3.28}$$

If we let W be the diagonal matrix with main diagonal elements,

$$W = \frac{\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2}{var(y_i)},$$

the inverse matrix $I(\beta)$ becomes

$$I(\beta) = X^T W X.$$

and the asymptotic covariance matrix becomes,

$$cov(\hat{\beta}) = \left(X^T W X\right)^{-1}$$

and the score equation $\bigcup_{\beta_j}$ 3.21 reduces to,

$$\bigcup_{\beta_j} = \frac{\partial l(\beta_j)}{\partial\beta_j} = \sum_{i=1}^{n}(y_i-\mu_i)W_i\frac{\partial\eta_i}{\partial\mu_i}x_{ij}.$$

By equating the score function to zero, the maximum likelihood estimates $\hat{\beta}$ can be obtained by applying the Iterative Re-weighted Least Squares(IRLS), the Newton Raphson(NR) or a Fisher Scoring(FS) method (McCullagh, 1989; Olsson, 2002). Using IRLS or FS method will give the same parameter estimate, while the FS and NR methods give similar results. This is because FS estimated covariance matrix of parameters may be slightly different since FS is based on the expected information matrix while NR is based on the observed information matrix (Heeringa et al., 2017). The method of finding the maximum likelihood estimators using FS, NR and IRSL methods are further discussed in appendix B.

3.3 Model Selection and Diagnostic of Generalized Linear Models

3.3.1 Assessing the fit of GLMss

In generalized linear models, deviance and Pearson Chi-square test are two tests used in assessing the goodness of fit of the model. During data analysis, a model is said to be the best if it can statistically fit data well and it minimizes discrepancy between the expected values under the model and observed values (Olsson, 2002).

The deviance

In GLMs the quality of the model fitted to data can be assessed through the deviance and thus the goodness of fit in the nested models (Olsson, 2002), it measures the discrepancy of fit between the maximum likelihood of the saturated model and the log-likelihood of the fitted model. The deviance denoted by D can be defined as follows,

$$D = 2l(y, \phi, y) - 2l(\hat{\mu}, \phi, y)$$

where $l(y, \phi, y)$ is the log-likelihood of the saturated model or full model (model with n parameters) and $l(\hat{\mu}, \phi, y)$ is the log-likelihood of the reduced model and $\hat{\mu}$ is the maximum likelihood estimator of the model of interest.

But for inference purposes, scaled deviance is used rather than simply deviance only. All parameters that have scaled parameters, for example, binomial distribution, poison distribution, the scaled deviance is defined as follows (Olsson, 2002),

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

For binomial and poison distributions, the deviance and the scaled deviance are the same. In addition, deviance is very important when comparing competing models (Olsson, 2002). If the model fits data well, the deviance will asymptotically tend towards chi-square distribution as degrees of freedom increase (or as n increases).

The generalized Pearson Chi-square (χ^2) statistic

The Pearson χ^2 goodness of fit test is the alternative method used to test for the goodness of the fit of the model and comparing competing models, which is defined as (Allison, 2012)

$$\chi^2 = \sum \frac{(y_i - \hat{\mu})^2}{\hat{v}(\hat{\mu})}$$

where $\hat{v}(\hat{\mu})$ is the estimated variance function (Allison, 2012). In the case of normal distribution $\hat{v}(\hat{\mu})$ is equal to the residual sum of squares of the model. In this case, the Pearson goodness-of-fit test statistic and the deviance coincide. However, this is not the case with other distributions, deviance, and the Pearson χ^2 test statistic have different asymptotic properties and hence, they produce different results (Olsson, 2002; Allison, 2012).

3.3.2 Model selection and model checking

Model selection is a crucial step during statistical analysis. It involves the selection of the best model that fits data well from amongst the other competing models. One way to evaluate a model is to use the Information criterion(IC). In GLMs, two criteria are used during the model selection process, Akaike's information criterion(AIC) and Schwarz criterion(SC). Schwarz criterion is also known as the Bayesian Information Criterion (BIC)(Agresti, 2003, 2018).

Akaike's information criterion (AIC)

Akaike (1974), proposed this criterion as a very useful statistic for comparing the relative fit of different models, it is expressed as,

$$AIC = -2loglikelihood + 2p = -2l(\beta) + 2p$$

where $l(\beta)$ is the maximum likelihood function and p is the number of parameters in the model. A model with smaller AIC is preferred (Lindsey, 2000).

The Schwarz information criteria

Schwarz et al. (1978) proposed an alternative method that is used during the model selection process. The SC or BIC method mainly focuses on the asymptotic behaviour of the Bayes estimators, and it takes into account the sample size (Schwarz et al., 1978). According to Schwarz et al. (1978) BIC or SC is expressed as

$$BIC = -2l(\beta) + plog(n)$$

where $l(\beta)$ is the maximum likelihood function, p is the number of parameters in the model, and n is the sample size. According to Schwarz, the smaller the BIC, the better the overall model performance.

3.4 The Components of Generalized Linear Models

Components of GLMs and their functions are briefly discussed below (McCullagh, 2019; Olsson, 2002; Dobson & Barnett, 2008): The logistic regression model is one of the examples of a well-known class of models, known as GLMs. Linear regression, ANOVA, Poisson regression, are also examples of GLMs. Generalized Linear Models are made up of three components called, the systematic component, random component and the tink function.

- **The random component**
  It specifies the probability distribution function of the response variable, Y. For example, binomial distribution for Y in the binary logistic regression. If Y is continuous then a normal distribution is used, but if Y is assuming to be binary, the suitable distribution is binomial, and if Y is poisson or negative binomial distribution (response variable are counts) the suitable distribution is for non-negative counts

- **The systematic component**
  It refers to the explanatory variables $(X_1 + X_2 + \cdots + X_p)$ as a combination of linear predictors, such that $\eta_i = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$, for $i = 1, 2, \ldots, n$

- **The link function**, η or $g(\mu)$
  It combines the systematic and the random component. If the response variable Y is continuous the link function is given by

  $$\eta_i = g(E(Y_i)) = E(Y_i) = E(Y_i)$$

  for $i = 1, 2, \ldots, n$. but, if Y is binary the suited link function is the logit link function given by

  $$\eta_i = logit(\pi_i) = log\left(\frac{\pi_i}{1 - \pi_i}\right)$$

  where, π_i is the probability bounded between 0 and 1 $(0 \leq \pi \leq 1)$

Looking at how the response variables depends on the explanatory variables in general and GLMs, we begin with the following examples (Dobson & Barnett, 2008),

Example 1: Simple linear regression model

Simple linear regression models express how the mean expected value of continuous response variable depends on a set of explanatory variables:

$$Y = \beta_0 + \beta_i X_i$$

$$E(Y) = \beta_0 + \beta_i X_i$$

- Random component: Y is a response variable and follows a normal distribution. Generally the random error part is assumed to follow a normal distribution with mean zero and constant variance σ^2 , i.e $\varepsilon \sim N(0, \sigma^2)$.
- Systematic component: Explanatory variables $X's$ can be continuous, discrete or both and are linear in the parameter $\beta_0 + \beta_i X_i$
- Link function: Identity link, $\eta = g(E(Y_i))$

Example 2: Binary logistic regression

- Random component: Y follows a binomial distribution
- Systematic component: Explanatory variables X's can be continuous, discrete or both and are linear in the parameter $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_p X_p$
- Link function: Logit, $\eta_i = log\left(\frac{\pi_i}{1-\pi_i}\right)$

3.5 Logistic Regression Model

The logistic regression model is one of the classical applications of the GLMs. Logistic regression models the relationship between the predictor and a categorical (binary) response variable (Allison, 2012; Dobson, 2002), for example in this study our response is defined as either a child is anaemic (Hg<11 g/dl) or non-anaemic (Hb$\geq$ 11 g/dl).

The purpose of logistic regression, linear regression can not be used to deal with the categorical response variable. Many regression techniques exist in the literature for analyzing data with a categorical response variable, including logistic regression and discriminant analysis. Logistic regression is often used rather than discriminant analysis when the data consist of only two variables. Logistic regression could be easily applied in statistical soft wares rather than discriminant analysis when there is a mixture of numerical and categorical predictor variables because procedures involving dummy variables are automatically generated, with fewer assumptions and is more statistically robust (Dobson & Barnett, 2008).

Types of logistic regression applied based on the nature of the categorical response variable are as follows:

- Binary logistic regression: is only applicable, when the response variable is binary(dichotomous).

- Nominal logistic regression: is applied when the response variable consists of three or more categories that have no natural ordering.
- Ordinal logistic regression: is also applicable when the response variable has three or more categories but in this case order matters.

The logistic regression model assumptions are as follows:

- Binary logistic regression: is only applicable when the response variable is binary (dichotomous).
- The response variable should be made of two or more categories.
- The predictor variable need not be an interval, nor normally distributed, nor linearly related and should not necessarily be of equal variance within each group.
- The categories (groups) must be disjoint.
- Larger sample sizes are required, unlike in linear regression because maximum likelihood coefficients are large sample estimates. A minimum of 50 subjects is recommended.

Simple logistic regression model

The residuals from the logistic model are assumed to have a binomial distribution since the outcome variable of interest is binary (Nelder & Wedderburn, 1972). Therefore, for a single predictor, the simple logistic regression model is used. The response variable is categorical with two levels (dichotomous response variable). Y is said to be Bernoulli distributed, where the probability distribution function is given by:

$$P(Y = y) = \pi(x)^y (1 - \pi(x))^{1-y}$$

where $\pi(x)$ is the probability of obtaining a success (event of interest), which is bounded between 0 and 1, And, $1 - \pi(x)$ is the probability of failure; for example, in this study, it is a probability of finding that a child is not anaemic. In this study, our response variable Y is binary:

$$Y = \begin{cases} 1, & \text{Child is anaemic} \\ 0, & \text{Child is not anaemic} \end{cases}$$

In linear regression, the mean $E(Y)$ is modelled, but in logistic regression, the probability is modelled as a function of the predictor variables (Olsson, 2002). Therefore,

with the logit transformation (link function), the logistic regression model is well behaved, where $\pi(x)$ is between 0 and 1. Unlike in linear regression, the relationship between $\pi(x)$ and X is non-linear. Thus, the errors can be modelled using the binomial distribution.

Multiple logistic regression model

Similar to multiple linear regression, multiple logistic regression extend the simple logistic regression in the same manner. It is applicable if we have more than one predictor variable. Here, we have a set of predictors and hence a set of model parameters, given as (McCullagh, 1989):
$\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})^T$
$X_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip-1})^T$
And, $X_i^T\beta = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip-1}$

Now we are modelling the logit function η_i against the linear predictor:

$$\eta_i = logit(\pi(x_i)) = {X_i}^T\beta. \tag{3.29}$$

Then solving for $\pi(x_i)$, it yields

$$\pi(x_i) = \frac{exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip-1})}{1 + exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip-1})}. \tag{3.30}$$

where, $\pi(x)$ probability of success and $1-\pi(x)$ is the probability of failure. $X_1, X_2, \ldots, X_p$ is the set of explanatory variables, β_1 is the intercept constant and $\beta_1, \ldots \beta_p$ is the set of unknown parameters, one for each predictor variable (X). These unknown parameters are usually estimated by the method of maximum likelihood. The errors are no longer normally distributed; they are now binomially distributed, thus the OLS method for estimation of the coefficients is no longer an appropriate one. The shorthand of the multiple logistic model (3.30) is:

$$\pi(x) = \frac{exp(X_i^T\beta)}{1 + exp(X_i^T\beta)}. \tag{3.31}$$

3.6 Estimation of Model Parameters Using Maximum Likelihood

Maximum likelihood method is the procedure of obtaining one or more parameters (coefficients) for a given statistic which maximizes the known likelihood distribution. It is used to estimate and draw conclusions about the parameters of the

model. Assuming that the outcomes are random and independent variables the maximum likelihood function is defined as the product of the joint probability distribution (Allison, 2012; Wood, 2017). For anaemia status, the response variables Y_i, for $(i = 1, 2, 3, \ldots n)$ is dichotomous and thus, follow a Bernoulli distribution $(Y_i \sim Benoulli(\pi_i))$. The probability distribution function of Y_i is given by :

$$P(Y_i = y_i) = f(y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$$

for $i = 1, 2, 3, \ldots, n$
And since the $Y_i's$ are assumed to be independent, the likelihood function is defined as

$$\begin{aligned} L(\pi_i, y_i) &= \prod_{i=1}^{n} f(y_i) \\ &= \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}. \end{aligned} \tag{3.32}$$

taking the natural log on 3.32

$$\begin{aligned} l = log_e(L) &= log_e\left(\prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}\right), \\ &= \sum_{i=1}^{n} \{y_i ln(\pi_i) + (1 - y_i) ln(1 - \pi_i)\}. \end{aligned} \tag{3.33}$$

but, $\pi_i = \frac{exp(\beta_0 + X_{ij}^T \beta)}{1 + exp(\beta_0 + X_{ij}^T \beta}$ and $1 - \pi_i = \frac{1}{1 + exp(\beta_0 + X_{ij}^T \beta}$.

Therefore, substituting π_i back to the log-likelihood function it yields:

$$\begin{aligned} l = ln(\pi_i, y_i) &= \sum_{i=1}^{n} \left[y_i ln\left(\frac{exp(\beta_0 + X_{ij}^T \beta)}{1 + exp(1 + X_{ij}^T \beta)}\right) + (1 - y_i) ln\left(\frac{1}{1 + exp(\beta_0 + X_{ij}^T \beta)}\right)\right], \\ &= \sum_{n}^{n} \left[y_i ln\left(\frac{exp(\beta_0 + X_{ij}^T \beta)}{1 + exp(\beta_0 + X_{ij}^T \beta)}\right) + (1 - y_i) ln(1 + exp(\beta_0 + X_{ij}^T \beta))^{-1}\right], \\ &= \sum_{n}^{n} \left[y_i ln\left(\frac{exp(\beta_0 + X_{ij}^T \beta)}{1 + exp(\beta_0 + X_{ij}^T \beta)}\right) - y_i ln(1 + exp(\beta_0 + X_{ij}^T \beta))^{-1} + \left(1 + exp(\beta_0 + X_{ij}^T \beta)^{-1}\right)\right. \end{aligned} \tag{3.34}$$

The maximum likelihood estimates can be obtained by differentiating the score function $\bigcup_{\beta_j}$, (for $j = 0, 1, 2, \ldots, p$) with respect to β_0 and after with β_j and after equate it to a zero. Where $\bigcup_{\beta_j}$ is given by:

$$\bigcup = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \left(y_i x_{ij} - x_{ij} \frac{exp(\beta_0 + X_{ij}^T \beta)}{1 + exp(\beta_0 + X_{ij}^T \beta)} \right) = \sum_{i=1}^{n} (y_i x_{ij} - \pi_i x_{ij}).$$

and

$$\frac{\partial l}{\partial \beta_0} = \frac{\partial l}{\partial \beta_0} \sum_{i=1}^{n} \left(y_i - \frac{exp(\beta_0 + X_{ij}^T \beta)}{1 + exp(1 + X_{ij}^T \beta)} \right).$$

These equations can be solved iteratively using the Newton-Raphson or by the Fisher's scoring methods because they are non-linear likelihood equations. These algorithms are available in statistical software such as SAS, R, and STATA. Several statistical software, including SAS, use the Fisher scoring algorithm as a default iterative technique. Then the maximum likelihood estimators of $\theta = (\beta_1 \beta_2 \dots \beta_p)$. can be obtained by solving for $\bigcup(\theta)_{\beta_0} = 0$ and $\bigcup(\theta)_{\beta_j} = 0$, (Wood, 2017).

3.6.1 Hosmer and Lemeshow goodness-of-fit test

After fitting the logistic regression model we do not quickly make inferences or predict future outcomes, but we have to, check as far as possible, that the model we have assumed is correctly specified. We want to verify if the probabilities from the output show the true outcome of interest in the data and this is known as the goodness-of-fit test (Hosmer et al., 1997). For binary logistic regression, it is the most popular modelling approach. However, general linear regression use Pearson chi-squared and deviance goodness-of-fit test to assess for the goodness of fit of the model (Hosmer Jr et al., 2013).

The logistic regression model, with Y as the binary response variable with covariates $X_1, X_2, \dots, X_p$, assumes that for,

$\pi(x) = P(Y = 1|X_1, X_2, \dots, X_p)$ the logit $(\pi(x))$ is given as:

$$logit(\pi(x)) = log\left(\frac{\pi(x)}{1 - pi(x)}\right) = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p.$$

The Hosmer and Lemeshow test involves groupings (a minimum of 10 groups), and the expected and observed number of events in each group. Specifically, the groupings are based on the estimated parameter values $\beta_(1), \beta_2, \dots, \beta_p$, for each observation in the group and the probability that an event will occur is calculated , based on each observations covariates values (Hosmer Jr et al., 2013):

$$\hat{\pi}(x) = \frac{\hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_p X_p}{1 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_p X_p}.$$

The Hosmer and Lemeshow test statistic is based on the following hypothesis:

$$H_0 : \text{The model is satisfactory to fit the data}$$
$$H_1 : \text{The model is not satisfactory to fit the data}$$

and the test statistic is given by:

$$\chi^2_{HL} = \sum_{j=1}^{g} \sum_{i=1}^{1} \frac{(O_{ij} - E_{ij})^2}{E_{ij.}} \tag{3.35}$$

Equation 3.35 is chi-squared distributed with g-2 degrees of freedom. where g is the number of groups, O_{ij}: is the observed frequency of the outcome variable $(y = 1)$ or $(y = 0)$ in the j^{th} group, and E_{ij}: is the expected frequency of the outcome variable $(y = 1)$ or $(y = 0)$ in the j^{th} group.

Decision rules:
According to Hosmer et al. (1997), a model is said to be the best model for the data if the computed p-value of the test statistic with $g - 2$ degrees of freedom is more than 0.05 (p>0.05); if not, the model is said to be the poor model for the data. It is not necessarily true that if the p-value for the test statistic is large, then the mean model fits the data well, but it implies that there is a lack of evidence to reject the null hypothesis in favour of the alternative hypothesis (Hosmer Jr et al., 2013).

3.7 Model Interpretation and Inferencing

In logistic regression modelling, one can conclude the significance and importance of each predictor variable in several ways. A Wald chi-square test can be used to test for the null hypothesis that a single coefficient is equal to zero (Heeringa et al., 2017), $H_0 : \beta_j = 0$, or for an even more complex hypothesis concerning multiple parameters in the fitted model. A confidence interval can be as well used to draw inferences concerning the significance of model predictors and to give information on the potential magnitude and uncertainty associated with the estimated effect of individual predictor variables.

A design based confidence interval for logistic regression ($j = 0, 1, 2 \ldots, p$) parameter is computed as follows, (Heeringa et al., 2017)

$$Cl_{1-\alpha} = \hat{\beta}_j \pm t_{df,1-\alpha} SE\beta_j$$

where $\alpha = 0.05$ is typically used among the design- based degrees of freedom. where

df=number of parameters in a model minus a one. The resulting is called a 95% confidence interval for the parameter. Inferences regarding the significance of predictor variables can be performed directly for β_j on the log-odds scale. When considering only one predictor variable in a logistic regression model, an estimate of the odds ratio corresponding to a one-unit increase in the value of the predictor variable can be computed by exponentiation of the estimated logistic regression coefficient:

$$\hat{\psi} = exp(\hat{\beta})$$

If we consider the effect of only one predictor variable (i.e equation consists of only one predictor variable), the resulting is an estimate of unadjusted odds ratio or so-called crude odds ratio (Allison, 2012). However, if the fitted model is made up of more than one predictor variable, that is

$$logit(\pi(x)) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i.$$

the resulting $\psi|\beta_{k\neq j}$ is an adjusted odds ratio (Sedgwick, 2013). In general, the adjusted odds ratio is the representative of the multiplicative impact of one unit increase in the predictor variable X_j while all other predictors in the model are kept constant. Furthermore, the confidence interval can be adjusted for adjusted odds ratios. The logit model is unlike the linear regression model; the basic problem is that the logit model assumes a nonlinear relationship between the probability and the explanatory variables and the interpretation is not straight-forward. Logistic regression is easily understandable when the coefficients are interpreted in terms of odds or odds ratios (Allison, 2012)

Odds and odds ratios

According to Allison (Allison (2012)), to appreciate the logistic model, it is helpful for a person to have understood the concept of odds and odds ratios (denoted by OR). Odds ratios are widely used by professional gamblers and are defined as the chance, or likelihood, of an event to take place. While odds are defined as the expected number of times an event will occur to the number of times it will not occur. Odds are defined as follows:

$$O = \frac{\pi(x)}{1-\pi(x)} = \frac{\text{probability of occurence}}{1 - \text{probability of occurence}}$$

Odds have a lower boundary of zero like probability, but unlike probability, odds have no upper boundary, $0 \leq$ odds $<\infty$. When looking at odds ratios, OR >1 corre-

sponds to a probability more than 0.5 and OR$<$ 1 corresponds to a probability less of than 0.5. Odds ratios are mathematically defined as the ratio of odds and given as follows:

$$OR = \frac{\frac{\pi_1(x)}{1-\pi_1(x)}}{\frac{\pi_2(x)}{1-\pi_2(x)}}$$

Hosmer Jr et al. (2013), defined the odds ratios as a measure of association which is found to be useful especially in epidemiology, because it approximates the likeliness or the unlikeness of the event occuring, OR=1 implies no association between the exposure and outcome.

3.8 Survey Logistic Regression

Introduction

Statistical texts and software packages used by researchers in various fields, including social sciences, health sciences, and other applied fields, are implemented methods of analysis based on the assumption of independent identically distributed data. In most cases, researchers do not discuss the difficulties resulting from analyzing data collected from complex survey designs, that is, sampling involving clustering, stratification and unequal probability selection. However, the application of classical statistical methods to complex survey data, without first dealing with the stratification, clustering and the problem of unequal probability selection accordingly can lead to false conclusions (Heeringa et al., 2017).

In particular, ignoring the design of the data can lead to a serious underestimation of standard errors of parameter estimates and the associated confidence interval convergence rates, and could also inflate test levels and result in misleading model diagnostics (Roberts et al., 1987; An et al., 2002; Lee et al., 2013). Thus, it is highly advisable to take into consideration the nature of the data before any model fitting. A Survey logistic regression model is very similar to the ordinary logistic regression in the SAS system (Hosmer & Lemeshow, 2000). When dealing with data that is from a complex survey design, survey logistic regression has a high capability fitting such data because of its ability to take into account of complex sample surveys (Hosmer & Lemeshow, 2000; Heeringa et al., 2017). Thus, ordinary logistic regression modelling would not be appropriate to fit the DHS data set.

Furthermore, both the models (ordinary logistic and survey logistic regression models) approaches has the same theory (Heeringa et al., 2017; Hosmer & Lemeshow, 2000; Roberts et al., 1987). The difference between the two approaches is that or-

dinary logistic regression assumes that the data are collected using simple random sampling while in real-life that is not necessarily true (Hosmer & Lemeshow, 2000; Heeringa et al., 2017). 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 (Roberts et al., 1987). Dividing the population into strata could reduce the variance of the estimator of a population total (Heeringa et al., 2017; Hosmer & Lemeshow, 2000; An et al., 2002). The section below discusses the methods of parameter estimation for survey logistic regression models.

3.8.1 Estimation of survey logistic model parameters and the standard errors

For a simple logistic regression the maximum likelihood method is used to estimate the unknown regression parameters as explained in the previous sections. However, a straightforward application of the maximum likelihood method fails when data is collected from complex sample survey designs. The procedure is no longer possible for several reasons (Heeringa et al., 2017). For example: Firstly, in the case of complex survey data, the probability of selection and responding for the $i = 1, 2, 3, \ldots, n$ sample observations are no longer necessarily equal. Secondly, when the assumption of independence of observations which is very crucial to the standard maximum likelihood estimation (MLE) is violated, approach to estimating the sample variances of the model parameters and choosing a reference distribution for the likelihood ratio test statistic Heeringa et al. (2017).

There are two general ways that have been developed to estimate the logistic regression model unknown parameters and standard errors for a survey sample data. The first method of estimation was developed by Grizzle et al. (1969), (Heeringa et al., 2017). This approach is based on a weighted least squares (WLS) method of estimation. Years after this method was proposed, Binder (1983) proposed a second method for fitting logistic regression models and other generalized linear models with complex survey settings. A Pseudo-maximum likelihood estimation (PLME) was proposed as a technique for estimating the model parameters. The PMLE approach was combined with a linearised estimator of the variance-covariance matrix for the parameter taking into account the complexity of sample design (Roberts et al., 1987). The PMLE technique was further presented by Rao et al. (1989). To date, the PMLE approach is the standard method for estimating parameters in logistic regression models in all major software systems that support the analysis of complex

survey data (Heeringa et al., 2017).

3.8.2 The Pseudo maximum likelihood estimate method for estimating the unknown model parameters

For Y, the binary dependent variable (Heeringa et al., 2017), the population likelihood can be defined as

$$L(\beta|x) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

where $\pi(x_i)$ under a logit link it is evaluated using the logistic CDF and the parameters, specifying the logistic regression model.
The estimate of the finite population regression parameters is obtained by maximizing population likelihood that follows, which is the function of the weighted observed sample data and the $\pi(x_i)$ values:

$$\prod_{i=1}^{N} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \tag{3.36}$$

The pseudo maximum likelihood estimators can be solved by the application of either the Fisher scoring algorithm or the Newton-Raphson algorithm, available in SAS statistical software. Variances of the regression parameters and odds ratios are computed the using Taylor series expansion approximation, (Binder, 1983; Morel, 1989). A PROC SURVEYLOGISTIC regression method available in the SAS and the R-software would be used instead of simply the PROC LOGISTIC method. An advantage of PROC SURVEY LOGISTIC over PROC LOGISTIC method, it allows one to specify class variables as explanatory in the model by using the same syntax for main effect and interaction. The PROC SURVEYLOGISTIC method was designed to handle complex sample survey data with unequal weights, clustering and stratification.

Similar to the ordinary logistic regression model, where data is assumed to be collected using simple random sampling. For Y, the binary dependent variable and binomial data likelihood, the maximization of the pseudo-likelihood approach to the logistic regression parameters and their variance-covariance matrix requires the solution to the vector of estimating equations that follows (Heeringa et al., 2017),

$$S(\beta) = \sum_{h}\sum_{\alpha}\sum_{i} W_{h\alpha i} D'_{h\alpha i} \left[(\pi_{h\alpha i}(\beta))\left(1-\pi_{h\alpha i}(\beta)\right)\right]^{-1}. \tag{3.37}$$

where, $D_{h\alpha i}$ is the vector of partial derivatives,

$$\frac{\partial(\pi_{h\alpha i}(\beta))}{\partial(\beta_j)}$$

, and $j = 0, 1, 2 \ldots, p$. h, is a stratum index. α, is a cluster (SECU) index within stratum h and i is an index for individual observations within cluster α. $\pi_{h\alpha i}(\beta)$ is the probability that the outcome variable is equal to 1 as a function of the parameter estimate and observed data according to the stratified logistic model (Heeringa et al., 2017). $W_{h\alpha i}(\beta)$,is the sampling weight for observation i.

For Y, logistic regression model of binary variable these results to a system of $p + 1$ equations. Where p is the number of predictor variables and there is one additional parameter corresponding to the intercept,

$$S(\beta)_{logistic} = \sum_h \sum_\alpha \sum_i W_{h\alpha i} \left(y_{h\alpha i} - \pi_{h\alpha i}(\beta)\right) x'_{h\alpha i} = 0, \tag{3.38}$$

where, $x_h \alpha i$ is a column vector of p+1 design matrix elements for case $i = [1, x_{1,h\alpha i}, \ldots x_{p,h\alpha i}]'$. For the probit regression model the estimating equations reduces to,

$$S(\beta)_{logistic} = \sum_h \sum_\alpha \sum_i \frac{W_{h\alpha i} \left(y_{h\alpha i} \pi_{h\alpha i}(\beta)\right) \phi x'_{h\alpha i} \beta * x'_{h\alpha i}}{\left(\pi_{h\alpha i}(\beta)\right) \left(1 - \pi_{h\alpha i}(\beta)\right)}, \tag{3.39}$$

where $\phi_{h\alpha i}$ is the standard normal probability density function evaluated at $x'_h \alpha i \beta$. The weighted parameter estimates are computed by using the Newton-Raphson method to derive a solution for $S(\beta) = 0$ (Sloane & Morgan, 1996). The vector of weighted parameter estimates based on the pseudo-maximum likelihood estimate is consistent for β even when the sample design is complex. Thus the bias of the estimator is of the order $\frac{1}{n}$, such that, as the sample size increases (which is the often the case for survey data), the bias of the estimator approaches zero.

3.8.3 Maximization of the pseudo- likelihood function

According to Heeringa et al. (2017), if we let $\pi_{kjih} = p(y_{kjih} = 1)$ denote the probability that an event will occur on h^{th} individual within i^{th} household, within j^{th} primary sample units nested within k^{th} stratum. Thus, $1 - \pi_{kjih} = p(y_{kjih} = 0)$ will denote the probability of non-occurrence of the event on the h^{th} individual within i^{th} household, within j^{th} primary sample units nested within k^{th} stratum. Therefore the likelihood will be the product of individual contributions (Hosmer & Lemeshow, 2000). For a contribution of a single observation, a pseudo-maximum likelihood is

defined as,

$$\pi_{kjih}^{W_{kjih}Y_{kjih}}(1-\pi_{kjih})^{1-W_{kjih}Y_{kjih}}$$

Thus a pseudo-likelihood function is given by

$$L(\beta,Y)=\prod_{k=1}^{k}\prod_{j=1}^{m_k}\prod_{i=1}^{n_{kj}}\prod_{h=1}^{H_{kji}}\pi_{kjih}^{W_{kjih}Y_{kjih}}(1-\pi_{kjih})^{1-W_{kjih}Y_{kjih}}. \tag{3.40}$$

Applying a natural log-function to the log-pseudo-maximum likelihood function will be given as follows:

$$l(\beta,Y)=\sum_{k=1}^{k}\sum_{j=1}^{m_k}\sum_{i=1}^{n_{kj}}\sum_{h=1}^{H_{kji}}W_{kjih}Y_{kjih}log\left(\frac{\pi_{kjih}}{1-\pi_{kjih}}\right)-log\left(\frac{1}{1-\pi_{kjih}}\right). \tag{3.41}$$

Differentiating $l(\beta,Y)$ with respect to unknown regression coefficients, we obtain $p+1$ vector score equations which can be written as

$$X'W(y-\pi)=0 \tag{3.42}$$

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

$$logit(\pi_{kjih})=log\left(\frac{\pi_{kjih}}{1-\pi_{kjih}}\right)=X'_{kjih}\beta \tag{3.43}$$

X_{kjih} is the vector that correspond to the characteristic of the h^{th} individual within i^{th} household, within j^{th} primary sample units nested with k^{th} stratum and β is the vector of the unknown regression coefficients.

Variance estimation

For data with unequal weights, clustering and stratification (complex survey sample design), the computation of standard error of parameter estimates used in the construction of confidence interval and performing statistical test is very complicated. A solution to this problem was proposed by Binder (1983), who applied a multivariate version of the Taylor series expansion (TSL). The result is a sandwitch-type variance estimator of the form

$$Var(\hat{\beta})=J^{-1}var\left[S(\hat{\beta})\right]J. \tag{3.44}$$

where J is a matrix of the second derivative with respect to $\hat{\beta}_j$ of the pseudo-log-likelihood, and $var[s(\hat{\beta})]$ is the variance-covariance matrix for the sample totals of weighted score functions for each observation to fit the model (Heeringa et al., 2017). The application of the Binder linearized variance estimation, $var[s(\hat{\beta})]$, can be found in Appendix 6.

3.9 Assessing the Model

Goodness-of-fit

In an ordinary logistic regression model, the responses (observations) are assumed to be independent and identically distributed. However, in complex survey design (e.g. survey logistic regression models) that is not really the case, there are higher chances that observations that are from the same cluster are more correlated compared to observations from different clusters. Thus, the goodness of fit test should be considered regarding the design of the study. The Hosmer and Lemeshow goodness of fit test was originally proposed for ordinary logistic regression, but Archer & Lemeshow (2006) and Archer et al. (2007) extended it to avoid possible problems associated with the asymptotic distribution of the chi-square tests. The method is based on grouping the observations in "deciles of risk" (Roberts & Matthews, 2016), where the observations are partitioned into ten equal-sized groups based on their ordered estimated probabilities, $\hat{\pi}_i$. The Hosmer-Lemeshow test statistic is given by

$$\hat{C} = \sum_{l=1}^{10} \frac{(O_l - E_l)}{E_l \left(1 - \frac{E_l}{n_l}\right)}, \tag{3.45}$$

where,n_l is the number of observations in the l^th. and $O_l = \sum_i y_i$ and $E_l = \sum_i \hat{\pi}$ are respectively the observed and expected number of cases in the l^th decile.

The test statistics (3.45) follows a chi-square distribution (Hosmer & Lemesbow, 1980). The extension of this method Archer & Lemeshow (2006) is called the F-adjusted mean residual test, which is asometimes called the Archer and Lemeshow goodness-of-fit test ad it can be estimated as follows.

Suppose the study is designed such that there are m PSUs (clusters) each containing a total of n_i observations. Using a fitted survey logistic regression model, the residual for the j^{th} observation in thei^{th} PSU is calculated as follows:

$$\hat{r} = y_{ij} - \hat{x_{ij}} \tag{3.46}$$

This grouping strategy was proposde by Graubard et al. (1997) the observations are grouped into deciles of risk according to their residuals and weights (Archer & Lemeshow, 2006). The size of the first decile group will be equal to number of observations with the smallest residuals such that the sum of the corresponding weights represent one tenth of the total weights of all the observations. In a similar manner, the size of the rest of the decile groups can be calculated. The mean residuals by decile of risk $\hat{M}' = (\hat{M}_1, \hat{M}_2, \ldots, \hat{M}_{10})$ are obtained where

$$\hat{M}_g = \frac{\sum_i \sum_j w_{ij} \hat{r_{ij}}}{\sum_i \sum_j w_{ij}} \tag{3.47}$$

is the mean residual for the g^th percentile of the weighted residual values for $g = 1, \ldots, 10$ and w_{ij} is the sampling weight associated with observation y_{ij} .

The Wald test statistic for testing g categories is given by

$$\hat{W} = \hat{m}'[\hat{var}(\hat{M})]^{-1}\hat{M}. \tag{3.48}$$

Where $\hat{var}(\hat{M})$ is the variance-covariance matrix of $\hat{M}$ obtained using linearization (Archer et al., 2007). This test statistic is approximately chi-squared distributed with $g - 1 = 9$ degrees of freedom, since $g = 10$ in this case. However, this chi-square test has been found to be not an appropriate reference distribution. Therefore, the f-corrected Wald test statistic has been suggested instead (Archer & Lemeshow, 2006). This test is given by

$$F = \frac{(f - g + 2)}{fg} W \tag{3.49}$$

is approximately F-distributed with $g-1$ numerator degrees of freedom and $f-g+2$ denominator degrees of freedom, where f is the number of clusters in the sample minus the number of strata and g is the number of categories. Therefore, based on this test statistic, the F-adjusted mean residual test statistic is

$$\hat{Q}_m = \frac{f - 8}{10f} \hat{M}'[\hat{var}(\hat{M})]^{-1}\hat{M} \tag{3.50}$$

as $g = 10$ deciles of risk. For further reading on the goodness of fit test for complex survey design refer to work by Archer & Lemeshow (2006) and Archer et al. (2007).

Model selection

The ordinary logistic regression (PROC LOGISTIC) model uses the forward; backward elimination and stepwise selection procedures to select the variables that fit

the data set in the SAS environment. Unfortunately, for complex survey designs, these steps are not yet implemented in the PROC SURVEY LOGISTIC. However, for complex survey designs Hosmer & Lemesbow (1980) suggested the following three steps that could be used for model selection.

- Perform univariate analysis between the dependent variable and the independent variables one at a time. This can be done through a contingency table of the outcome and the nominal or ordinal predictor variable or through fitting a univariate survey logistic regression.
- Discard non-significant variables and consider only significant ones in the multivariate survey logistic regression, in addition to the other variables known from the literature to be important when modelling certain outcome.
- Include the relevant predictors in the multivariate survey logistic regression analysis, one at a time.

The importance of each variable is verified through the Wald chi-square statistic and also by comparing it to the estimated coefficient with the one from the univariate analysis. This is repeated until only the significant variables are left in the model. Furthermore, interaction terms are then included amongst the variables in the model. In addition, the AIC and the BIC discussed in Section 3.3.2 are also important measures that can be used to compare two nested models when determining the better one that describes the data set.

3.10 Survey Logistic Regression Applied to TDHS and ADHS Data Sets

In this thesis, SAS 9.4 was used to fit SLR model into the two data sets (TDHS and ADHS) using PROC SURVEYLOGISTIC. By default, the PROC SURVEYLOGISTIC method in SAS uses the Taylor series expansion to estimate the variance of the SLR model.

The study variables as discussed in Chapter 2 Section 2.2 were used to model the outcome variable (anaemia status). The sampling weights were adjusted for non-response and to represent only those households included in the data set used in this thesis, where only the households that had children under the age of five years old tested for anaemia were included in the sample. The two data sets (TDHS and ADHS) were fitted separately and a two-way interaction term was considered. However, none of the possible two-way interaction was found to have a significant effect;

hence, it was discarded.

After performing all the steps discussed in Section 3.9 the following tables (Table 3.1 and Table 3.3) are the results for univariate analysis. These tables consist of the p-values from (type 3 analysis, SAS output), the unadjusted odds ratios and 95% confidence interval which was used to determine the significance of the variable and if the variable would be considered for multivariate analysis.

Table 3.1: Unadjusted odds ratios (OR) from the univariate survey logistic regression analysis (TDHS)

Variable	P-value	Odds ratio (95% CI)
Child's age(in months)	< 0.0001	
0 - 11		5.20(4.22,6.42)
11-23		3.81(3.23,4.50)
24-35		1.83(1.56,2.15)
36-47		1.08(0.90,1.29)
Gender	0.0104	
Female		0.88(0.79,0.97)
Type of place of residence	0.0033	
Rural		1.23(1.07,1.42)
Highest education level	<0.0001	
No education vs secondary		1.58(1.29,1.92)
Primary vs secondary		1.05(0.89,1.23)
Higher vs secondary		0.66(0.40,1.11)
Wealth index	0.0006	
Middle vs rich		1.22(1.03,1.44)
Poor vs rich		1.35(1.56,1.58)
stunting	<0.0001	
Moderate vs severe		0.79(0.634,0.98)
Normal vs severe		0.67(0.55,0.82)
Internet	0.0002	
Have access to internet vs no access		0.61(0.47,0.80)
Under five slept under mosquito bed next last night	0.1865	
No ve yes		0.92(0.80,1.03)
Gave child meat(Pork,beef,lamb, etc)	0.6955	
No vs yes		0.95(0.73,1.24)
Had cough the last two weeks prior to survey	0.0958	
No vs yes		0.89 (0.77,1.02)
Iron supplementation during pregnancy	0.2414	
No vs yes		1.06 (0.90,1.26)
Size of child at birth	0.9530	
Average vs small		0.99 (0.85,1.15)
Large vs small		0.97 (0.77,1.21)

Table 3.2: TDHS univariate continues from the previous page

Variable	p-value	Odds ratio (95% CI)
Age of the household head	0.0370	1.004 (1.000, 1.008)
Currently breast feeding	<0.0001	
No vs yes		0.51 (0.46,0.57)
Type of mosquito bed nets	0.0097	
No nets vs only untreated nets		1.15 (0.91,1.45)
only treated nets vs only untreated nets		1.34 (1.07,1.68)
Wasting	0.0009	
moderate vs severe		0.53 (0.35,0.81)
normal vs onlysevere		0.48 (0.33,0.72)
Underweight	0.0071	
moderate vs severe		0.53 (0.35,0.81)
normal vs onlysevere		0.48 (0.33,0.72)
Vitamin A supplementation	0.0271	
No vs yes		1.14 (1.01,1.28)
Sex of the household head	0.7578	
Female vs male		1.02(0.88,1.19)
Household has a television	0.0009	
no vs yes		1.35 (1.15,1.58)

Table 3.3: unadjusted odds ratios (OR) from the univariate survey logistic regression analysis (ADHS)

Variable	p-value	Odds ratio (95% CI)
Child's age (in months)	< 0.0001	
0-11		4.40 (3.21 6.04)
12-23		3.02(2.41 3.79)
24-35		1.70(1.36 2.11)
36-47		1.34 (1.05 1.70)
Gender	0.0370	
Female		0.84(0.72 0.99)
Type of place of residence	0.0033	
Rural		1.23(1.07,1.42)
Highest education level	(0.1444)	
No education vs secondary		1.21(0.94, 1.54)
Primary vs secondary		1.18(0.89,1.23)
Higher vs secondary		0.69(0.37, 1.28)
Wealth index	0.3400	
Middle vs rich		1.20(0.92 ,1.56)
Poor vs rich		1.02 (0.81, 1.28)
stunting	<0.0001	
Moderate vs severe		0.65 (0.49, 0.84)
Normal vs severe		0.58 (0.46, 0.73)
Internet	0.9430	
Have access to internet vs no access		1.01 (0.73, 1.39)
Under five slept under mosquito bed next last night	0.1865	
No ve yes		0.92(0.80,1.03)
Gave child meat(Pork,beef,lamb, etc)	0.9856	
No vs yes		1.00 (0.76 1.32)
Had cough the last two weeks prior to survey	<0.0001	
No vs yes		0.89 (0.77,1.02)
Iron supplementation during pregnancy	<.0001	
No vs yes		1.14(0.90, 1.45)
Size of child at birth	0.9530	
Avarage vs small		0.99 (0.85,1.15)
Large vs small		0.97 (0.77,1.21)

Table 3.4: ADHS univariate analysis continues from the previous page

Variable	p-value	Odds ratio (95% CI)
Age of the household head	0.4672	0.99(0.99, 1.00)
Currently breast feeding(ref=yes)	<0.0001	
No		0.71 (0.62, 0.83)
Type of mosquito bed nets (ref=only untreated nets)	0.0097	
No nets		1.19(0.83 1.72)
Only treated nets		1.06 (0.71 ,1.56)
Wasting (ref=severe)	0.0012	
Moderate		0.67 (0.40,1.12)
Normal		0.54 (0.35 0.82)
Underweight(ref=severe)	0.1848	
Moderate		1.74(0.83 ,3.62)
Normal		1.18 (0.65,2.13)
Vitamin A supplementation(ref=yes)	0.0271	
No		1.14 (1.01,1.28)
Sex of the household head(ref= male)	0.6550	
Female		0.959 (0.80, 1.15)
Household has a television(ref=yes)	0.0009	
No		1.13 (0.93, 1.38)

All variables that were found to be significantly associated with childhood anaemia (p-value < 0.05) from the univariate analysis were considered for multivariate analysis. From the above tables we can observe that child age, type of place of residence, gender (sex of the child), stunting, , currently breastfeeding, type of mosquito bed nets, wasting, and household has a television were commonly significantly associated with childhood anaemia in both countries and hence were considered for multivariate analysis. However,variables named: highest education level, wealth index, age of household head, sex of household head, iron supplementation during pregnancy and gave child meat (pork, beef, lamb, etc) were not commonly significantly associated with anaemia but because of their importance when modelling anaemia they were considered for multivariate analysis. Thus, the final model fitted is as follows:

$$g(\mu) = \beta_0 + \beta_1(child \quad gender) + \beta_2(residence \quad type) + \beta_3(Region) + \beta_4(Wealth \quad index) + \beta_5(Currently \quad breasfeeding) + \beta_6(Stunting) + \beta_7(Television) + \beta_8(Education \quad Level) + \beta_9(Child \quad age) + \beta_{10}(household \quad head \quad age) + \varepsilon_i. \tag{3.51}$$

The final multivariate survey logistic regression (SLR) analysis

The tables below (Table 3.5, Table 3.6 and Table 3.7) are the outputs from the final multivariate survey logistic regression analysis. The variable named, age of household head, was fitted as a continuous variable, whereas the rest of the variables are categorical. Table 3.5 is the type 3 analysis of effect; Table 3.6 and Table 3.7 are the combination of analysis of maximum likelihood estimates and odds ratio estimates tables from the SAS system.

Table 3.5: Type 3 Analysis of Effects (TDHS and ADHS)

	Tanzania		
Effect	**DF**	**Chi-square value**	**p-value**
Gender	1	4.34	0.0373
Child age(in months)	4	378.29	<.0001
Type of place of residence	1	1.25	0.2637
Wealth index	2	5.80	0.0550
Age of household head	1	0.26	0.6120
Currently breastfeeding	1	11.32	0.0008
Stunting	2	22.79	<.0001
Region	29	192.41	<.0001
Mother's highest education level	3	22.47	<.0001
Household has a television	2	0.91	0.6351
	Angola		
Gender	1	2.75	0.0971
Child age(in months)	4	147.58	<.0001
Type of place of residence	1	0.24	0.6274
Wealth index	2	11.56	0.0031
Age of household head	1	0.068	0.8037
Currently breastfeeding	1	0.60	0.4386
Stunting	2	23.87	<.0001
Region	17	86.07	<.0001
Mother's highest education level	3	6.99	0.0721
Household has a television	2	13.68	0.0011

Table 3.5 above shows the final survey logistic regression model after being fitted to the two data sets. The model reveals that the predictor variables, age of the child (in months), stunting and the region were the common determinants of childhood anaemia in both countries. Their p-values in both countries were all < 0.0001. Looking at each country, in Tanzania, sex of the child (p-value = 0.0373), currently

breastfeeding (p-value = 0.0008) and the mother's or guardian highest education level (p-value $<$ 0.0001) predictors were also found to be significantly associated with childhood anaemia at 5% level of significance. However, in Angola, there were two additional variables that were significantly associated with childhood anaemia, the standard of living or so-called wealth index (p-value = 0.0031) and the variable household has a television (p-value = 0.0011)

Table 3.6: Adjusted odds ratios from the final survey logistic regression model for the TDHS data set

Parameter	Estimate	Standard error	t-value	P-value	aOR(95% CI)
Intercept	0.5293	0.4221	1.25	0.2104	
Gender (ref=male)	-0.1232	0.0592	-2.08	0.0378	0.884(0.787, 0.993)
Child's age (in months) (ref = 48-59)					
0-11	1.6563	0.1259	13.16	<0001	5.240 (4.092, 6.710)
12-23	1.2567	0.0933	13.47	<0001	3.514 (2.925, 4.220)
24-35	0.5595	0.0908	6.16	<0001	1.750 (1.464,2.091)
36-47	0.0325	0.0954	0.34	0.7338	1.033 (0.856, 1.246)
Age of household head	0.00122	0.00240	0.51	0.6122	1.001 (0.997, 1.006)
Mother's highest education level (ref =No education)					
Higher	-2.5850	1.1026	-2.34	0.0194	0.075 (0.009, 0.658)
Primary	-0.3729	0.0894	-4.17	<.0001	0.689 (0.578, 0.821)
Secondary	-0.4248	0.1592	-2.67	0.0079	0.654(0.478,0.894)
Wealth Index (ref = rich)					
Middle	0.1108	0.0913	1.21	0.2259	1.117 (0.934,1.337)
Poor	0.2182	0.0906	2.41	0.0164	1.244 (1.041,1.486)
Residence type (ref = urban)					
Rural	0.1252	0.1120	1.12	0.2642	1.133 (0.909, 1.412)
Stunting (ref =severe)					
Moderate	-0.2114	0.1212	-1.74	0.0817	0.809 (0.638, 1.027)
Nourished	-0.4707	0.1131	-4.16	<.0001	0.625 (0.50,0 0.780)
Currently breast-feeding (ref=yes)					
No	-0.2473	0.0735	-3.36	0.0008	0.781 (0.676, 0.902)
Household has televison (ref=yes)					
No	-0.1374	0.1984	-0.69	0.4889	0.872 (0.590, 1.287)

Table 3.7: Adjusted odds ratios from the final survey logistic regression model for ADHS data set

Parameter	Estimate	Standard error	t-value	P-value	aOR(95% CI)
Intercept	1.0759	0.3337	3.22	0.0013	
Gender (ref=male)	-0.1442	0.0869	-1.66	0.0976	0.866(0.730,1.027)
Child's age (in months) (ref = 48-59)					
0-11	1.4986	0.1683	8.91	<.0001	4.475 (3.216,6.228)
24-35	1.1316	0.1206	9.39	<.0001	3.101 (2.447,3.929)
36-47	0.5088	0.1143	4.45	<.0001	1.663 (1.329,2.082)
48-59	0.2532	0.1269	1.99	0.0466	1.288 (1.004,1.653)
Age of household head	-0.000769	0.00396	0.19	0.8462	0.999 (0.993,1.005)
Mother highest education level (ref = No education)					
Higher	-0.5422	0.3354	-1.62	0.1065	0.581 (0.301,1.124)
Primary	-0.0607	0.1057	-0.57	0.5662	0.941 (0.765 1.158)
Secondary	-0.3289	0.1451	-2.27	0.0238	0.720 (0.541, 0.957)
Residence type (ref = urban)					
Rural	0.0675	0.1391	0.49	0.6276	1.070 (0.814,1.406)
Wealth index (ref = rich)					
Middle	-0.00564	0.1697	-0.03	0.9735	0.994 (0.713,1.388)
Poor	-0.5305	0.2130	-2.49	0.0131	0.588 (0.387,0.894)
Stunting (ref =severe)					
Moderate	-0.4142	0.1395	-2.97	0.0031	0.661 (0.502,0.869)
Nourished	-0.5528	0.1150	-4.81	<.0001	0.575 (0.459,0.721)
Currently breast feading (ref = Yes)					
No	-0.0661	0.0853	-0.77	0.4389	0.936 (0.792,1.107)
Household has television (ref = Yes)					
No	0.3425	0.1305	2.63	0.0089	1.409 (1.090,1.820)

From Table 3.6 and Table 3.7 the model reveals that children who were aged 0-11 months (< 1 year) had higher chances of being anaemic compared to all other age groups and that the chances of a child being anaemic decreases with the increase in age. Children aged less than one from Tanzania and Angola were respectively, 5.2 (OR = 5.24, 95% CI (4.09;6.71)) and 4.48 (OR = 4.475, 95% CI (3.22;6.23)) times more likely to be anaemic than children aged four. Those who were aged 12-23 months (< 2 years) in Tanzania and Angola were respectively, 3.51(OR=3.514, 95% CI(2.93;4.22)) and 3.10 (OR = 3.101, 95% CI (2.45;3.93)) times more likely to be anaemic compared to children aged 4 years. Childhood anaemia was significantly higher for children aged two compared to those aged four in Tanzania OR = 1.75, 95% CI (1.46,2.09) and in Angola OR = 1.663, 95% CI (1.33;2.08). Children who were aged three were 3.3% (OR = 1.033, 95% CI (0.86;1.25)) more likely to be anaemic in comparison to children aged four in Tanzania and in Angola they were 28.8% (OR = 1.288, 95% CI (1.004;1.65)) more likely to be anaemic compared to those aged four. Childhood anaemia was significantly low for female children in Tanzania (OR = 0.884, 95% CI (0.79;0.99)) compared to male children, but gender was not significantly associated (p-value = 0.055) with childhood anaemia in Angola according to to the magnitude of the p-value but looking at other effect sizes such as the odds ratios, we can tell that female children were also less likely to be anaemic compared to male children (OR = 0.866, 95% CI (0.73;1.03)),with an OR of less than one.

Severely stunted children were more likely to suffer from childhood anaemia compared to nourished (non-stunted) and moderately stunted children in both countries. In Tanzania, moderately stunted children had reduced odds of being anaemic by 19.1% (OR = 0.809) with a 95% CI (0.64;1.03) and nourished children had reduced by 37.5% (OR = 0.625, 95% CI (0.5;0.78)) in comparison to severely stunted children. In Angola, moderately stunted children were 33.9% (OR=0.661, 95% CI (0.50;0.87)) less likely to be anaemic and normal children were 42.5% (OR = 57.5, 95% CL (0.46;0.72)) less likely to be anaemic than severely stunted children. Children who were born in middle class and rich families in Angola had reduced odds by, 0.16% (OR = 0.994, 95% CI (0.713;1.39)) and 41.2% (OR=0.588, 95% CI (0.38;0.89)) respectively, compared to children from poor families. In Tanzania, wealth index was not significantly associated with anaemia (p-value = 0.055) but looking at the results, we can see that children from poor families are still more likely to be anaemic than children from moderate (OR=1.12, 95% CI (0.93;1.34)) and rich (OR = 1.24, 95% CI (1.04;1.49)) families. Furthermore, the level of education was also found to be a significantly associated with childhood anaemia in Tanzania, where, children who had educated mothers were less likely to be anaemic compared to those who had uneducated mothers; OR = 0.08 (95% CI (0.009;0.66)) for highly educated mothers;

OR = 0.65 (95% CI (0.480.89)) for mothers with secondary education and OR=0.69 (95% CI (0.58;0.82)) for mothers with primary education.

3.11 Summary and Discussion

In this chapter we covered generalized linear models, survey logistic regression was fitted to our data sets. GLMs are parametric regression models with underlying assumption of normality. At first, a univariate analysis was performed for each data set, and then significant effects were passed into the SLR model. The backward and forward variables selection steps for logistic regression were applied as suggested by Hosmer & Lemesbow (1980) and a final SLR model resulted as represented above.

The findings from this thesis are not much different from many other similar studies existing in the literature (Habyarimana et al., 2017; Allali et al., 2017; Foote et al., 2013; Schellenberg et al., 2003). As discussed above, the SLR model revealed that the explanatory variables, as listed: stunting, child age, and region were the common determinants (in both countries) of childhood anaemia. Children at a young age have a poor immune system, and hence they are easily affected by diseases (De Pee et al., 2002; Leal et al., 2011; Sanou & Ngnie-Teta, 2012). As we can observe from the two tables of the final SLR models, the odds for a child being anaemic decreas with the increase in age; children less than two years are seemingly to have very high odds of being anaemic compared to older children. These results are very similar to many studies conducted about predictors and prevalence of anaemia in Africa (Habyarimana et al., 2017; Allali et al., 2017; Foote et al., 2013; Schellenberg et al., 2003; de Savigny et al., 2003). Several studies have shown that children with educated mothers or caretakers are less likely suffer from curable diseases, such as: anaemia, fever, diarrhea, malaria infections, etc Habyarimana et al. (2017).

The odds of the child being anaemic decrease with the increase of the mother's (caretaker) level of education; children with uneducated mothers or with primary education are highly likely to be anaemic compared to others. These findings are consistency to many other existing studies in literature (Getaneh et al., 2017; Foote et al., 2013; Pektaş et al., 2015; Mehta, 2004). Level of education plays a major role in child health, nutrition, growth, and development. A caretaker lacking education may easily fail to understand the nutritional requirements and the recommended feeding practices. Another factor that is highly associated with childhood anaemia is stunting. Stunting (height for age) is defined as impaired growth and development that children experience from poor nutrition, repeated infection, and inadequate psychosocial stimulation. Although it globally affect children, there is high

prevalent in Africa (Habyarimana et al., 2017). In this study, stunting was found to be a common determinant of anaemia in both countries with severely stunted children having higher odds of being anaemic compared to moderately and non-stunted children.

Furthermore, the standard of living in a household was found to be significantly associated with childhood anaemia in Angola with children from poor homes being more likely to be anaemic compared to those from rich homes. This may be associated with the fact that rich children are more likely to have quality education and their parents can afford nutritious food to feed their children and maintain a healthy lifestyle. In both countries, the type of place of residence (rural/urban) is not significantly associated with childhood anaemia. Although this is the case, this study does show higher odds of being anaemic for children from rural compared to urban areas. This may be because most people residing in urban areas are wealthier, can afford a quality education and they can easily access hospitals for child health; hence, their children will have reduced odds of being anaemic. The next chapter focuses more on non-parametric (semi-parametric) models, as they are believed to be more precise compared to parametric models.

Chapter 4

Semi-Parametric Regression Models

4.1 Introduction

Thus far, we considered GLMs (ordinary logistic and survey logistic regression model approaches) which are known to be parametric. Generalized linear model regression approaches are used to linearly describe the effect of the covariates on the outcome variable of interest. Unlike non-parametric models, in parametric regression models, the functional form of the model is known in advance. For this reason, the results produced might be biased or misleading, thus non-parametric models that assume no functional form of the model prior to modelling would be more useful (Hastie, 1990).

We now consider the non-parametric regression that relaxes the assumption of linearity between the covariates and the response variable. The non-parametric (or semi-parametric) approach, specifically generalized additive mixed models (GAMM) were used to investigate the relationship between the effect of explanatory variables on childhood anaemia. Non-parametric regression models are the flexible statistical approaches for modelling non-linear forms of the data that have no functional forms prior to being defined (Zhang & Lin, 2003). Although parametric regression models are easy to compute and interpret, they are too restrictive compared to non-parametric regression approaches. Instead of using one of the two methods, the combination would be even more powerful, as would combining the two methods results in so-called semi-parametric additive mixed models (SAMMs) (Härdle et al., 2012; Zhang & Lin, 2003), which are a special case of GAMMs.

4.1.1 Model structure

Supposing we have a pair of n random data points Silverman (1985), $\{x_i,y_i\}$ for $i = 1, 2...n$, a semi-parametric model is given as,

$$Y_i = g(x_i) + \varepsilon_i. \tag{4.1}$$

As usual Y_i, is a vector of the response variable. g is the unknown regression function, which can be estimated by the roughness penalty method, as suggested by Green & Silverman (1993). In the literature there exist a number of non-parametric regression models and smoothing for dependent data; for example, kernel estimators, smoothing splines, running mean line smoothers, bin smoothers, wavelets and local weighted scatter plot smoothing (LOWESS) (Hastie, 1990; Härdle & Kneip, 1999; Green & Silverman, 1993). One such model is called a project pursuit regression model as suggested by Friedman & Stuetzle (1981). It fits the model of the form,

$$Y = \sum_{j=1}^{p} S_j(\hat{\alpha}_j X) + \varepsilon. \tag{4.2}$$

where $\hat{\alpha}_j X$ is a one-dimension projection of the vector X, S_j is the arbitrary smooth function and ε is the error term which is assumed to be identical and independent normally distributed with mean 0 and a constant variance σ^2. For large p, these models are difficult to interpret, although they are a parsimonious smooth surface (Hastie and Tibshirani, 1990).

Breiman & Friedman (1985) suggested an alternative conditional expectation which is also a non-parametric approach for estimating nonlinear regression, of the form,

$$Y = \sum_{j=1}^{p} S_j(x_j). \tag{4.3}$$

where $S_j$is an unspecified (arbitrary) non-parametric smooth function. The response variable Y is estimated as the transformation of the form $\theta(Y)$. Readers can find intensive literature about non-parametric regression approaches in the work by (Silverman, 1985; Izenman, 1991; Faraway, 2016).

Although the non-parametric regression approach is highly recommended, in cases of high dimensional data it becomes difficult to deal with and often results in biased estimators and unreliable interpretations of the fitted model. Thus, semi-parametric models were developed which combine the properties of both the regression approaches, non-parametric and parametric (Härdle et al., 2012; Faraway, 2016; Zeger

& Diggle, 1994).
Unlike parametric regression models, semiparametric regression models fit the data without prior specifying of the functional form of the model and the fitted models are not necessarily linear. This chapter focuses more on such a regression approach using generalized additive models and generalized additive mixed models (GAMMs). Specifically, GAMMs was used to fit the data to help in investigating the factors that are significantly associated with childhood anaemia in Tanzania and Angola. However, beforehand, a discussion of additive models in general and its generalization is a necessity.

4.2 Additive Regression Model

The additive models were suggested by Friedman & Stuetzle (1981). The term additive is self-explanatory, it is simply the sum of the terms in the model. The additive model is the generalization of the ordinary linear regression models after considering problems associated with the estimation and interpretation of fully general regression surfaces (Hastie, 1990). Additive models are more flexible and more interpretable in comparison to the ordinary regression models (Hastie, 1990). For the moment, we restrict our attention to the standard multiple regression problem given in the form

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon. \tag{4.4}$$

where α is the model intercept, $X_1, X_2, ...X_p$ is the set of explanatory variables, Y is the response variable and ε is the error term, $\varepsilon \sim (0, \sigma^2)$. The main goal is to model the dependence of the response variable Y on X for certain reasons which include inference, description and prediction. The dependence of $E(Y)$ on X is assumed to be linear, unlike in additive models. Standard regression can be generalized in several ways: one class is surface smoothing, which is well discussed in the book by Hastie (1990) among others. A standard regression (non-additive) approach assumes that the response variable is linearly correlated to the covariates while in additive models that is not necessarily the case.

In additive modelling, the response is modelled as the sum of the smooth functions, for example

$$E(Y|U, V, W) = f_1(U) + f_2(W) + f_3(W)$$

For the three covariates U, V and W, the function $f(\cdot)$ are unspecified in form and are commonly estimated using linear smoothers in an iterative method algorithm known as "back-lifting".

4.2.1 Model overview

Supposing we have n observations of random variable Y and p dimensional design values denoted by $Y = (y_1, y_2, ..., y_n)^T$, $X = (x_1, x_2, ..., x_n)^T$, respectively, the additive models as suggested by Friedman & Stuetzle (1981), take the form

$$Y_i = \alpha + \sum_{j=1}^{p} f_i(x_{ij}) + \varepsilon_i. \tag{4.5}$$

Thus,

$$E(Y_i|x_{i1}, x_{i2}, \ldots, x_{ip}) = \alpha + \sum_{j=1}^{p} f_i(x_{ij}). \tag{4.6}$$

for $E(\varepsilon_i) = 0$ and $var(\varepsilon_i) = \sigma^2$. The $\varepsilon_i's$ are independent of the $X_{ij}'s$, the predictor variables, Y_i is the response variable and $f_i(\cdot)$ are unknown arbitrary univariate functions for each covariate X_{ij}. The arbitrary function $f_i(\cdot)$ is commonly estimated using linear smoothers by the use of the iterative method known as back-lifting. However, there are alternative methods proposed for estimation, such as the marginal integration estimation methods (Linton & Nielsen, 1995), the Fourier series approximation (Sardy & Tseng, 2004), The linear wavelets method of estimation (Sardy & Tseng, 2004). The back-lifting method is well discussed in the work of (Buja et al., 1989; Hastie, 1990).

The additive model is a special type of project pursuit regression (PPR) model suggested by Friedman & Stuetzle (1981). The main goal in additive modelling is to estimate the unknown functions $f_i(\cdot)$, which are said to be smooth and non-parametric to obtain the best fit model for our data. The smoothing process involves two steps: firstly, identifying the smoothing technique, and secondly, determining the smoothing parameter that controls the trade-off between underestimating and overestimating (Wood et al., 2013). The following section focuses on some of the smoothing techniques for estimating the unknown functions $f_i(\cdot)$.

4.3 Smoothing Techniques

A smoother is a tool for summarizing the behaviour of a response measurement Y, as a function of one or more predictor variables $x_1, x_2, \ldots, x_p$. Smoothers produce an estimate of the trend that is less variable than the response measurements, Y itself (Hastie, 1990). One important property of a smoother is that, they assume no distribution in advance, and are thus non-parametric. Smoothers have two main

uses: (1) For data description purposes and (2) to estimates the dependence of the mean on Y, on the set of predictor variables $x_1, x_2, \ldots, x_p$. There are a number of smoothing approaches that exist in the literature, and few of them are discussed below.

4.3.1 Linear smoothers

Supposing we have n response measurements $(Y = y_1, y_2, \ldots, y_n)^T$defined at design points $X = (x_1, x_2, \ldots, x_p)^T$. In standard regression if the design points $x_i{}'s$ are univariate real values, one can assume that the dependence of the response variable on the covariates is smooth, correspondingly, non-parametric regression approaches are often called scatter plot smoothers (Buja et al., 1989). In scatter plot smoothing the assumption is, each of y and x represents measurements of variables Y and X. By definition, scatter plot smoothing is a function of X and Y, that results in a smooth function S which is independent of Y, S has the same domain as the values in X, where $S = S(X|Y)$. Examples of linear smoothers are running mean, locally weighted running lines, kernel smoothing, smoothing splines, bin smoothers and the least-squares line (Buja et al., 1989).

Running-mean and running line smoothers

A **running mean smoother** produces a fit at each point x_i by averaging the data points in a neighbourhood N_i around x_i, thus commonly uses neighbourhoods that are symmetric nearest neighborhoods. Buja et al. (1989) defined a running mean smoother as follows

$$S(x_i) = ave_{j \epsilon N^{s(x_i)}}(y_i). \tag{4.7}$$

The symmetric nearest neighbourhoods method works by choosing k points to the left and right of point x_i that are closest in X-value to x_i. $N^{s(x_i)}$ denote the indices points. Failing to choose k points to the left and right of point x_i, results to choosing as many points as possible. The symmetric nearest neighbourhoods method can be formally defined as

$$N^{s(x_i)} = \{max(i+k, 1), \ldots, min(i+k, n)\} \tag{4.8}$$

In nature running mean smoothers are simple to execute and fast, which sometimes produces wiggly functions and are biased at the endpoints (Buja et al., 1989).

Running line smoother, fit a line by least squares to the data points in the symmetric neighbourhood N_i around each design point x_i. They are considered to be better than the average smoothers because the estimated smooth at each x is a value around

the fitted line at x_i, a process that is done at each point x_i, Thus, line smoothers reduces the bias near endpoints.

Bin smoothers

A bin smoother is similar to a running mean smoother, the difference being that the average is computed in a non-overlapping neighbourhoods. In bin smoothing, we choose cut-points $\zeta_0 < \zeta_1 < \zeta_2, \ldots, < \zeta_k$, where $\zeta_0 = -\infty$ and$\zeta_k = \infty$. $R_k = \{i, \zeta_k \leq x_i \leq \zeta_{k+1}\}$ for $k = 0, 1, \ldots, n-1$ is defined as the indices of the data points in each region. Hastie (1990) defined a bin smoother as follows:

$$S(x_i) = ave_{j \epsilon R_k} y(x_i). \tag{4.9}$$

Kernel Smoother

A kernel smoother uses a set of weights explicitly defined by kernels to produce the estimate at each large value (Hastie, 1990). Usually, the weights are given to the j^{th}point producing the estimate at x_i and are defined by

$$S_{ij} = \frac{\zeta_0}{\lambda} d\left(\frac{|x_i - x_j|}{\lambda}\right) \tag{4.10}$$

where S_{ij} are equal to weights, the parameter λ is known as the window-width (bandwidth) parameter and ζ_0 is usually chosen such that the weights, $S_{ij}{}'s$ sum to unity (Hastie & Tibshirani, 1987). Kernel smoothers decrease in a smooth fashion as one moves away from the targeted value X, and $d(t)$ is an even function decreasing in $|t|$ The value of d is chosen in such a manner that the kernel function is optimal (Härdle, 1990). A commonly used function is the parabolic function,

$$d(t) = \begin{cases} 0.75(1-t^2) & \text{for} \quad |t| \\ 0 & \text{elsewhere} \end{cases}$$

Härdle (1990) defined the kernel smoother referring to the shape of it function as a kernel K that is continuous, bounded and symmetrically real and it integrates to one.

$$\int k(u)du = 1. \tag{4.11}$$

Furthermore, the weight sequence defined as $\{W_{ni}(x)\}_{i=1}^{n}$described in the shape function of weight $\{W_{ni}(x)\}$ are the weights sequence for kernel smoothers (for one-

dimensional x). The weights sequence is defined by

$$W_{n_i}(x) = \frac{K_{h_n}(x - X_i)}{\hat{f}_{h_n}(x)} \tag{4.12}$$

where,

$$\hat{f}_{h_n}(x) = n^{-1} \sum_{i=1}^{n} K_{h_n}(x - x_i) \tag{4.13}$$

and where

$$K_{h_n}(u) = h_n^{-1} K\left(\frac{u}{h_n}\right)$$

is the kernel scale factor h_n. The function $\hat{f}_{h_n(\cdot)}$ is the Rosenblatt–Parzenkernel density estimator of the marginal density of X (Härdle, 1990).

Locally weighted running line smoother (LOWESS)

Cleveland (1979)implemented the locally weighted running line smoother (LOWESS), so-called LOESS), which is more powerful because it combines both the strict nature of local running lines and the smooth kernel weight smoother. It is easy to compute, involving three steps (Buja et al., 1989)

- Find the symmetric nearest neighbourhood ($N(x_i)$) of x_i,
- It calculates the distance to the k^{th} nearest neighbour denoted by d_i, and
- It assigns a tri-cube weight function to each point in $N(x_i)$:

Hastie (1990) defined the LOWESS function as follows

$$W_{ij} = \left(1 - |\frac{x_j - x_i}{d_i}|^3\right). \tag{4.14}$$

Because of the robustness of LOWESS, it automatically down weights outlying responses during the smoothing process, which causes it to be a non-linear smoother if it is used. On a subtler note, LOWESS uses nearest neighbours, whereas the running means and lines described earlier use symmetric nearest neighbours.

The function $S(x_i)$ in 4.3.1 is a regression function fitted at point x_i obtained by fitting the weighted least squares line. The smooth function $S(x_i)$ can also be estimated using splines, as they are briefly described in the following section.

4.4 Spline Smoothing

A spline is defined as a joint function (piecewise) from a polynomial function, for example a sequence of knots defined by $\xi_0 \leq \xi_1 \leq \ldots \leq \xi_k$ that join smoothly (Buja et al., 1989). Smoothing splines are assumed to have no pre-specified function, thus, they are a flexible approach for estimating regression curves with observed values of x_i at the knots (Wood, 2017; Hastie, 1990; Härdle, 1990).
A common measure of fidelity of the data (smoothness of data) for a curve fitted in the data is the residual sum of squares (Härdle, 1990), which can be written in the form,

$$\sum_{i=1}^{n} (y_i - g(x_i))^2$$

where, g is any curve of unrestricted function form, that interpolates the data such that the distance reduces to zero. However, using any curve g (non-unique), is not recommended because can be is too wiggly for unstructured-orientation interpolation(Härdle, 1990). For this reason, the spline smoothing approach was developed and used instead to avoid the questionable interpolation of the data. The main aim of smoothing is to produce a good fit for the data and to produce a curve with less local variation. There are several ways to avoid local variation one of which is defining the roughness based, for example, on a first derivative, second derivative and so on. The fitted curve g to the data, for analysis the integrated second derivative is the most convenient, that is roughness penalty(Härdle, 1990)

$$\int \left(g^{''}(x)\right)^2 dx$$

is used to avoid local variation.

4.4.1 Natural cubic splines

Supposing we have n collected data points $\{W_{ni}(x)\}_{i=1}^{n}$ (Härdle, 1990), the regression relationship can be modelled as:

$$Y_i = g(x_i) + \varepsilon_i. \tag{4.15}$$

where $g(\cdot)$ is a non-parametric regression function and ε_i is the random error term, $\varepsilon \sim (0, \sigma^2)$. $g(x_i)$ is defined as the regression between Y_i and X_i. On this setting (Härdle, 1990) an optimal solution to this problem will be attained by minimization

of the penalized sum of squares (PSS), given by

$$S_\lambda(g) = \sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int_{-\infty}^{\infty} \left(g^{''}(x)\right)^2. \tag{4.16}$$

where λ is a smoothing parameter that controls the trade-off between the smoother of the curve and it proximity to the response values, y_i. Larger values of λ produces smoother curve while smaller values of λ produces more wiggly curves. Equivalently, when $\lambda \to \infty$, the penalty term dominates, it force $g^{''}(x) = 0$ everywhere and the solution is standard least squares, but when $\lambda \to 0$ the solution interpolates a twice-differentiable function. The solution to $S_\lambda(\cdot)$ is a cubic polynomial whose derivations are continuous at the boundary points say $x(\star)$ and $x(\star\star)$ (Härdle, 1990).

4.4.2 Regression splines

Regression spline is the projection method for fitting the splines, (Buja et al., 1989). Regression spline uses fewer number of knots compared to others smoothers, specifically, natural cubic spline. If $\xi_0 \leq \xi_1 \leq \ldots \leq \xi_k$ denote the set of knots and $\beta_1(z), \beta_2(z), \cdot, \beta_k(z)$ set of basis functions, then the smooth function $g(x)$ in 4.16 can be estimated as

$$g(x) \approx \sum_{i=1}^{k} \beta_k(z)\alpha_k. \tag{4.17}$$

where α_k^T is a vector which can be estimated by fitting a parametric model via ordinary least squares regression, of the form

$$g(x) = \sum_{i=1}^{k} \beta_k(z)\alpha_k + \varepsilon_i. \tag{4.18}$$

4.4.3 P-Splines

P-Splines or penalized spline regression was introduced by Eilers & Marx (1996), while the concept of using fixed a spline basis combined with a penalty for modelling complexity was first introduced by Silverman (1985) and was elaborated by O'Sullivan et al. (1986). Penalized splines smoothing are widely used in practice (Ruppert et al., 2003) because of it advantage of being less computer intensive especially in large sample sizes.

In spline smoothing choosing the smoothing parameter well is very important (Härdle, 1990; Zhang et al., 1998). There are indeed several methods for selecting smoothing parameter featured in literature. To ensure effectiveness during selection of the smoothing parameter there are selection criteria's that needs to be considered (Ay-

dın et al., 2013), these include:an improved version of the AIC criterion; a robustified cross-validation method (RCV); an average predictive square error (PSE); a parallel of the AIC criterion (GFAIC); a generalized cross-validation (GCV); and cross-validation (CV), among others.
In the following subsection a discussion of few on the criterion or methods for choosing a smoothing parameter are discussed below.

The predictive squared error (PSE) criterion

During the process of choosing smoothing parameter, it is not really important to minimize the mean square error at each point, but instead the focus should be shifted to global measures such as the PSE (Aydın et al., 2013).

$$PSE(\lambda) = \left\{1 + \frac{tr(S_\lambda S_\lambda^T)}{n}\right\}\sigma^2 + \frac{\|(I - S_\lambda)f\|^2}{n}. \tag{4.19}$$

where $f = f(x_1), \cdots, f(x_n)$ is the vector of knot points $x_1, x_2, x_3, \cdots, x_n$, $tr(S_\lambda S_\lambda^T)$ is the trace of matrix $(S_\lambda S_\lambda^T)$ and $\|(I - S_\lambda)\|$ is the norm of matrix $(I - S_\lambda)$. In practice, if σ^2 is not known,an estimate is used, given by

$$\hat{\sigma}^2 = \frac{RSS(\lambda^*)}{\{n - tr(2S_{\lambda^*}S_{\lambda^*}^T)\}} = \frac{\|(I - S_\lambda)y\|^2}{\{n - tr(2S_{\lambda^*}S_{\lambda^*}^T)\}}. \tag{4.20}$$

where $RSS(\lambda^*)$ is the residual sum of squares from a smooth $S_{\lambda^*}y$ and λ^* is the pilot λ selected by using any of methods discussed by Aydın et al. (2013).

Cross validation (CV) criterion

The basic idea in the cross validation (CV) method is to exclude one data point (x_i, y_i) during the process of choosing λ, (Zeger & Diggle, 1994; Rice & Silverman, 1991). The smoothing parameter that minimizes the residual sum of squares is $\sum_{i=1}^{n}(y_i - g(x_i))^2$. Thus, estimation of the squared residual for a smooth function at x_i is based on the remaining $(n-1)$ points. The CV score is given by

$$CV(\lambda) = n^{-1}\sum_{i=1}^{n}\{y_i - \hat{g}_\lambda^{-1}(x_i)\} \equiv n^{-1}\sum_{i=1}^{n}\left\{\frac{(y_i - \hat{g}_\lambda)^2}{1 - (S_\lambda)_{ii}}\right\}. \tag{4.21}$$

where, $\hat{g}_\lambda$ is the fit (spline smoother) for the n points $\{x_i, y_i\}_{i=1}^n$ with smoothing parameter λ, $\hat{g}_\lambda^{-1}$ is the smoothing spline calculated from the $(n-1)$ remaining points and $(S_\lambda)_{ii}$ is the i^{th} diagonal element of smoother matrixS_λ.

Generalized cross validation (GCV) Criterion

Generalized cross validatio is the generalization of a cross validation selection method for choosing the smoothing parameter (λ),and is thus more advanced than CV (Craven & Wahba, 1978). The GCV score function which is constructed by analogy of ordinary cross-validation can be written in the form

$$GCV(\lambda) = n^{-1}\frac{\sum_{i=1}^{n}\{y_i - \hat{g}_\lambda(x_i)\}}{\{1 - n^{-1}tr(S_\lambda)\}^2} = \frac{n^{-1}\|1 - S_\lambda\|^2}{\{n^{-1}tr(I - S_\lambda)\}^2}. \tag{4.22}$$

There is intensive theory in the work by Hastie (1990) about the value of λ that optimizes the function by minimizing the function $S_\lambda(g)$ in 4.16. The curve g is continuous on the interval [a,b]. If $\hat{g}$ is a cubic spline with knots in each x_i we obtain a smoothing matrix.

$$\left(y - \sum_{i=1}^{p}\right)^T \left(y - \sum_{i=1}^{p}\right) + \sum_{i=1}^{n} \lambda_i g_\lambda^T K_i g_i$$

.

The $K_i's$ are the penalty matrices. For further information, readers can refer to (Aydın et al., 2013; Buja et al., 1989; Hastie & Tibshirani, 1987; Hastie, 1990).

4.5 Generalized Additive Regression

As the class of generalized linear models is described in detail by Nelder & Wedderburn (1972) and fully developed by McCullagh (1989) and, Hastie (1990) extended generalized linear models to generalized additive models (GAMs) in the same manner of development of GLM from ordinary linear regression. Unlike generalized linear models, GAMs are data driven rather than model driven (Yee & Mitchell, 1991). Additive models extend ordinary linear models by replacing the linear form $\alpha + \sum_j X_j\beta_j$ with the additive form $\alpha + \sum_j f(X_j)\beta_j$, thus generalized additive models (Chen, 2000). For example, a logistic regression model is a special case of GLM where the response variable is binary, and has the form

$$log\left(\frac{\pi}{1-\pi}\right) = \alpha + \sum_{j=1}^{p}\beta_j X_j. \tag{4.23}$$

Alternatively, a generalized additive logistic regression model is of the form

$$log\left(\frac{\pi}{1-\pi}\right) = \alpha + \sum_{j=1}^{p} f_j(X_j). \tag{4.24}$$

where $\pi = p(y = 1)$ as discussed in the previous chapter on GLMs, and Y is the indicator variable denoting the occurrence of an event of interest. The non-parametric function$f(\cdot)$ makes the generalized additive logistic regression model more flexible. Suppose that a random variable Y_i is a response variable with a distribution amongst of the exponential distributions is, the general form of GAM is as follows,

$$g(u_i) = X_i{}^*\theta + \sum_j f_j(x_j). \tag{4.25}$$

where $g(u)$ is a monotonic, invertible and a differentiable link function, $.X_i{}^*\theta$ is an estimate of X^* that can be parametrically estimated and $f_j(\cdot)'s$ are the non-parametric smooth function to be estimated.

4.5.1 Estimating the generalized additive model

Estimation of GAM is a two-step process first is, estimation of the smoothing parameters and secondly, finding the model coefficients of the maximum penalized likelihood function. Therefore, correctly choosing the basis function and the smoothing parameter in GAM is central (Zhang et al., 1998), where penalized regression smoothers are the common choice of basis that is based on smoothing splines. Thus, smooth terms can be represented as a linear combination of the basis functions, b_{jk} and the unknown regression parameter, β_{jk} such that

$$f_j(x_j) = \sum_{j=i}^{q_k} \beta_{jk} b_{jk}(x_j)$$

And substituting each smooth term in 4.25 by their basis will result in Equation 4.26

$$g(u_i) = X_i\beta. \tag{4.26}$$

where X_i contains the columns of $X_i{}^*$and the columns containing the spline basis are evaluated at each covariate, x_i, and β is the column vector that contains θ^* and all the other smooth coefficients vectors of β. Now we can relate 4.26 to a GLM class fitted by using iterative re-weighted least squares procedures; however due to the additive structure in GAM that is not really the case; instead, the penalized likelihood function is maximized using a penalized iterative re-weighted least squares method (P-IRLS). Furthermore, the optimization problem can be achieved by maximizing 4.27

$$||W^k(Z^k - X\beta)||^2 + \sum_j \lambda_j \beta^T S_j \beta. \tag{4.27}$$

where K is a constant representing the iteration index and λ_j is the smoothing parameter. W is a diagonal matrix of weights,where ${W_i}^k = W^{0.5}\left(\frac{V({u_i}^k)^{-0.5}}{(g'(u_i))^k}\right)$,$V(u_i) = \phi^{-1}var(y_i)$ and $Z^k = X\beta^k + G^k(y - u^k)$. G^k is a diagonal matrix such that the diagonal elements $G^k_{rr} = g'(u^k)$. A broad theory on iterative methods can be found on the work by (Keen & Engel, 1997)

4.6 Generalized Additive Mixed Moddels(GAMMs)

The generalized linear mixed models (GLMMs) of Breslow & Clayton (1993) provided an intensive framework for parametric regression of overdispersed and correlated outcomes. Overdispersed and correlated outcomes data frequently arise from longitudinal studies, survey sampling studies, clinical trials and disease mapping (Lin & Zhang, 1999). Although GLMMs (Breslow & Clayton, 1993) are powerful in handling such data, they have that one limiting factor,that is the assumption of the parametric mean function used to model covariate effects (Zhang & Lin, 2003). For this reason, (Lin & Zhang, 1999) proposed GAMMs as an extension of GLMMs to deal with over dispersed and correlated data with complex covariate effects. Generalized additive mixed models uses an additive non-parametric functions to model covariate effects, while accounting for overdispersion and correlation by adding random effects to the additive linear predictor (Lin & Zhang, 1999).

4.6.1 Model overview

Suppose $y = (y_1, \ldots, y_n)^T$ is the set of independent outcome (response) variables and $x_i = (x_{i1}, \ldots, x_{in})^T$ are the covariates associated with the fixed effects and Z_i is a $q \times 1$ covariate vector associated with the random effect, provided a $q \times 1$ vector of random effect b is known (or given), a general form of GAMM as proposed by Lin & Zhang (1999) is given as

$$g({u_i}^b) = \beta_0 + \sum_{j=1}^{p} f_j(x_i) + \sum_{k=1}^{q} z_k b_k. \tag{4.28}$$

where $g(\cdot)$ is a monotonic differentiable link function, $f_j(\cdot)$ is a centred twice differentiable smooth function, the vector of random effect b is assumed to be normally distributed, $b \sim (0, D(\gamma))$, where γ is a $q \times 1$ variance component vector. The outcomes variables, $y_i's$, are conditionally independent with means $E(y_i|b) = u_i^b$ and variance, $var(y_i|b) = \phi m_i v(u_i^b)$, ϕ is a dispersion parameter and $v(\cdot)$ is a prior known

variance function and m_i is a weight that is also prior known. One key fixture of GAMM, is the ability of the additive non-parametric functions to model covariate effects and the correlation between observations is modelled by the random effects (Lin & Zhang, 1999) . Therefore, 4.28 can be expressed in a matrix form (Lin & Zhang, 1999) and it simplifies to

$$g(u^b) = \mathbf{1}\beta_0 + \sum_{j=1}^{p} N_j f_j + Zb. \tag{4.29}$$

where $\mathbf{1}$ is an $n \times 1$ vector of ones, N_j is an $n \times r_j$ incidence matrix with the i^{th} component $N_j f_j$ defined by $f_j(x_j)$. $u^b = ({u_1}^b \ldots, {u_n}^b)$, $g(u^b) = \{g\left({u_1}^b \ldots, {u_n}^b\right)\}^T$ and $Z = (z_1, z_2, \ldots, z_n)^T$ A comprehensive literature of GAMM can be found in the work by Lin & Zhang (1999).

4.7 Estimating the Generalized Additive Mixed Model

It can be seen that GAMM extends GAM by means of adding the random effect term to account for correlated outcomes, thus GAMMs are estimated to be very similar to GAMs, except that GAMM further considers the inferencing of the variance component, γ. The infinite dimensional unknown parameters, $f(\cdot)$, are to be estimated using a cubic smoothing spline as they are discussed in the previous section. The smoothing parameter, λ and the variance component, γ, are jointly estimated by the marginal quasi-likelihood (Lin & Zhang, 1999)
Suppose the values of γ and λ are known, the natural cubic smoothing spline estimator of $f(\cdot)$ maximize the penalized log-quasi likelihood(Lin & Zhang, 1999)

$$l\{y, \beta_0, f_1(\cdot), \ldots, f_p(\cdot), \gamma\} - 0.5 \sum_{j=1}^{p} \lambda_j \int_{s_j}^{t_j} {f_j}''(x^2)dx = l\{y, \beta_0, f_1(\cdot), \ldots, f_p(\cdot), \gamma\} - 0.5 \sum_{j=1}^{p} \lambda_j {f_j}^T k_j f_j \tag{4.30}$$

where the j^{th} covariate is defined over a range of (s_j, t_j), the vector $\lambda = (\lambda_1, \ldots, \lambda_p)^T$ is a smoothing parameter that controls the trade-off between goodness of fit and the smoothness of the estimated function. The roughness penalty of the penalized sum of squares $\sum_{j=1}^{p} \lambda_j \int_{s_j}^{t_j} {f_j}''(x^2)$ can be estimated by $\int_j^T k_j f_j$.
To maximize the function 4.30, numerical integration methods, as discussed by Breslow & Clayton (1993), need to be applied. Lin & Zhang (1999) proposed an approximation of 4.30 using a double penalized quasi-likelihood (D-PQL) within the framework of generalized linear mixed models (GLMMs) to obtain the cubic spline smoothers.

4.7.1 Double penalized quasi-likelihood method

The double penalized quasi-likelihood method is defined as (Lin & Zhang, 1999)

$$l_{dpql} = -0.5\sum_{i=1}^{n} d_i(y_i, {u_i}^b) - 0.5\sum_{i=1}^{n} b^T D^{-1} b - 0.5\sum_{j=1}^{r} \lambda_j {f_j}^T k_j f_j. \tag{4.31}$$

where the penalty term $\sum_{i=1}^{n} b^T D^{-1} b$ results from approximation of the integrated log-quasi likelihood based on the Laplace method, $\lambda_j {f_j}^T k_j f_j$ is the penalty term that determines the smoothness of the function $f(\cdot)$ that depends on the estimates of the smoothing parameter,λ_j
Since the centred vector f_j can be re-parametrized in terms of the basis function such that $f_j = X_j\beta_j + \beta_j a_j$, the maximization of 4.31 with respect to $(\beta_j; f_1 \ldots, f_p)$ and b to obtain the cubic spline smoothers, 4.31 becomes

$$l^*_{dpql} = -0.5\sum_{i=1}^{n} d_i(y_i, {u_i}^b) - 0.5 b^T D^{-1} b - 0.5 a^T \wedge^{-1} a. \tag{4.32}$$

where x_j is an $r_j \times 1$ vector centred at r_j distinct values of x_{ij}, while $\beta_j = L_j({L_j}^T L_j)^{-1}$ with the $r_j \times (r_j - 2)$ full rank matrix L_j, which meets the condition $L_j {L_j}^T$, and ${L_j}^T X_j$ results from the identity ${f_j}^T k_j f_j$ in the parametrized DPQL. For $a = (a_1^T, \ldots, a_p^T)$. The vector $\wedge = diag(\tau_1 I, \ldots, \tau_p I)$ with $\tau = \frac{1}{\lambda_j}$, thus, the matrix equation of GAMM 4.29 can be generalized to

$$g(u^b) = X\beta + Ba + Zb. \tag{4.33}$$

where a^T and b^T are the random vector effects and are both multi-normally distributed with mean zero and variance $\wedge$ and D, respectively. $\beta = (\beta_1, \ldots, \beta_n)^T$ is a $(p+1) \times 1$ vector of the model coefficients, $X = (1, N_1 X_1, \ldots, N_p X_p)^T$ and $B = (N_1 B_1, \ldots, N_p B_p)^T$, 4.33 is simply a GLMM and ${f_j}'s$ can be obtained by fitting the model.
The maximization of 4.33 with respect to β, a and b results to the normal Equations 4.34 that can be solved by applying the iterative methods such as Fisher's scoring algorithm

$$\begin{pmatrix} X^T W X & X^T W B & X^T W Z \\ B^T W X & B^T W B + \wedge^{-1} & B^T W Z \\ Z^T W X & Z^T W B & Z^T W Z + D^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ a \\ b \end{pmatrix} = \begin{pmatrix} X^T W Y \\ B^T W Y \\ Z^T W Y \end{pmatrix} \tag{4.34}$$

to obtain estimators $\hat{f}_j$ and $\hat{a}$ and $\hat{b}$ the random effect estimators.

When comparing the covariance matrix of $\hat{f}_j$, it is more convenient to calculate the values of β and a by using

$$\begin{pmatrix} X^T R^{-1} X & X^T R^{-1} B \\ B^T R^{-1} X & B^T R^{-1} B + \wedge^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ a \end{pmatrix} = \begin{pmatrix} X^T R^{-1} Y \\ B^T R^{-1} Y \end{pmatrix} \tag{4.35}$$

where $R = W^{-1} + ZDZ^T$. Denoting the left hand-side vector of 4.35 by $H$and $H_0 = (X, B)^T R^{-1}(X, B)$, therefore, the approximate covariance matrix of $\hat{a}$ and $\hat{b}$ random estimators is

$$cov(\hat{\beta}, \hat{a}) = H^{-1} H_0 H. \tag{4.36}$$

Thus, the covariance matrix of $\hat{f}$ can be written as

$$cov(\hat{f}) = (X_j; \beta_j) cov(\hat{\beta}_j \hat{a}_j)(X_j; \beta_j)^T. \tag{4.37}$$

Here, we assume that the non-parametric functions $f_j(\cdot)'s$ are fixed in calculating the estimates $\hat{f}_j$. Furthermore, when estimating $f_j(\cdot)$ it is equally important to also estimate the smoothing parameter λ and the unknown vector of fixed regression parameters. To ensure that $\hat{f}_j$ performs well, the choice of the smoothing parameter $\hat{\lambda}$ has to be good.

In addition, on the estimation of the smoothing parameter, λ and the variance component γ there exist a number of method for approximation. One of the famous data-driven method is the cross-validation (CV) method discussed in Section 4.4.3 (Rice & Silverman, 1991; Zeger & Diggle, 1994). Zeger & Diggle (1994) pointed out that the CV is often expensive and takes too much time for variance component approximation. The second method of approximation, generalized maximum likelihood (GLM) was proposed by Wahba et al. (1985) under the non-parametric modelling of independent data (Zhang et al., 1998). Robinson et al. (1991) found that a GLM estimator of say $\tau = \frac{1}{\lambda}$ is equal to a REML estimator of τ under linear mixed models. The REML is another well-known and yet one of the most recommended (Harville, 1977) methods for estimation of λ and γ. It performs better than many methods of approximation that exist in the literature (Kohn et al., 1991) that include CV and GCV methods as discussed in section 4.4.3. For a comprehensive theory about the approximation of the smoothing parameter and the variance component, readers can refer to (Lin & Zhang, 1999; Zhang et al., 1998), among others.

4.8 Advantages and Disadvantages of GAMMs in R statistical Software

Generalized additive mixed models (GAMMs) extend generalized linear mixed models (GLMMs) by allowing the predictor variables to depend linear to the unknown smooth function of some covariates (Lin & Zhang, 1999). For this reason, GAMMs are non-parametric (or semi-parametric) because the unknown functions are modelled without any distribution specification in advance. Like GLMMs, GAMMs are not considered as the best model to fit binary response data, as they may produce unreliable results and are possibly misleading (Wood, 2017). Instead, for binary response and low mean count data GAMM must be fitted using gamm4 package in R instead of gamm, gamm4 is more preferable because of its numerical robustness. gamm4 is based on gamm from package mgcv in the statistical R-software, gamm4 uses lme4 rather than nlme as the underlying fitting engine via a trick due to Fabian Scheipl. Furthermore, gamm4 performs even better when PQL is avoided (Wood, 2017). As preferable as it might be, it cannot deal with multi-penalty smooths and it has no facility to accommodate for nlme style correlation structure, thus this remains its main disadvantage. When fitting GAM without the random effects, gam would perform much than gamm4 or gamm, whereas gamm4 would produce worse MSE than gam with REML smoothness selection (Wood et al., 2013). For large data sets, fitting GAMMs with modest numbers, gam is much faster (or better) than gamm4 when the random effects are identically and independent distributed (i.e i.i.d). Thus gamm4 is most useful when the random effects are not i.i.d or when there are large numbers of random coefficients (more than several hundred), each applying to only a small proportion of the response data. Furthermore, GAMMs are more computer-intensive compared to GLMMs (Zhang & Lin, 2003).

4.9 Application of GAMM to the TDHS and ADHS data sets

4.9.1 Introduction

As mentioned earlier, the survey logistic model under the GLM class is a parametric model and, hence, it has a number of assumptions, as discussed in 3.5; as powerful and as easy as it is to deal with complex survey data, it cannot deal with correlated data. For these two reasons a GAMM was fitted to handle the issue of possibly correlated outcomes from these two data sets with complex survey designs. A GAMM was fitted over a GLMM because of its statistical robustness and the fact that it is one of the non-parametric (semi-parametric) mixed models, which are said to be non-parametric. Thus, it has fewer assumptions compared to GLMM or SLR mod-

els, which made it a more preferable model to be fitted into the two data sets (TDHS and ADHS) to help the study's investigation of the factors that are said to be significantly associated with childhood anaemia.

4.9.2 Model fitting and interpretation of the results

The response variable is anaemia status, which is binary (anaemic or not-anaemic). The associated covariates are: age of the child, gender of the child, wealth index, highest education level mother, sex of the household head, age of the household head, stunting status, household has television, type of place of residence and the respective regions in each country. The age of the child and the age of the household head were modelled non-parametrically; hence the age of the child is no longer categorically for us to me the necessary requirements of fitting non-parametric terms. While, all the other covariates were modelled parametrically in this study. Non-parametrically terms are continous variables,the child age is no longer treated as a categorical variable. The same variables that were used in the final survey logistic regression model were used here, with an addition of two variables, with the purpose of comparing which of the two models do better when fitting the two data sets.

A gamm formula available in the mgcv package in R-statistical software was used to fit the GAMM model. The mgcv package has several options for estimations of the smooth terms, and as discussed earlier, for more details on smoothing refer to (Ruppert et al., 2003; Green & Silverman, 1993; Härdle & Kneip, 1999; Hastie, 1990). By default mgcv uses the shrinkage smoothers, which have several advantages. One of them is that: shrinkage smoothers circumvent knots placement. Furthermore, this method is constructed in such a way that it smoothes any number of covariates (Wood, 2017). However, the shrinkage smoothers are very slow and use a large amount of memory, specifically, for large data sets. For this reason, in this study the cubic splines "cr" were used to estimate the smooth terms in GAMM. In the gamm formula, the distribution was specified (binomial) and the link function (logit link) was used when fitting the model. The random effect variable (clusters) was also specified on the "random= ()" statement. Both backward and forward methods were used for variable selection, non-significant variables (p-value > 0.05) were discarded from the model and the changes on the AIC were observed. Therefore, the final GAMM model is as follows:

$$g(\mu_i) = \beta_0 + \beta_1(child \quad gender) + \beta_2(type \quad of \quad place \quad of \quad residence) + \beta_3(Region) + \beta_4(Wealth \quad index) + \beta_5(Stunting) + \beta_6(Household \quad has \quad television) + \beta_7(Motherhighest \quad education) + s(child \quad age) + s(age \quad of \quad the \quad household \quad head) + b_{0j}. \tag{4.38}$$

where $g(\mu_i)$ is the logit link function, β's are parametric regression coefficients, s_j 's are centred smooth functions and b_{0j} is the random effect distributed as $N(0, D(\gamma))$. The common widely used methods for estimating additive models are: cubic smoothing splines, locally-weighted running line smoothers, and kernel smoothers (Ruppert et al., 2003; Green & Silverman, 1993; Härdle & Kneip, 1999; Hastie, 1990) as discussed in section 4.3.

Looking at the overall significance for each effect in childhood anaemia after fitting GAMM to the TDHS and ADHS data set, we can observe from the ANOVA Table 4.1 that the variable that was found to be commonly significantly associated with childhood anaemia in both countries, when the two data sets were fitted on an SLR model, have changed. There are two more additional variables, the highest education level of the mother and the gender of the child, (while the SLR model showed only three variables, child age, regions, and the stunting covariate) that had a significant effect on anaemia. Looking at the significant factors (p-value < 0.05) by country, we can observe that in Tanzania, the model suggests that childhood anaemia was significantly associated with child gender (p-value = 0.0387), mother's highest education level (p-value < 0.0001), stunting (p-value < 0.0001), region (p-value < 0.0001) and child age, modelled non-parametrically (continuous variable), p-value < 0.0001. In Angola, GAMM reveals that childhood anaemia was significantly associated with child gender (p-value = 0.0178), household has television (p-value = 0.00014), mother's highest education level (p-value = 0.0003), stunting (p-value < 0.0001), wealth index (p-value = 0.0043), region (p-value = 0.0041), the age of the child and the age of household head, which were modelled non-parametrically (continuous variables) with their respective p-values of, < 0.0001 and < 0.0001.

Table 4.1: ANOVA results (TDHS and ADHS) for the parametric terms

	Tanzania		
Effect	**DF**	**F-value**	**p-value**
Gender	1	5.62	0.03878
Child's age(in months)	5.10	99.35	<0.0001
Type of place of residence	1	1.10	0.294
Wealth index	2	1.59	0.204
Stunting	2	18.77	<0.0001
Region	29	3.831	<0.0001
Mother's highest education level	3	8.45	<0.0001
Household has a television	2	0.36	0.699
	Angola		
Gender	12	5.62	0.0178
Child age(in months)	2.62	91.55	<0.0001
Type of place of residence	1	0.207	0.649
Wealth index	2	5.46	0.0043
Stuntinng	2	21.053	<0.0001
Region	17	2.15	0.0041
Mother highest education level	3	6.322	0.0003
Household has a television	2	8.87	0.0014

Table 4.2: Final GAMM model fitted to TDHS data set

Parameter	Estimate	Standard error	t-value	p-value	aOR(95%CI)
Intercept	0.664	0.223	2.991	0.0028 **	1.943(1.258; 3.006)
Gender (ref=male)	-0.116	0.0557	-2.068	0.0387 *	0.890(0.799;0.994)
Mother's highest education level (ref =No education)					
Primary	-0.342	0.071	-4.800	<0.0001 ***	0.710(0.618; 0.817)
Secondary	-0.443	0.129	-3.447	0.0006 ***	0.642(0.499;0.826)
Higher	-2.053	2.554	-0.803	0.422	0.128(0.001; 19.191)
Residence type (ref = urban)					
Rural	0.117	0.112	1.049	0.2944	1.124(0.903;1.399)
Wealth Index (ref = Poor)					
Middle	-0.055	0.081	-0.685	0.493	0.946(0.808;1.108)
Rich	-0.142	0.081	-1.757	0.0789	0.868(0.739;1.016)
Stunting (ref =severe)					
Moderate	-0.233	0.095	-2.454	0.0141 *	0.792(0.658;0.954)
Nourished	-0.493	0.088	-5.604	<0.0001 ***	0.611(0.513;0.725)
Household has television (ref = yes)					
No	0.005	0.159	0.030	0.9757	1.005(0.735;1.373)

Approximate significance of smooth terms

Parameter	edf	f-value	P-value
S(Child's age)	5.102	99.355	<0.0001 ***
S(Age of the household head)	1.501	2.091	0.231

Table 4.3: Final GAMM model fitted to ADHS data set

Parameter	Estimate	Standard error	t-Value	P-value	aOR(95%CI)
Intercept	1.363	0.300	4.540	<0.0001 ***	3.91(2.142;6.903)
Gender (ref=male)	-0.154	0.065	-2.371	0.0178 *	0.847(0.755; 0.974)
Mother highest education level (ref =No education)					
Primary	-0.070	0.089	-0.783	0.4337	0.932(0.783;1.113)
Secondary	-0.389	0.114	-3.420	0.0006 ***	0.678(0.540;0.844)
Higher	-0.726	0.231	-3.142	0.0017 **	0.483(0.307; 0.759)
Residence type (ref = urban)					
Rural	0.0601	0.134	0.455	0.6488	1.062(0.823; 1.388)
Wealth Index (ref = rich)					
Middle	0.449	0.134	3.218	0.0013 **	1.567(1.194;2.063)
Poor	0.504	0.173	2.910	0.004**	1.655(1.183;2.33)
Stunting (ref =severe)					
Moderate	-0.482	0.110	-4.380	<0.0001 ***	0.618(0.497; 0.765)
Nourished	-0.659	0.102	-6.478	<0.0001 ***	0.517(0.424; 0.632)
Household has a television (ref = yes)					
No	-0.342	0.112	-3.065	0.0022 **	0.710(0.574; 0.888)

Approximate significance of smooth terms

Parameter	edf	f-value	P-value
S(Child's age)	2.616	91.55	<0.0001 ***
S(Age of the household head)	1.00	20.64	<0.0001 ***

a

[a]Note: The odds ratios as shown on the above tables (4.2 and 4.3) are adjusted for the regions in each country

From the above two tables (4.2 and 4.3) we can observe that in both countries the

effect of education level is significantly associated with childhood anaemia, children whose mothers (or household head/caretaker) had no education had higher odds of being anaemic compared to other children whose mothers (or household head/caretaker) had primary education (OR=0.932, 95% CI (0.783;1.113)), secondary (OR=0.678, 95% CI (0.540;0.844)) and higher education (OR = 0.483, 95% CI (0.307;0.759)), in Angola and in Tanzania, we observe the very similar trend, mothers with primary education (OR=0.710, 95% CI (0.618;0.817)), secondary (OR=0.642, 95% CI (0.499;0.826)) and higher (OR=0.128, 95% CI (0.001;0.19.191)) education had less chances of having anaemic children compared to uneducated mothers. Male children from both countries were more likely to be anaemic compared to female children, OR= $\frac{1}{0.847} = 1.18$ (95% CI(0.755;1.113))in Angola and OR =$\frac{1}{0.890} = 1.12$ (95% CI (0.799;0.994))in Tanzania.

The odds of children from both countries (Angola and Tanzania) not suffering from stunting (normal stunting status) were reduced by 48.3% (OR = 0.517, 95% CI (0.424;0.632)) in Angola and by 38.9% (OR=0.611, 95% CI (0.513;0.725)) in Tanzania compared to children who were severely stunted. In Angola, children who were moderately stunted also had reduced odds of being anaemic compared to severely stunted children, OR = 0.618, 95% CI (0.497;0.765), while children with normal stunting status were 48.3% less likely to be anaemic compared to severely anaemic children. Furthermore, in Angola, the availability of television in the household (OR = 0.710, 95% CI (0.574;0.888)) and wealth index predictors were also found to have a significant effect on childhood anaemia. Children who were from poor and middle families were respectively, 56.7% and 65.53% more likely to be anaemic in comparison to children from rich families (Angola).

The effect of child age and household head or caretaker's age was fitted non-parametrically. Only the child age effect (p-value < 0.0001) was found to be significantly associated with childhood anaemia in Tanzania (edf=5.102), while in Angola both the child age (edf=2.616) and age of the caretaker (edf=1.00) had a significant effect, with p-values for both < 0.0001.

The plots for smooth effects are presented in the figures below, HW1- is the child age and V012-is the caretakers age:

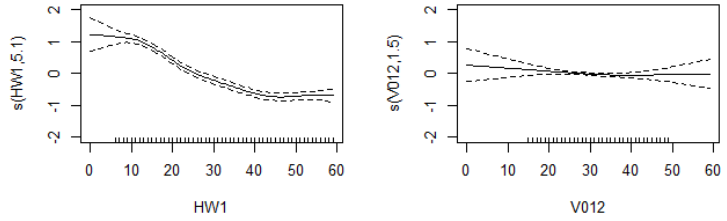


Figure 4.1 – Smoothing components for anaemia with Child age and household head age (TDHS data set)

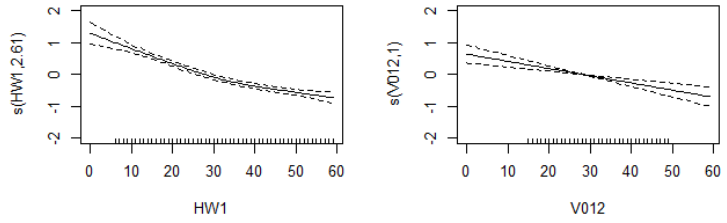


Figure 4.2 – Smoothing components for anaemia with Child age and household head age (ADHS data set)

The two figures presented above are smoothing components of childhood anaemia with child age and household head. In each panel, the smooth line is the estimated trend from a generalized additive mixed model. The y-axis represents the effect of the age (child and household age) term, where s is a smoother term and the number

in parentheses is the estimated degrees of freedom (edf). We can observe in Figure 4.1, and 4.2 specifically, for HW1 (child age) there is a decreasing trend as the x-axis (age of the child in months) increases; children at a young age had higher chances of being anaemic. The test statistic was respectively, 99.35 and 91.55 in TDHS and ADHS with p-values both less than 0.0001.

4.10 Summary and Discussion

In this section, a generalized additive mixed model (semi-parametric mixed model) was fitted to the two data sets for the purpose of investigating the demographic and socioeconomic factors that are significantly associated with childhood anaemia. The GAMM is a flexible statistical approach because of its ability to fit both the non-parametric and parametric terms simultaneously. There are a number of statistical models that could have been used to analyse these data sets (for example, generalized estimating equation, generalized linear mixed models, etc.), but because of the robustness of GAMM in dealing with data that assumes complex survey designs (data that involves stratification, clustering and with unequal weights), it was recommended for this thesis.

Looking at the two models fitted on the data sets (SLR and GAMM), we can identify a huge difference in the way they fit the data. One of the most notable differences is that SLR is a fully parametric model that fitted the data sets assuming that there was a linear relationship between the predictor variables and the response variable, while GAMM relaxes that assumption. Secondly, other variables were fitted non-parametrically. The two models agree on the factors that are said to have a significant effect on childhood anaemia in both countries; however, GAMM further revealed that the mother's highest education level and the child gender are in fact also significantly associated with childhood anaemia (p-values < 0.05 in both countries).

The model highlights that, in both countries, severely stunted children and children with highly educated mothers (higher education, e.g. attended university) were less likely to be anaemic compared to non-stunted and children with mothers who had no education, respectively. Although the model shows non-significant effect of the type of place of residence and the standard of living in the household in Tanzania, considering other effect sizes besides the p-value, we can tell that children from poor homes were more likely to suffer from anaemia compared to children from rich families; children from rural areas were also more likely to be anaemic compared to children from urban areas. Furthermore, the effect of child age (fitted non-

parametrically) is the common cause of childhood anaemia in both countries. Moreover, the age of the child was fitted non-parametrically, with an effective degrees of freedom in both results not equal to one (edf $\neq 1$), this simply means the relationship between childhood aneamia and the child age is not linear. The $edf = 1.00$ for caretaker age in Angola, suggests that there was a linear relationship between childhood anaemia and the age of the caretaker in Angola, while in Tanzania that is not the case. In the next chapter, our focus is shifted, the data sets were analysed using spatial regression.

Chapter 5

Spatial Regression Analysis

5.1 Introduction

Thus far we have fitted generalized linear models, specifically, survey logistic regression (SLR) and generalized additive mixed models (GAMMs). However, none of these modelling techniques take into account the effect of spatial variability. Hence, this chapter was added to investigate if there is an effect of spatiality and further to produce maps illustrating the predicted (interpolated) prevalence of childhood anaemia in different regions in each country.

Similar to the general linear regression, spatial regression is a broad class of statistical modelling that focuses on space. Generally, statistical analysis describes the cause and effect relationship between a dependent variable and one or more independent variables. One kind of data set used for analysis is a cross-sectional data set which is sometimes referred to as spatial or area data (Zurnila et al., 2019). Unlike other data sets, cross-sectional data sets state the observation and its location, that is, where the subject resides. Hence, there are high chances that subjects are spatially dependent, meaning that observations in a region could be related to observations from other areas. All things are related , but something adjacent is more influential (Tobler, 1979). Hence, in this thesis spatial regression analysis was included because DHS data sets are a type of cross-sectional (spatial) data.

Spatial data can be divided into three methods (Schabenberger & Gotway, 2005). These are point pattern analysis, point referenced data and areal data. Point pattern analysis is defined as the methods for lattice data and geostatistics, point referenced data is often called geocoded or geostatistical data set and aerial data is also called lattice data (Schabenberger & Gotway, 2005). On rare occasions, you find that a data set that features both point and areal-level data. Most of the time in point pattern

data, the response occurrence of an event is fixed and only the areas (location) where it occurs are thought of as random. In epidemiological analysis, the geostatistical approach is said to be the most relevant and is conducted at the landscape scale and based on remote sensing (Goovaerts et al., 2005, 1997; Allard, 2013). In this study, since we are working with cross-sectional data sets, the point referenced method is the most appropriate to apply.

5.2 Model Structure

A common way to deal with spatially correlated data with random effect and continue with maximum likelihood estimation approaches for covariate estimation and covariogram parameters is based on the theory of generalized linear mixed models of Breslow & Clayton (1993). Generalized linear mixed models have the form (Schabenberger & Gotway, 2005)

$$g(u) = X\beta + Zb + \varepsilon$$

where X and Z are the design matrices of fixed (with $N \times p$ dimensions) and random effect (with $N \times q$), respectively. β is a $p \times 1$ vector coefficients of fixed effects and b is a $q \times 1$ vector of random effects. In spatial problems, b, is assumed to be normally distributed with a $mean = 0$ and a $variance = \sum_b(\theta)$ ($b \sim N(0, \sum_b(\theta))$) and the random error terms are also Gaussian distributed with mean of zero and variance of $\sigma^2 \times I$, $\varepsilon \sim N(0, \sigma^2{}_\varepsilon I)$, where I is an identity matrix. The spatial correlation is parametrized by θ in $\sum_b(\theta)$. To incorporate the location s_i, we assume that $y(s_i|b)$ is conditionally independent for any location with a mean $E(y(s_i|b)) = \mu(s_i)$. The parameter b is used to define the distribution of s (Schabenberger & Gotway, 2005). Similar to the classical generalized linear models, $g(u)$ is a classical link function which is normally the a function of the location parameter.

Basically, non-normal (non-Gaussian problem) spatial problems can be analysed in the context of generalized linear mixed models. Supposing $Y(s_i)$ is a spatial random variable at location s_i for $i = 1, 2, ..., n$ and assuming that, $Y = (Y(s_1), \ldots, Y(s_n))^T$ are conditionally independent given the latent variable $\delta(s_i)$ with probability density function (Hosseini Shojaei et al., 2018)

$$f(y(s_i), \eta(s_i), \psi(s_i)) = \exp\left(\frac{y(s_i)\eta(s_i) - b(\eta(s_i))}{a(\psi(s_i))}\right) + c\left(y(s_i), \psi(y(s_i))\right). \qquad (5.1)$$

for some specific function $a(\cdot)$,$b(\cdot)$ and $c(\cdot)$. where $\eta(s_i)$ and $\psi(s_i)$ are canonical and

scale parameters, respectively. for $X(s_i) = (x(s_1), x(s_2), \ldots, x(s_n))^T$, with $x(s_i) = 1$ a spatial generalized linear mixed model (SGLM model) can be defined as follows (Hosseini Shojaei et al., 2018)

$$g(E(Y(s_i))) = \beta^T X(s_i) + \varphi(s_i). \tag{5.2}$$

where $g(\cdot)$ is a link function and $\beta \in \mathbb{R}^p$ are the unknown regression parameters. The latent variables $\delta(s_i) = (\delta(s_1), \ldots, \delta(s_n))^T$ are assumed to be multi normally distributed in SGLM with mean zero and a covariance matrix $\sum_\theta = \sigma^2 \mathbb{R}(\psi)$, where $\mathbb{R}$ is the correlation matrix with elements $\mathbb{R}_{ij} = \rho(s_i - s_j, \psi)$ and $\psi(\cdot, \psi)$ is a valid spatial correlation function on $\mathbb{R}^2$ is indexed by a parameter ψ. The parameters $\theta = (\sigma^2, \psi)$ are sometimes called the partial sill and range, respectively. Furthermore, specification of the likelihood of the random effect $y(s)$ is required.

In the literature, there exist several functions that could be used to study spatial dependence. For the study of spatial correlation, there are three major functions used in geostatistics. These functions are correlogram, the covariance and the semivariogram (also called variogram). In geostatistics, the variogram is the main function that is used to fit a model for spatial correlation in the data. By definition (Sherman, 2011), a variogram is given by

$$2\gamma(h) = \frac{1}{N(h)} \sum_{i=1}^{N(h)} \{Z(s_i) - Z(s_i + h)\}^2. \tag{5.3}$$

where $2\gamma(h)$ is a variogram which represents the average variance between observations separated by distance h, for $h = s_i - s_j$, $\gamma(h)$ is the semivariogram, $Z(s_i)$ is the measurement at location s_i and $N(h)$ is the number of sampled data points of distance (lag) of length h. The shape of a semivariogram has the form which is presented in Figure 5.1 (Ayele, 2013).

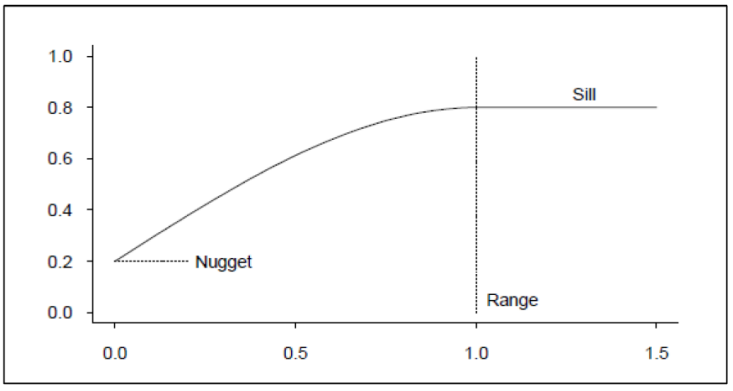


Figure 5.1 – Idealized form of variogram function, illustrating the nugget, sill and range

The model found in the variogram is used in kriging estimation. Moreover, variogram models are used to understand the maximum distances of spatial autocorrelation which can be further used in the construction of search parameters for different interpretation techniques (Ayele, 2013). A variogram represents both the structural and the random aspects of the data. Moreover, a variogram has certain criterias to meet, for instance, to develop a variogram $\mu(s)$ is assumed to constant (Sherman, 2011) and define

$$Var(Z(s_1) - Z(s_2)) = 2\gamma(s_1 - s_2)$$

the variance of s_1 and s_2 is through their difference $s_1 - s_2$ and the process which satisfies the property is called intrinsically stationary (IS). If $\gamma(h)$ depends only on its vector argument h through its height $||h||$ then the process is called isotropic. However, the process that is said to be both intrinsically stationary and isotropic is known as homogeneous process (Sherman, 2011; Schabenberger & Gotway, 2005). For some function, $\gamma_0(h) = \gamma(h)$, it is more convenient to deal with isotropic processes because there are a number of widely used parametric funtions for $\gamma_0(h)$. A semivariogram increases monotonically to reach a peak (called sill) and range (r) with spatial variance called partial sill σ^2, and a non-random variance starting at $h > 0$ referred to as nugget, as shown in Figure 5.1. Several examples are shown as follows:

1. Linear

$$\gamma_0(h) = \begin{cases} 0, & \text{if} \quad |h| = 0 \\ C_0 + C_1 h, & \text{if} \quad |h| > 1 \end{cases} \tag{5.4}$$

Where C_0 and C_1 are positive constants. As $t \to \infty$, $\gamma_0(h) \to \infty$. Thus, the assumption of stationary is not satisfied. h is the lag distance interval.

2. Spherical

$$\gamma_0(h) = \begin{cases} 0, & \text{if} \quad |h| = 0 \\ C_0 + C_1 h \left(\frac{3}{2} \frac{h}{R} - 0.5 \left(\frac{h}{R} \right)^3 \right), & \text{if} \quad 0 < |h| \leq R \\ C_0 + C_1 h & \text{if} \quad |h| \geq R \end{cases} \tag{5.5}$$

This is valid in $(\mathbb{R}^d, for \quad d = 1, 2, 3)$. The spherical function reaches the sill at $|h| = R$. The model is almost linear at small lags. In practice, spherical models are commonly used variogram structures (Schabenberger & Gotway, 2005), particularly for modelling spatial correlation that decreases with an increase in $|h|$ which is

simply the spatial distance.

3. Exponential

$$\gamma_0(h) = \begin{cases} 0, & \text{if} \quad |h| = 0 \\ C_0 + C_1(1 - \exp(-\frac{h}{R})) & \text{if} \quad |h| \geq 1 \end{cases} \tag{5.6}$$

This is simpler in functional form compared to the spherical case. Exponentials cases are valid for all d. However, it reaches the sill asymptotically as $|h| \to \infty$.

4. Gaussian

$$\gamma_0(h) = \begin{cases} 0, & \text{if} \quad |h| = 0 \\ C_0 + C_1(1 - \exp(-\frac{h^2}{R^2})) & \text{if} \quad |h| > 1 \end{cases} \tag{5.7}$$

Similar to the exponential case, Gaussian cases are also valid for all dimension. The Gaussian model reaches the sill asymptotically. Moreover, they are applicable if the data is continuous at a short lag distance. Equivalently, they are applicable when spatial correlation nearby points are very high.

5. Exponential Power form

$$\gamma_0(h) = \begin{cases} 0, & \text{if} \quad |h| = 0 \\ C_0 + C_1 h(1 - \exp(-|\frac{h}{R}|^p)) & \text{if} \quad |h| \geq 1 \end{cases} \tag{5.8}$$

where $< p \geq$ lies in the interval. Note: Gaussian and exponential forms are special cases of the exponential power form.

6. Relation quadratic

$$\gamma_0(h) = \begin{cases} 0, & \text{if} \quad |h| = 0 \\ C_0 + C_1 h(1 - \exp(\frac{h^2}{R})) & \text{if} \quad |h| > 1 \end{cases} \tag{5.9}$$

7. Wave forms

$$\gamma_0(h) = \begin{cases} 0, & \text{if} \quad |h| = 0 \\ C_0 + C_1 h(1 - \frac{R}{h} sin(\frac{h}{R})) & \text{if} \quad |h| \geq 1 \end{cases} \tag{5.10}$$

Generally, the wave or hole forms are used when there is presence of periodicity in the data resulting in a hole effect.

8. Power law

$$\gamma_0(h) = \begin{cases} 0, & \text{if} \quad |h| = 0 \\ C_0 + C_1 h^{\lambda} & \text{if} \quad |h| \geq 1 \end{cases} \tag{5.11}$$

The power laws forms are valid for all dimensions, but the power does not reach the sill. Non-positive definiteness requires $0 \leq \lambda < 2$, which is an example of a semivariogram that does not correspond to the stationary process Ayele (2013).

9. The Matern class
This method was proposed by Matérn (1960), which was neglected in favour of the simple analytic forms. Handcock & Stein (1993); Handcock & Wallis (1994), demonstrated the flexibility of this method in handling several spatial data sets. The class is best defined in terms of isotropic covariance. Therefore,

$$C_0(t) = \frac{1}{2^{\theta_2 - 1}\gamma(\theta_2)} \left(\frac{2\sqrt{\theta_2 t}}{\theta_1} \right)^{\theta_2} K_{\theta_2} \left(\frac{2\sqrt{\theta_2 t}}{\theta_1} \right). \tag{5.12}$$

where θ_1 is the spatial scale parameter and θ_2 is the shape parameter, both θ_1 and θ_2 are greater than zero.$\gamma(\cdot)$ is a gamma function and K_{theta_2} is the modified Bessel function. For further theory, readers can refer to Schabenberger & Gotway (2005) and Sherman (2011).

The other problem to deal with in the spatial process is that of anisotropic processes, there are several ways for direct generalization. The simplest of them all is the geometric anisotropy. A semivariogram with the form of geometric anisotropy is given by

$$\gamma(h) = \gamma_0(||Ah||)$$

where γ_0 is an isotropic semivariogram and A is a $d \times d$ matrix representing a linear transformation of $\mathbb{R}^d$. If matrix A is the identity matrix this reduces to an isotropic case; the process is isotropic in some linearly transformed space. Moreover, if A is positive definite, the contours of equal variance are ellipses and not circles. To generalize anisotropy (Ayele, 2013), let the simple independent intrinsically stationary process be $Z_1, \ldots, Z_p$. Therefore we can define

$$Z = Z_1 + \cdots + Z_p$$

which is also intrinsically stationary. Thus, the corresponding semivariogram is given by

$$\gamma(h) = \gamma_1(h) + \cdots + \gamma_p(h)$$

where $\gamma_1(h), \ldots, \gamma_p(h)$ are the semivariograms of $Z_1, \ldots, Z_p$, respectively. Hence

$$\gamma(h) = \sum_{i=1}^{p} \gamma_0(A_i h). \tag{5.13}$$

γ_0 is an isotropic semivariogram and $A_1, \ldots, A_p$ are matrices.

5.2.1 Valid Covariance and Semivariogram Functions

Consider isotropic models for the covariance function and semivariogram of spatial process. Let $C(h)$ be isotropic covariance of second order stationary (SoS) field and $\gamma(h)$ be the isotropic semivariogram of second order stationary or intrinsically stationary (IS) field (Sherman, 2011). Therefore, a valid covariance $C(h)$ is a positive definite function given as:

$$\sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j C(S_i - S_j) \geq 0. \tag{5.14}$$

For any finite configuration of spatial location $(s_1, s_2, \ldots, s_m)$ and all for all real numbers $(a_1, a_2, \ldots, a_m)$. According to the Bochner's equation 5.14 can be represented in spectral form

$$\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} exp(iw'h) dS(w). \tag{5.15}$$

for $S(w) = S(w)dw$ integrated over $\mathbb{R}^d$ and S is a positive bounded spectral measure (Schabenberger & Gotway, 2005).

$$C(h) = \int_0^d \phi_d(h_w) dH(w)$$

and

$$\phi_d = \left(\frac{2}{t}\right)^v r(\frac{d}{2}) J_v(t)$$

Where ϕ_d is commonly known as the basis function of the covariance model in $\mathbb{R}^d$ and $V = \frac{d}{2} - 1$, J_v is known as a Bessel function of the first kind of order V and H is a non-decreasing function on $[0\ , \infty\ )$ interval, with $\int_0^\infty dH(w) < \infty$. Furthermore, there is a corresponding theory for variogram that we can refer on by Schabenberger & Gotway (2005) to check the model validity. For SoS process of semivariogram $\gamma(\cdot)$,

if $(a_1, a_2, \ldots, a_m)$ are constants with $\sum a_i = 0$, therefore

$$\sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j C(S_i - S_j) \leq 0. \tag{5.16}$$

A semivariogram as in the case of covariance has also a spectral representation of the following form (Schabenberger & Gotway, 2005):

$$\gamma(h) = \frac{1}{2}\int_0^\infty w^{-1}\left(1 - \phi_d(w_h)\right) dH(w). \tag{5.17}$$

with $\int_0^\infty (1+w^2)^{-1} dH(w) < \infty$. For $\gamma(h)$ to be a valid semivariogram, $2\gamma(h)$ should grow slowly than $||h||^2$ which is often referred to as the intrinsic hypothesis (Schabenberger & Gotway, 2005).

5.3 Estimation

In the previous section we looked at and defined the concept of spatial covariance and variogram, now the main goal is to find a valid variogram that as a measure of spatiality (Schabenberger & Gotway, 2005), it is the closest to the spatial dependence present in the data $Z(s_1), \ldots, Z(s_n)$.
To estimate a variogram, there are several methods that exist in the literature. These methods are: Matheron's (method of moments) estimator, the Cressie-Hawkins robust estimator, and estimators based on order statistics and quantiles. For variogram estimation, the method of moments is known as the simplest estimator.

Let $Z(s_1), \ldots, Z(s_n)$ be a set of spatial data, where one can plot the squared difference $(Z(s_i) - Z(s_j))^2$ against the lag distance h and the resulting graph is know as the semivariogram cloud (Schabenberger & Gotway, 2005). However, the squared difference $(Z(s_i) - Z(s_j))^2$ unbiased estimation of the variogram at lag h, $h = s_i - s_j$ provided the mean of the random field is assumed to be constant. A more useful estimator is obtained by summarizing the squared differences. The moment estimator or sample variogram is given by (Sherman, 2011)

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|}\sum_{N(h)} (Z(s_i) - Z(s_j))^2. \tag{5.18}$$

where $N(h) = (s_i, s_j)$: $||s_i - s_j|| = ||h||$. The set of all location separated by vector h and $|N(h)|$ is the number of all districts in $N(h)$. This assumes that for lag vectors h of interest, there are sufficient pairs of points separated by vector h. The estimator $\hat{\gamma}(h)$ in $\hat{\gamma}(h)$ is well known as the classical Matheron estimator (Habyarimana, 2016)

For sparse data set, it is highly advisable to group the distances into bins according to distance lag and lag tolerance (Cressie & Hawkins, 1980). Hence, the corresponding average squared is $\frac{(Z(s_i)-Z(s_j))^2}{2}$ in each bin which is referred to as a semivariogram estimate for the distance lag. The lag tolerance must be chosen such that adequate spatial resolution and stability are retained. Furthermore, the lag distance tolerance must be chosen such that at least 30 location-to- location pairs wall with each bin (Journel & Huijbregts, 1978).

Cressie & Hawkins (1980) proposed an estimator that alleviates the non-negative impact of the underlying observations by eliminating squared differences from the evaluation. This estimator is known as the robust semivariogram estimator the Cressie Hawking (CH) estimator.

Robust estimation of semivariogram

The moment estimator is the average of the squared differences and thus can be greatly influenced by small number aberrant values (Sherman, 2011). It is advisable to consider the robust estimator, to lessen the importance of any large squared differences. Cressie & Hawkins (1980) proposed the following estimator

$$\bar{\gamma}(h) = \frac{1}{|2(h)|} \frac{\sum_{N(h)} \{(Z(s_i) - Z(s_j))^{\frac{1}{2}}\}^4}{0.457 + \frac{0.494}{N(h)}}. \tag{5.19}$$

This estimator (5.19) is robust in the sense that it is resistant to contaminated normal distributions and outliers resulting from skewed distributions (Sherman, 2011).
Another case that should be taken into consideration when estimating a variogram (semivariogram) is when the data are unequally spaced. Often, there are no pairs of points separated by any particular spatial lag h. Thus to obtain such an estimate in these cases some amount of smoothing is necessary (Sherman, 2011).

The Kernel smoothing in estimation of semivariogram

Let $w(u)$ be a non-negative symmetric bivariate density function centred at 0, with $\int w(u) = 1$. The kernel estimator of $\gamma(h)$ is given by (Sherman, 2011)

$$\hat{\gamma}_\delta(h) = \frac{\sum_{i,j} \{Z(s_i) - Z(s_j)\}^2 w_\delta(h - h_{ij})}{\sum_{i,j} w_\delta(h - h_{ij})}. \tag{5.20}$$

where $h_{ij} = s_i - s_j$ is the observed spatial lag between i and j, δ is the bandwidth parameter that determines the amount of averaging that goes into the estimate at each lag h and $w_\delta(u) = \frac{1}{\delta} w(\frac{u}{\delta})$.

Both the Cressie Hwakins (CH) and the Matheron estimators (5.18 and 5.19) have unbounded influence functions and a breakdown point of 0%. The influential function of an estimator measures the effect off infinitesimal contamination of the data on the statistical properties of the estimator (Hampel et al., 2011) and the point is the percentage of the data that can be replaced by arbitrary values without explosion of the estimator. Rousseeuw & Croux (1993) proposed the median absolute deviation (MAD), a robust estimator scale with a 50% breakdown point and smooth influence function. For a set of numbers $\{x_1, \ldots, x_n\}$, the MAD is given by

$$MAD = d \quad median_i(||x_i - median_j(x_j)||). \tag{5.21}$$

where the $median_j(x_j)$ is the median of the x_j and d is chosen such that its produces approximately and consistently. Furthermore, Genton (1998, 2001) proposed a modified version of 5.20 and 5.21 to develop a robust estimator of the variogram based on Q_n.Their Q_n estimator is given by the k^{th} order statistic of the $\frac{n(n+1)}{2}$ inter-points distance. Assigning $h = \frac{n}{2} + 1$ and $k = \begin{pmatrix} h \\ 2 \end{pmatrix}$, therefore Q_n would be defined as $Q_n = c\{||x_i - x_j|| : i < j\}_k$. Let $N(h)$ denote pairwise difference, $T_i = Z(s_i) - Z(s_i + h)$ for $i = 1, 2, 3, \ldots, \frac{n(n-1)}{2})$ for observed spatial data $\{Z(s_1), \ldots, Z(s_n)\}$. After, calculate $Q||N(H)||$ for T_i and return the semivariogram estimator at lag h

$$\bar{\gamma}(h) = \frac{1}{2} Q^2_{|N(h)|}. \tag{5.22}$$

Thus $\gamma(h)$ also has 50% breakdown points since Q_n has 50% breakdown points, but not necessary in terms of $Z(s_i)$.

The empirical semivariogram estimator could alternatively be robustly estimated using quantiles of the distribution squared differences $(Z(s_i) - Z(s_j))^2$, equivalently, $|Z(s_i) - Z(s_j)|$, instead of considering the arithmetic averages (Schabenberger & Gotway, 2005). Let $(Z(s_i), Z(s_i + h))'$ denote a bivariate Gaussian with common mean, therefore

$$\frac{1}{2}\{Z(s_i), Z(s_i + h)\}^2 \sim \gamma(h)\chi^2_1$$

$$\frac{1}{2}|Z(s_i), Z(s_i + h)|^2 \sim \sqrt{\frac{\gamma(h)}{2}}|U|, U \sim G(0, 1).$$

Let $q^p_{|N(h)|}$ denote the p^{th} quantile, therefore

$$\hat{\gamma}(h) = q^{p}_{|N(h)|}\{\frac{1}{2}(Z(s_i) - Z(s_i + h))^2\}. \tag{5.23}$$

estimates $\gamma(h) \sim \chi^2_{p,1}$. If $p = 50\%$, then 5.23 reduces to median estimator as:

$$\begin{aligned} \hat{\gamma}(h) &= \frac{1}{2} median_{|N(h)|} \frac{\{\frac{1}{2}(Z(s_i) - Z(s_i + h))^2\}}{0.455}, \\ &= \frac{1}{2} \frac{\left(median_{|N(h)|} \frac{1}{2}(Z(s_i) - Z(s_i + h))^{\frac{1}{2}}\right)^4}{0.455}. \end{aligned} \tag{5.24}$$

then $q^{p}_{|N(h)|}$ reduces the median-based estimator. These methods provide estimates at a finite set of lags or lag classes. In order to obtain estimates of $\gamma(h)$ at any arbitrary lag, the empirical semivariogram must be smoothed and, hence, the kernel smoothing estimator could be used as discussed (Schabenberger & Gotway, 2005).

The properties of the semivariogram estimates $\hat{\gamma}(h)$, $\bar{\gamma}(h)$ have been extensively studied, but on far single value h, as a function of over all h. The chances are, these estimators may not be appropriate to meet the condition of non-positive definiteness,and, as such, they lack a very important condition (Sherman, 2011). Thus, spatial predictions derived from such estimators might have negative variances. Furthermore, such problems could be avoided by replacing the empirical $\gamma(h)$ with some parametric form which is known to be conditionally non-positive definite. Hence, there is a necessity to seek a parametric family that adequately models the observed data (Sherman, 2011; Schabenberger & Gotway, 2005; Ayele, 2013). Generally, there are three methods used to estimate the empirical semivariogram $\gamma(h)$ parametrically: Least-squares estimation, maximum likelihood or restricted maximum likelihood and the Bayesian estimation. For this thesis, only the least-squares method, maximum likelihood and/or the restricted maximum likelihood method are discussed.

Least squares estimation

Suppose that a semivariogram $\gamma(h)$ has been estimated at a finite value of h and we desire to fit a model with a priori-defined form with function $\gamma(h, \theta)$ in terms of finite parameter θ. Assume that the method of moments and let is applied and let $\hat{\gamma}$ denotes the vector of estimates, $\gamma(\theta)$ the vector of model values at the same vector of h values. In the literature there are three versions of non-linear least squares estimators: ordinary least squares (OLS), generalized least squares (GLS) and the generalized weighted least squares (WLS) (Cressie, 1985). In OLS, the parameter θ can be minimized using $\{\hat{\gamma} - \gamma(\theta)\}'\{\hat{\gamma} - \gamma(\theta)\}$. In GLS, θ can be minimized as

$\{\hat{\gamma} - \gamma(\theta)\}' V(\theta)^{-1} \{\hat{\gamma} - \gamma(\theta)\}$, where $V(\theta)$ is the variance matrix of $\hat{\gamma}$. The GLS estimates depend on the unknown parameter θ, because the problem is not linear. Lastly, the parameter $theta$ in WLS can be minimized by $\{\hat{\gamma} - \gamma(\theta)\}' W(\theta) \{\hat{\gamma} - \gamma(\theta)\}$, where $W(\theta)$ is a diagonal matrix whose entries are the variance of the matrix $\hat{\gamma}$. Both OLS and GLS allow for use of variance and covariance of $\hat{\gamma}$ while GLS allows only the variance of $\hat{\gamma}$(Cressie, 1985).

Furthermore, these three estimators (OLS, GLS, and WLS) are expected to be in increasing order of efficiency, but in decreasing order of convenience to use (Habyarimana, 2016). However, it turns out that OLS is the most convenient estimator for non-linear least-squares procedure. While, the other two methods, GLS and WLS require specification of the matrices $V(\theta)$ and $W(\theta)$. Assuming a Gaussian process, the following expression is given (Ayele, 2013)

$$var\{(Z(s+h) - Z(s))^2\} = \{2\gamma(h)\}^2. \tag{5.25}$$

$$\begin{aligned} & cov\Big(\{Z(s_1+h_1) - Z(s_1)\}^2, \{Z(s_2+h_2) - Z(s_2)\}^2\Big), \\ = & \frac{(\gamma(s_1 - s_2 + h_1) + \gamma(s_1 - s_2 - h_2) - \gamma(s_1 - s_2 + h_1 + h_2) - \gamma(s_1 - s_2))^2}{4\gamma(h_1)\gamma(h_2)}. \end{aligned} \tag{5.26}$$

This equation can be useful in evaluation of the matrices $V(\theta)$ and $W(\theta)$. As one of the least-squares estimator, it is not guaranteed that the resulting minimization problem has a unique solution (Schabenberger & Gotway, 2005).

Cressie (1985) proposed the WLS criterion to solve this complication. Suppose $\hat{\gamma}$ is evaluated at a finite set of values of $\{h_j\}$ and choose θ to minimize

$$\sum_j |N(h_j)| \{\frac{\hat{\gamma}(h)}{\gamma(h_j, \theta)}\}^2. \tag{5.27}$$

then WLS can be derived under the approximation of Equation 5.27 and can be given as

$$var(|N(h_j)|) \approx \frac{8\hat{\gamma}(h)}{|N(h)|}. \tag{5.28}$$

where Equation 5.28 follows from Equation 5.27, assuming that $Z(s_i) - Z(s_j)$ are individual independent terms. Although this assumption is not satisfied this remains a reasonable estimate. Moreover, if the pairs (s_i, s_j) lying in $N(h)$ are widely spread

from one another over a sample space, then the assumption of independence would be a reasonable approximation (Schabenberger & Gotway, 2005). Therefore Equation 5.27 is not difficult to implement to OLS.

Maximum likelihood estimation (MLE)

The idea of maximum likelihood estimation of spatial data was first introduced by Kitanidis (1983). This method is only applicable for Gaussian distributed data under the assumption of $Z(s) \sim N(X(s)\beta, \sum(\theta))$, where $\sum = \alpha V(\theta)$. α is a scale parameter, θ is an unknown parameter and $V(\theta)$ is a vector of standard covariance. The maximum likelihood (ML) method estimates the mean and the covariance parameters altogether at one time. The ML estimates are a solution to the simultaneous solution to the problem of minimizing negative twice the Gaussian log likelihood given by (Schabenberger & Gotway, 2005) as:

$$\varphi(\beta, \theta, Z(s)) = ln(|\sum(\theta)|) + nln(2\pi) + (Z(s) - X(s)\beta)^{'} \sum(\theta)^{-1} (Z(s) - X(s)\beta) . \tag{5.29}$$

The estimate of β can be obtained by differentiating 5.29 with respect to β and solve. This results in generalized least squares (GLS) estimator

$$\hat{\beta} = \left(X'(s)\sum(\theta)^{-1}X(s)\right)^{-1} X(s)'\sum(\theta)Z(s). \tag{5.30}$$

Equation 5.29 and 5.30 yields an objective function for minimization of $\hat{\beta}$ given by

$$\varphi_{\beta}(\theta, Z(s)) = ln(|\sigma^2 \sum(\theta^*)|) + nln(2\pi) + \sigma^{-2} r \sum(\theta^*)^{-1} r. \tag{5.31}$$

for

$$r = Z(s) - \left(X(s)'\sum(\theta)^{-1}X(s)^{-1}\right)^{-1} X(s)'\sum(\theta)^{-1}Z(s)$$

where r is the GLS residual.σ^2 can be estimated from 5.31 , note that its MLE is

$$\hat{\sigma}^2 = \frac{1}{2} r' \sum (\theta^*)^{-1} r.$$

Thus, substituting again yields the negative of twice the profiled log likelihood as (Schabenberger & Gotway, 2005)

$$\varphi_{\beta,\sigma}(\theta^*, Z(s)) = ln\left(|\sum(\theta^*)|\right) + n\left(ln(2\pi) - 1\right). \tag{5.32}$$

Therefore, minimizing Equation 5.32 is an optimization with only $(q-1)$ parameters. One of the disadvantages of likelihood estimation is the ability to estimate

the variance-covariance matrix of the parameter estimates based on the observed or expected information matrix.

Restricted maximum likelihood estimation (REML)

This method of estimation was originally introduced by Patterson & Thompson (1971), specifically for estimating variance-covariance parameters from data that follow the Gaussian linear model. The REML estimates are frequently preferred over maximum likelihood estimates. Since the latter, they have exhibited a greater negative bias for estimates of covariance parameters. In the case of the spatial process:

$$Z(s) \sim N(X(s)\beta, \sum(\theta)),$$

The adjusted REML consists of performing maximum likelihood estimation not for $Z(s)$, but for $KZ(s)$, where K is a K matrix $((n-k) \times n)$, is chosen such that the mean of $KZ(s)$ is zero $E(KZ(s)) = 0$. The rank of $K = n - k$, K matrix is called error constant. An object around θ is given as follows

$$\varphi(\theta, KZ(s)) = ln\left(|K\sum(\theta)K'|\right) + (n-k)ln(2\pi) + Z(s)'K'\left(\sum(\theta)K'\right)^{-1}Z(s)K. \tag{5.33}$$

and

$$\hat{\beta}_{reml} = \left(X'\sum\left(\hat{\theta}_{reml}\right)^{-1}\right)X'\sum\left(\hat{\theta}_{reml}\right)^{-1}Z(s). \tag{5.34}$$

For $E(KZ(s) = 0$ implies that $KZ(s) = 0$ and if $\sum(\theta)$ is positive definite, then Equation 5.33 can be reduced to equation 5.35 (Searle et al., 2009)

$$K'\left(\sum(\theta)K'\right)^{-1}K = \sum(\theta)-1. \tag{5.35}$$

where $\sum(\theta) = \left(X(s)'\sum(\theta)^{-1}X(s)\right)^{-1}$ and $\sum X(s)'\sum(\theta)^{-1}Z(s) = \hat{\beta}.$, thus, $Z(s)'K'\left(K\sum(\theta)K'\right)KZ(s) = r'\sum(\theta)^{-1}r.$ where in GLS residuals

$$r = Z(s) - \left(X(s)'\sum(\theta)^{-1}X(s)^{-1}\right)^{-1}X(s)^{-1}\sum(\theta)^{-1}Z(s).$$

based on the identities as follows

$$KK' = I - X(s)\big(X(s)'X(s)\big)-1X(s)'$$

and

$$KK' = I$$

reduces Equation 5.33 reduces the minus twice the log-likelihood of $K(Z)$ to

$$\varphi_R(\theta, KZ(s)) = ln\left(|\sum(\theta)| + ln|K\sum(\theta)K'|\right) + (n-k)ln(2\pi) + Z(s)'K'\left(K\sum(\theta)K\right)^{-1}Z(s)K - ln(|X(s)'X(s)|) - r'\sum(\theta)^{-1}r + (n-k)ln(2\pi). \tag{5.36}$$

It was also pointed out that $(n-k)\times n$ rows of $I_X(s)(X(s)'X(s))^{-1}$ will meet a REML objective function that differs by a constant which is independent of θ and β (Harville, 1974). The REML objective for minimization is given as

$$\varphi_R(\theta, KZ(s)) = ln\left(|\sum(\theta)| + ln|K\sum(\theta)K'|\right) + (n-k)ln(2\pi) + Z(s)'K'\left(K\sum(\theta)K\right)^{-1}Z(s)K + r'\sum(\theta)^{-1}r + (n-k)ln(2\pi). \tag{5.37}$$

Equation 5.37 differs by the terms $ln(|X(s)'\sum(\theta)^{-1}X(s)|)$ and $kln(2\pi)$ from the REM log likelihood. Similar to the maximum likelihood estimation, a scale parameter can be profiled from $\sum(\theta)$ and the REML estimator of this parameter is given by

$$\hat{\sigma}_{reml} = \frac{1}{n-k}r'\sum(\theta^*)^{-1}r$$

and upon substitution, this will result in minus twice the profile REML log likelihood as follows

$$\varphi_R(\theta, KZ(s)) = ln\left(|\sum(\theta^*)| + ln|K\sum(\theta^*)K'|\right) + (n-k)ln(\hat{\theta}^2) + (n-k)(ln(2\pi)-1). \tag{5.38}$$

Minimum norm quadratic estimation

An alternative method of estimation was proposed by Rao (1979), called the minimum norm quadratic estimation (MINQ). When this method is compared to others, it is shown to be restricted in scope. Despite this limitation, it remains competitive (Kitanidis, 1983).

This method is for special cases, where the variance matrix of the data is linear in its parameters, is given by

$$\sum(\theta) = \theta_1\sum_1 + \cdots + \theta_m\sum_m$$

where $\hat{\theta} = W'F_jW$, for $W = A'Z$, and this is used to find the estimator of θ_j.The minimum norm estimator can be obtained by minimizing $E(\hat{\theta} - \theta)$. Generally, it is subjected to unbiased or variance restriction. This method is appropriate for variance component models; however, in spatial settings, the $\sum(\theta)$ might be a non-linear function of the small scale variation parameter θ.

Moreover, although this method might be easily applicable it is less motivated compared to general procedures of estimation such as maximum likelihood and restricted maximum likelihood (Stein, 1987).

5.4 Measures of Spatial Autocorrelation

There are two measures of spatial autocorrelation existing in literature, namely, Moran's I and Geary's C. These methods are used to investigate if there exist any spatial correlation in the data.

Moran's I

Moran (1950) proposed this method to test for global spatial autocorrelation. Moran's I autocorrelation coefficient is similar to a Pearson correlation coefficient and it quantifies the similarity of response variance among areas that are spatially related (Habyarimana, 2016). The Moran's I test is based on the cross-products of the deviations from the mean. Suppose the deviation is calculated for n observations on variable x at location i, j, then Moran's I coefficient is defined as

$$I = \frac{n \sum_i \sum_j W_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i \sum_j W_{ij} \sum_i (x_i - \bar{x})}. \tag{5.39}$$

where $\sum_i \sum_j W_{ij}$ is the sum of weights of the elements of the weight matrix. W_{ij} are the elements of the weight matrix and $\bar{x}$ is the mean of variable x. The method is approximately normally distributed with a mean of $\frac{1}{N-1}$, where N equals to the number of observations within that specific region. Moreover, Moran's I values lies between -1 and 1, including -1 and 1. When the Moran's I value is zero that indicates the null hypothesis of no clustering, whereas, positive and negative values, respectively, indicate positive (clustering of areas of similar attribute values) spatial correlation and negative (neighboring areas with no similar attribute values) spatial correlation.

Geary's C

Geary (1954) proposed an alternative method for measuring spatial autocorrelation.

This method is based on deviations in responses of each observation with one another and is given by (Ayele, 2013)

$$C = \frac{(n-1)\sum_i \sum_j W_{ij}(x_i - x_j)^2}{2\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}}. \tag{5.40}$$

where n is the number of polygons in the study area of interest. Geary's C method focuses on the similarities between two pairs of regions and the value C varies between 0 and 2. For $C = 0$, it indicates the highest value of positive autocorrelation and $C = 2$ indicates a strong negative autocorrelation.

Moran's I method of measuring spatial autocorrelation is a more global measurement and more sensitive to outlier values of x. While Geary's C is more sensitive to the differences in small neighbourhoods (Ayele, 2013). Generally, both approaches converge to similar conclusions. However, Moran's I is the more preferred compared to Geary's C method (Ayele, 2013).

5.5 Application of Spatial Modelling into the Data Sets

We now apply the spatial generalized linear mxed model into our data sets. Our model can be written as follows:

$$\begin{aligned} g(E(Y(s_i))) = g(\mu_k) = Child \quad age \times \beta_1 + Child \quad gender \times \beta_2 + Level \quad of \quad education \times \beta_3 \\ + Stunting \times \beta_4 + Wealth \quad Index \times \beta_5 + Household \quad has \quad a \quad television \times \beta_6 \\ + Type \quad of \quad place \quad of \quad residence \times \beta_7 \quad + Regions \times \beta_8 + Vitamin \quad A \times \beta_9 + \\ Age \quad of \quad the \quad household \quad head \times \beta_1 0 + \delta(s_i). \end{aligned} \tag{5.41}$$

In this model (5.41), the variables included for analysis are all those that were found to have a significant effect on childhood anaemia when the univariate analysis was performed.

5.5.1 Data Analysis Using Spatial statistics Approach

Our data sets were analysed by fitting generalized linear mixed models using SAS PROC GLIMMIX procedure. There are many covariance structures that were consideredduring the analysis; for example: SP(EXP) (Exponential); SP(EXPA) (Anisotropic Exponential); SP(EXPGA) (2D Exponential), Geometric Anisotropic; SP(GAU) (Gaussian); SP(GAUGA) (2D Gaussian, Geometrically Anisotropic); SP(SPH) (Spherical); SP(LIN) (Linear); SP(LINL) (Linear Log); SP(Matern) (Matron) and SP(SPHGA) (2D

Spherical; Geometrically Anisotropic); and SP(MATHSW) (Matrn(Handcoks-Stein-Wallis)). Furthermore, smooth maps for the prevalence of anaemia in each country made up of regions (provinces), were produced using ArcGIS.

5.5.2 Result interpretations

The scatter plots presented in Figure 5.2 and Figure 5.3 are observed data from Tanzania and Angola, respectively.

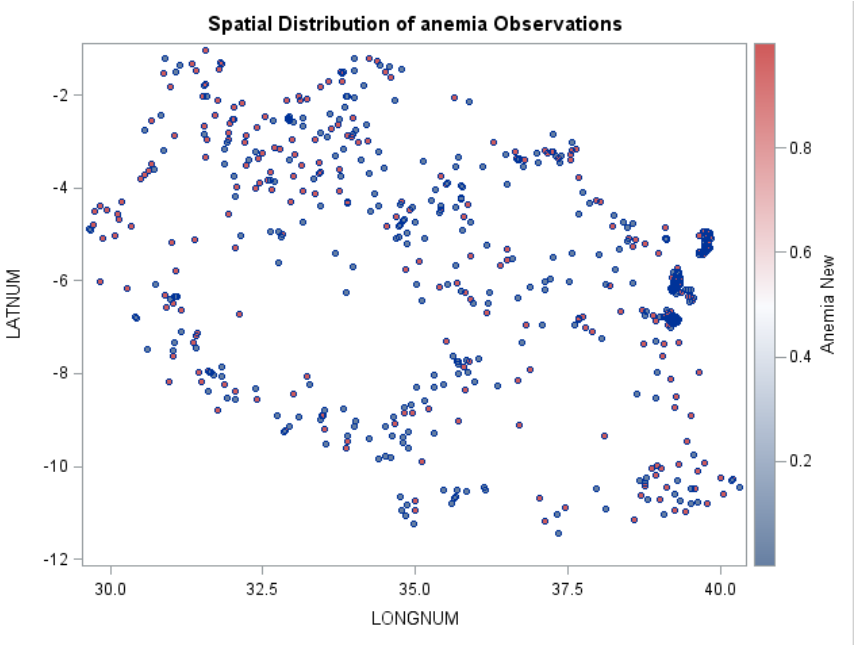


Figure 5.2 – Scatter Plot Anaemia Prevalence (Tanzania)

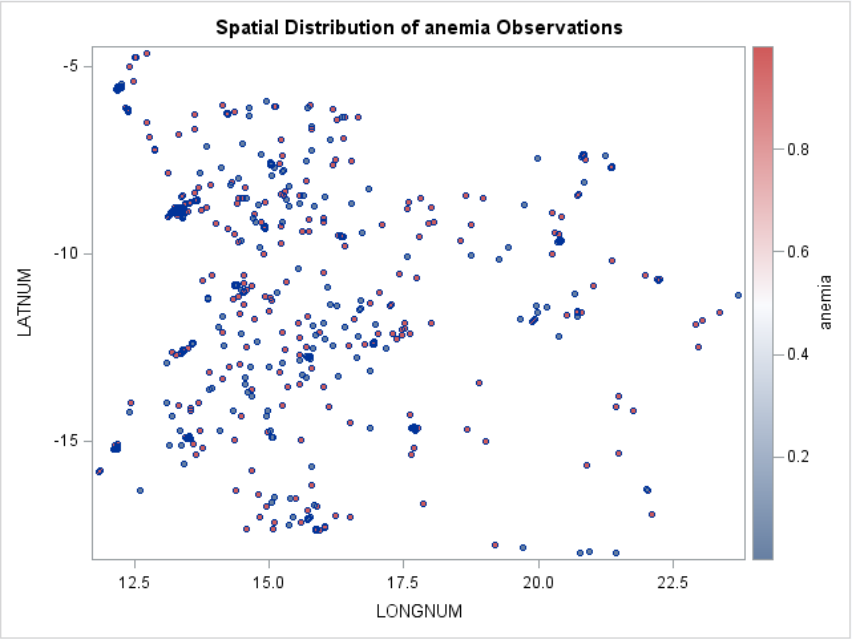


Figure 5.3 – Scatter Plot Anaemia Prevalence (Angola)

As we can observe from both figures, the plots suggest that the distribution is not uniform. Hence, there are direct inferences that can be made about the existence of a surface trend in the data. However, the distributions indicate a random spread response. The spatial autocorrelation is an inferential statistic tool, which is important to test for randomness. Thus, the results of the analysis are always interpreted within the context of its null hypothesis of a random occurrence of events, which can be stated as follows: The attribute being analysed is randomly distributed among the features in the study area. Alternatively, the null can be stated in this way: The spatial processes promoting the observed pattern of values is random chance (Ayele, 2013). For the randomness test, Moran's I and Geary's C tests can be used. Furthermore, The results from these tests of spatial autocorrelation are shown in the tables below.

Table 5.1: Autocorrelation Statistics (Moran's I and Geary's C)

Autocorrelation Statistics (Tanzania)						
Assumption	**Coefficient**	**Observed**	**Expected**	**Std Deviation**	**Z-value**	**Pr>|Z|**
Normality	Moran's I	0.0044	-0.0001	0.00011	39.58	< 0.0001
Normality	Geary's C	0.9939	1.00	0.0022	-2.79	0.0052
Autocorrelation Statistics (Angola)						
Assumption	**Coefficient**	**Observed**	**Expected**	**Std Deviation**	**Z-value**	**Pr>|Z|**
Normality	Moran's I	0.0023	-0.0002	0.00016	15.27	< 0.0001
Normality	Geary's C	1.001	1.00	0.0034	0.412	0.680

According to the results in both data sets, for the Moran's I test statistics, the p-values are very small (p-value < 0.001), which suggests a very strong autocorrelation. Thus, we can reject the null hypothesis of no autocorrelation. The Moran's I Z-values are positive (Z=39.58 in Tanzania and Z=15.27 in Angola), implying that we reject the null hypothesis and conclude that the spatial distribution of high values and/or low values in our data sets is more spatially clustered than would be expected if the underlying spatial process was random (Ayele, 2013). Among all the covariance structures used here for analysis, the Gaussian covariance structure was found to perform better than all the others (AIC=32.035 TDHS and AIC=28.11 in ADHS) in both data sets (Tanzania and Angola). However, inferences can also be drawn based on plotting the corresponding semivariogram. For instance, a graphical presentation of the semivariogram corresponding to the TDHS data is presented in the following diagram (Figure5.4).

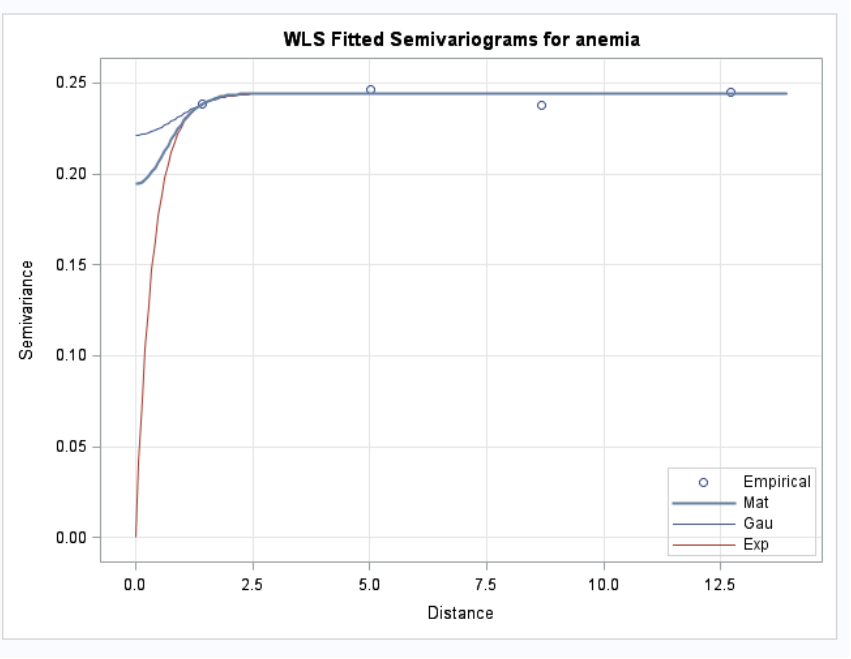


Figure 5.4 – Graphical presentation of the different semivariograms fitted for Anaemia TDHS data set

Now looking at the covariate used in fitting the model, at first, all the covariates fitted on the other SLR and GAMM models were fitted here and later were discarded because they had no significant effect on the outcome variable. The fitted model for both data sets (containing only the significant covariates) is expressed as follows:

$$g(\mu_k) = Child \quad age \times \beta_1 + Child \quad gender \times \beta_2 + Level \quad of \quad education \times \beta_3 + Stunting \times \beta_4 + Wealth \quad Index \times \beta_5 + \delta(s_i). \quad (5.42)$$

When taking into consideration the possibility of spatial dependence effect, child age, level of education, stunting and sex of the child are the significant predictors of childhood anaemia in both countries (p-values < 0.05), as can be seen in Table 5.2. Furthermore, the wealth index has a significant effect (p-value < 0.0001) on childhood anaemia in Tanzania after taking into account the spatial effect. However, in Angola, the standard of living was not found to have any significant effect on childhood anaemia.

Table 5.2: Type III Tests of fixed effects for the GLMM with spatial effect

Tanzania			
Effect	**Num DF**	**F Values**	**P-value**
Gender	1	13.26	0.0003
Child age	4	166.9	<0.0001
Level of education	3	5.46	0.001
Stunting	2	13.48	<0.0001
Wealth Index	4	1.67	<0.0001
Region	29	6.52	<0.0001
Angola			
Gender	1	4.44	0.035
Child's age	4	65.11	<0.0001
Level of education	3	5.30	0.0012
Stunting	2	19.97	<0.0001
Wealth index	4	1.67	0.155
Region	17	4.13	<0.0001

Table 5.3: Solution For fixed effects and the odds ratios (Tanzania)

Parameter	Estimate	Standard error	t-Value	P-value	aOR(95%CI)
Intercept	-0.574	0.218	-2.63	0.0088**	
Gender (ref=male)	-0.182	0.050	-3.64	0.0003 ***	0.847(0.754; 0.919)
Child's age in months (ref = 48-59 moths)					
0-11	1.83	-0.098	18.65	<0.0001***	6.21(5.13;7.53)
12-23	1.42	0.075	18.86	<0.0001***	4.14(3.57;4.80)
24-35	0.63	0.074	8.47	<0.0001***	1.88(1.62; 2.17)
36-47	0.092	0.076	1.22	0.2232	1.09(0.95;1.27)
Mother highest education level (ref =No education)					
Primary	-0.230	0.0672	-3.42	0.0006***	0.932(0.69;0.91)
Secondary	-0.296	0.097	-3.06	0.0022 **	0.678(0.615;0.89)
Higher	-0.730	0.291	-2.51	0.0012 **	0.482(0.27; 0.85)
Wealth index (ref = richest)					
Poorer	0542	0.109	4.79	<0.0001***	1.72(1.39;2.13)
Poorest	0.603	0.112	5.41	<0.0001***	1.83(1.47;2.28)
Middle	0.544	0.106	5.14	<0.0001***	1.72(1.40;2.12)
Richer	0.290	0.096	3.00	0.0027**	1.34(1.11;1.61)
Stunting (ref =severe)					
Moderate	-0.176	0.090	-1.94	<0.0521	0.84(0.70; 1.00)
Nourished	-0.385	0.083	-4.63	<0.0001 ***	0.68(0.58; 0.80)

[a]

[a]Note: The odds ratios as shown on the above Table (5.3) are adjusted for the regions

With reference to male children, in both countries the odds of positive anaemia for female children were lower, OR=0.847, with 95% CI (0.754;0.919) in Tanzania and OR=0.88 and 95%(0.78,0.99) in Angola. Children aged 0-23 months, that is children aged less than 2 years had much higher odds of being anaemic compared to children aged four (48-59 months) in both countries. In Angola, the results were OR

=4.95 and a 95% CI (3.86;6.33) for children aged 0-11 months (or less than 1 year), OR=3.15 , 95% CI (2.62,3.79) for those aged 12-23 months. In Tanzania, we can observe a very similar trend, the OR for children aged 0-11 months and 12-23 months are respectively, OR=6.21, 95% CI (5.13;7.53) and OR=4.14, 95% CI (3.57;4.80). The model reveals that there were higher odds of being anaemic for children who had a mothers with no education compared to children who had mothers with primary, secondary and higher education. In Tanzania, children who had mothers with primary, secondary and higher education had reduced odds of being anaemic respectively by, 7.8% (OR=0.932, 95% CI (0.69;0.91)), 32.2% (OR=0.678, 95% CI (0.615;0.89)) and 51.8% (OR=0.482, 95% CI (0.27;0.85)). In Angola the odds for children with mothers with secondary and higher education were respectively reduced by, 28% (OR=0.78, 95% CI (0.64;0.95)) and 47% (OR=0.53, 95% CI (1.15;1.63)).

Table 5.4: Solution for fixed effects and the odds ratios (Angola)

Parameter	Estimate	Standard error	t-value	P-value	aOR(95%CI)
Intercept	0.864	0.261	3.31	0.0010**	
Gender (ref=male)	-0.131	0.062	-2.11	0.035 *	0.88(0.78; 0.99)
Child's age in months (ref = 48-59 months)					
0-11	1.599	0.126	12.68	<0.0001***	4.95(3.86;6.33)
12-23	1.148	0.094	12.25	<0.0001***	3.15(2.62;3.79)
24-35	0.573	0.092	6.20	<0.0001***	1.77(1.48; 2.13)
36-47	0.313	0.090	3.46	0.0005	1.37(1.15;1.63)
Mother's highest education level (ref =no education)					
Primary	0.060	0.079	0.76	0.446	1.06(0.91;1.24)
Secondary	-0.249	0.101	-2.46	0.0022 **	0.78(0.64;0.95)
Higher	-0.637	0.251	-2.54	0.011*	0.53(032; 0.87)
Wealth index (ref = richest)					
Poorer	-0.115	0.156	-0.74	<0.459	0.89(0.66;1.21)
Poorest	0.022	0.167	0.13	<0.895	1.02(0.74;1.42)
Middle	0.114	0.144	0.79	<0.428	1.12(0.85;1.49)
Richer	-0.038	0.141	-0.27	0.786	0.96(0.73;1.27)
Stunting (ref =severe)					
Moderate	-343	0.104	-3.29	<0.0010**	0.710(0.58;0.87)
Nourished	-0.580	0.094	-6.14	<0.0001 ***	0.560(0.47; 0.67)

[a]

[a]Note: The odds ratios as shown on the above Table (5.4) are adjusted for the regions

Moreover, we can observe from these two tables (5.3 and 5.4) that children from very rich families were much less likely to be anaemic in comparison to children from the poorest families in both countries. In Angola, the children were 1.02 (OR=1.02, 95% CI (0.74;1.42)) times more likely to be anaemic compared to children from rich families and in Tanzania, they were 1.83 (OR =1.83, 95% CI (1.47;2.28)) times more likely

to be anaemic. Furthermore, severely stunted children had greater chance of being anaemic compared to non-stunted and moderately stunted children and this is similar in both countries. In Tanzania, The odds for moderate and non-stunting (nourished) children were respectively reduced by 16% (OR=0.84 , 95% CI (0.70;1.00)) and 32% (OR=0.68, 95% CI (0.58,0.80)). In Angola, the odds were reduced by 29% (OR=0.71 ,95% CI (0.58;0.87)) and 44% (OR=0.56, 95% CI (0.47;0.67)), respectively.

Spatial prediction

The purpose of modelling spatial data is not only to investigate the significant covariates, but also to produce smooth maps of the outcome by predicting at unsampled locations ($S_i, for i = 1, \ldots, n$). Spatial prediction is usually referred to as kriging(Ayele et al., 2013). Generally, spatial interpolation (prediction) is defined as the process of manipulating spatial information to extract new information and meaning from the original data. Usually, spatial analysis is carried out with a geographic information system (GIS). Generally, GIS provides spatial analysis tools for calculating feature statistics and carrying out geoprocessing activities such as data interpolation (Tim Sutton & Mthombeni, 2017; Mitas & Mitasova, 1999). Kriging is an optimal interpolation based on regression against observed values of surrounding data points, weighted according to spatial covariance values. It has many advantages (Ayele, 2013) such as: helping to compensate for the effects of data clustering, assigning individual points within a cluster less weight than isolated data points, giving an estimate of estimation error (kriging variance), along with estimate of the variables, ensuring availability of estimation error which provides a basis for stochasticity and its also allows simulation of possible realization. The maps for the outcome variable (anaemia) that resulted from the kriging interpolation process are presented below.

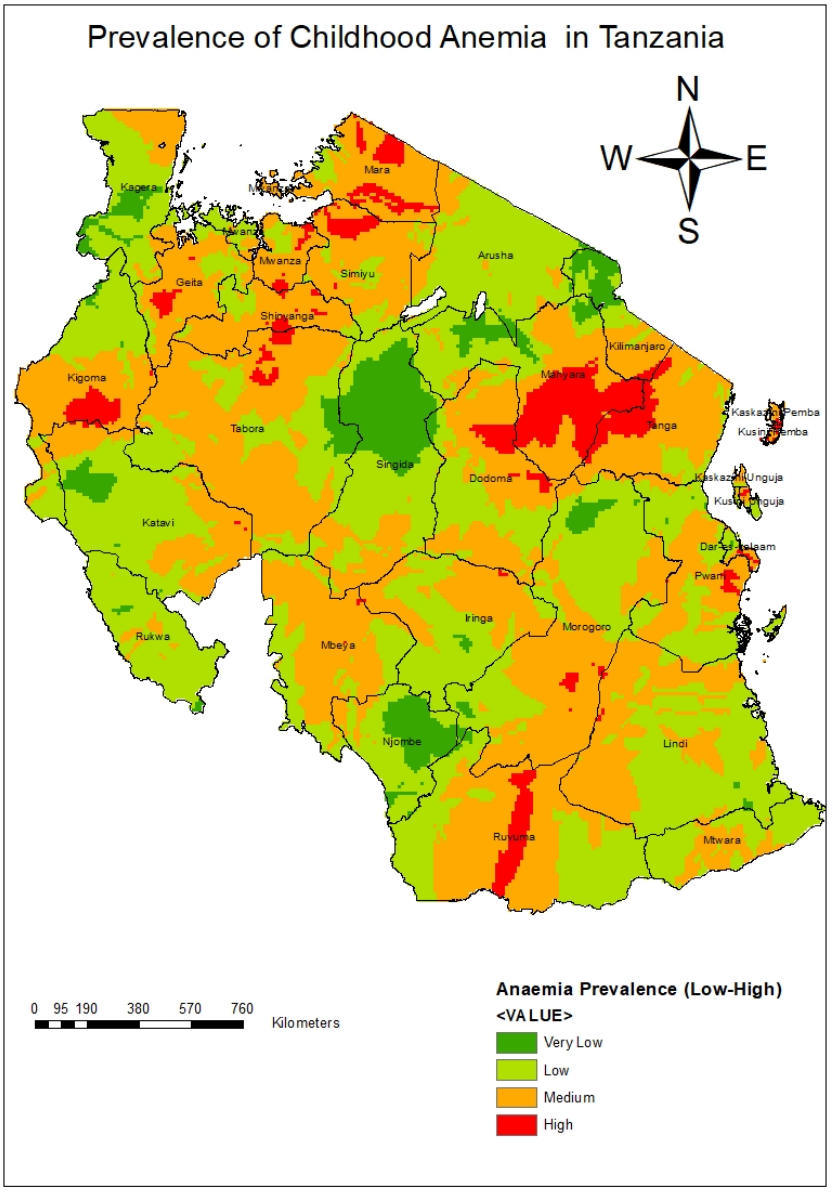


Figure 5.5 – Prevalence of childhood Anaemia in Tazania

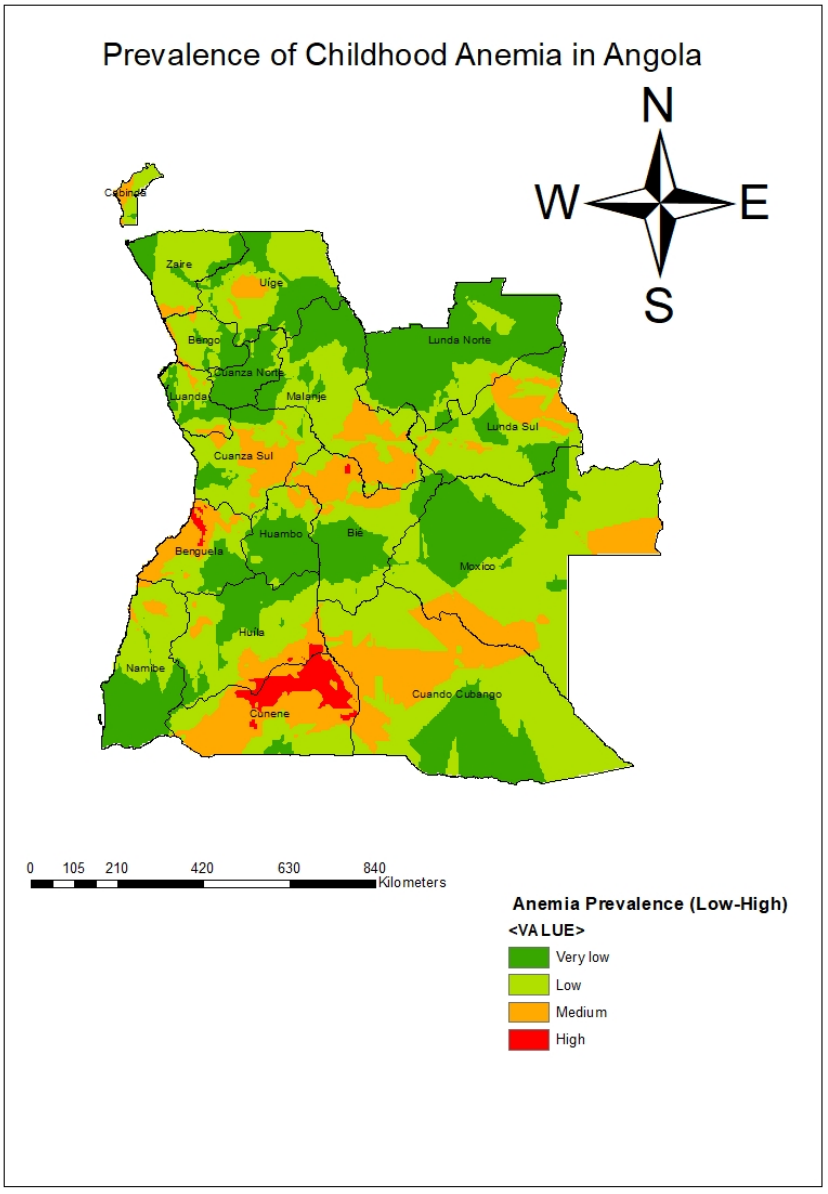


Figure 5.6 – Prevalence of childhood Anaemia in Angola

As we can observe from the above two figures (5.5 and 5.6), the results relate to the first law of geography by Tobler (1979), which states that: everything is related everything, but something adjacent is more influential. The prevalence of anaemia differs by area or region in both countries. After kriging interpolation, the method reveals that individuals who are very close to one another are very likely to behave similarly regarding anaemia status, unlike individuals who are far apart. The prevalence of anaemia was high, medium and low in different areas in the same region; for instance, looking at Tanzania, 5.5, focusing specifically in the Ruvuma region, we can observe three different colours in one region. The same occurs in

Tanga, Kigoma, Tabora, Dodoma, Geita, Mara, Shinyanga, and Manyara regions, but in Angola, only two regions show all the colours (Cunene and Benguela regions). These results correspond strongly to Tobler (1979). Moreover, drawing conclusions in terms of mean prevalence in each region, by inspection, in Tanzania the Ruvuma, Mara, Mwanza, Manyara Simuyu, Kigoma, Geita, Simuyu, Mwanza and Tanga regions show high prevalence of childhood anaemia, while the Mtwara, Iringa, Morogoro, Pwani, Mbeya, Tabora, Dar- es-Salaam and Katavi regions have a moderate prevalence of anaemia and the rest have low prevalence, which are the Lindi, Katavi, Singida, Arusha, Rukwa Kagera region, etc. These conclusions are drawn regarding the dominating colour(s) in each region. However, in Angola (Figure 5.4), we can conclude that the overall prevalence of anaemia in children in the country at the time of the study was low or moderate. There are only two regions (the Cunene and the Benguela regions) that show a high prevalence, with regard to the dominant colours in that region. Most of the regions are dominated by green and lime green, which, according to our legends, show low prevalence. For instance, we can see that the overall prevalence of childhood anaemia in the Lunda Norte, Guanza Norte, Namibe and Hambo regions were found to be very low, as green is dominant. While Lunda Sul, Moxico, Cuanza Sul, Bie, Malanje, and Cuonda Cubango had a moderate prevalence of childhood anaemia. The same trend of interpretation could be followed to draw a conclusion for the remaining regions.
Furthermore, maps for the stunting prevalence were produced to verify the significance of this factor to childhood anaemia. They are presented in the figure 5.7 below.

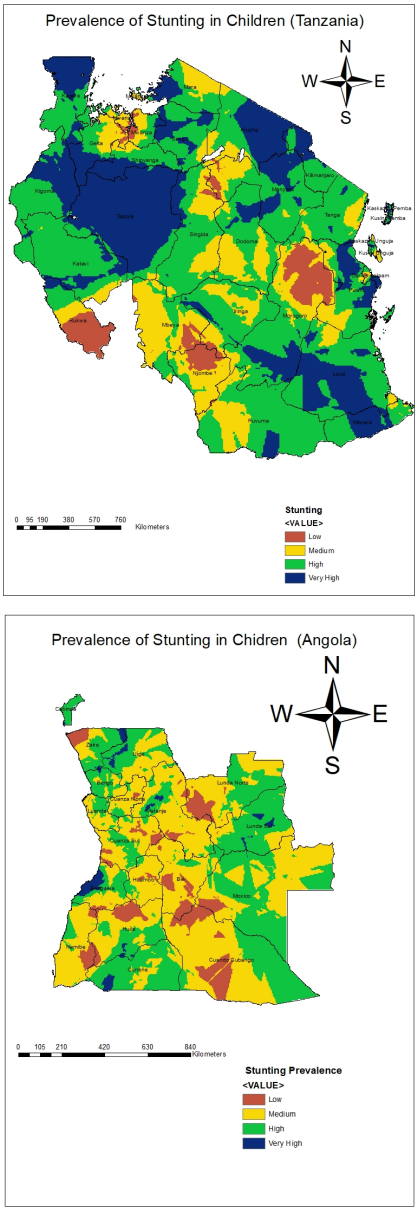


Figure 5.7 – Stunting Prevalence in Children under five in Tanzania and Angola

Following the same style of interpretation anaemia prevalence, we can see that in Tanzania the Tabora, Kigoma, Kagera, Ruvuma, Mtwara, Arusha, Mara, Katavi, Kilimanjaro and Lindi regions are dominated by green and lime green, which implies that there was a high prevalence of stunting there. It was found to be moderate in the Njombe, Geita, Mwanza, Singida, Dodoma, Morogoro, Tanga, Kaskazini Unguja and Kusini Uguja regions, and low in the Rukwa region and Mwanza region. In Angola, most of the regions are dominated by yellow, which implies a moderate prevalence of anaemia. Regions with a high prevalence of stunting are Cunene, Bengo and Lunda regions.

5.6 Summary and Discussion

In this chapter, a spatial model was fitted in the context of generalized linear mixed models as discussed in Section 5.2. A GLIMMIX procedure available in SAS 9.4 was used to analyze the data sets taking into account the different spatial covariance structures. The best covariance structure was confirmed by plotting the corresponding semivariogram rather than considering only the AIC's. There were small differences in the AIC's of exponential and Gaussian covariance structures; hence, the corresponding semivariogram was plotted. Non-significant terms in the model were discarded one at a time which resulted in the final model containing only six predictor variables.

The results from this chapter are consistent with the other results from the two previous chapters, the variables: age of the child, child gender, stunting and level of education of the mother are significantly associated with childhood anaemia. The trend remains unchanged for the age of the child, stunting, and level of education, for children aged less than 24 months (2 years) are still showing higher chances of having anaemia compared to children aged four (48-59 months). Moreover, we note from these results that the chances of a child being anaemic were inversely proportional to the child's age. As the child's age increases, the chances of being anaemic decreases, which can be confirmed by looking at the trend of the OR in both Tables (5.3 and 5.4) which decreases when child age increases. Children who were severely stunted are more likely to be anaemic compared to non-stunted children. Moreover, children with mothers who had no education had a higher chance of being anaemic compared to children born of mothers with primary, secondary and(or) higher education. We can note that the type of place of residence has no significant effect on children anaemia, but the standard of living of a child had a significant effect. The children residing in the poorest families were more likely to be anaemic, and it could be due to their poor dietary guidelines among many other factors.

Moreover, the results presented in the maps (Figure 5.7,) show strong agreement with the findings we obtained about the stunting factor when fitting SLR, GAMM, and SGLMM models. The two maps in Figure 5.7 are consistent with the two maps of anaemia prevalence Figure 5.6 and Figure 5.5. They are consistent in that the regions on the country that show a high prevalence of stunting, also have high prevalence anaemia. This result was highly expected, since stunting was found to be one of the highly significant factors of childhood anaemia. For instance, focusing in Tanzania in both maps, Figure 5.5 and Figure 5.7, we can see that the Ruvuma region, Tabora, Mwanza and Kigoma regions had a high prevalence of childhood anaemia and also

appear to have a high prevalence of stunting. Thus, this strongly emphasizes that children who were suffering from stunting had high chances of being anaemic compared to those who were free from anaemia. Very similar inferences can be drawn in Angola.

Chapter 6

Discussion and Conclusion

The main objective of this study was to flexibly fit statistical models to investigate the demographic and socio-economic factors which are significantly associated with childhood anaemia, in two African countries, Tanzania and Angola. The study is flexible in that we can fit any statistical model that can allow us to model a dichotomous response variable. Furthermore, we wanted to identify the common determinants of childhood anaemia in both countries, and our models successfully fitted the data sets and shown significant factors. The identification of the common causes of childhood anaemia is important in that it will help governments and policymakers in the African countries and other developing countries to know the factors they should focus on in order reduce the prevalence of childhood anaemia and meet the 2025 global targets. A guideline released by WHO for countries to reduce the prevalence of anaemia stated that the goal is to reduce the prevalence of anemia by 50% by 2025. However, the diversity of each country's charecteristics requires targeted handling. The baseline prevalence data in 2012 indicates that it requires continuous monitoring evry year (Rakanita et al., 2020).

To fulfil our objective, three different statistical models were fitted into our data sets: survey logistic regression (SLR) from the class of generalized linear models (GLMs), generalized additive mixed models (GAMMs) and a spatial generalized linear mixed models (SGLMMs). The SGLMMs were fitted to check if there is an effect of spatialiality on the results that were already found in SLR and GAMM and, in addition, to produce maps. The maps will help policymakers to understand the associated factor(s) that are high contributors to childhood anaemia in a specific region in a country and, hence, they will better able to deal with that specific region accordingly.

As explained in previous chapters, the DHS data set has a complex sampling design. Thus, an ordinary logistic regression would not be appropriate in fitting such data

since ordinary logistic regression suggests that data are collected using a simple random sampling method. Since, the DHS data sampling methods involve: clustering, unequal weighing, and stratification, thus survey logistic regression is a convenient model to fit the data from the class of generalized linear models. The data was further fitted to model the response using a generalized additive mixed model, a more flexible model, compared to GLM, and even if it compared to generalized linear mixed models (GLMMs). GAMM is non-parametric semi-parametric modelling approach, which allow some terms to be fitted non-parametrically. Furthermore, the functional form of the model is not specified prior to model fitting, thus making the model to be even more flexible (Lin & Zhang, 1999).

Similar to GLMM, GAMM explores the idea of a statistical model that incorporates random factors into generalized linear models. It adds a random effect or correlation among observations arising from a distribution that is from an exponential family. In addition, the use of GAMM also allows random effects to be properly specified and computed and errors can be correlated. Thus, GAMM exhibits non-constant variability, while simultaneously allowing more than one source of variation. Since the model allows other terms to be fitted non-parametrically (smooth terms), it ultimately more robust in modelling the dependent variable compared to SLR and GLMM. Moreover, because DHS data set is a spatial one, thus, observations closer to one another are more likely to have similar attributes compared to observations which are distant apart (Tobler, 1979). Hence, a spatial generalized linear mixed model was fitted to take into account of the effect of spatiality in the investigation of the predictors of childhood anaemia. The spatial generalized linear mixed model is a powerful model to fit when dealing with data assuming complex survey design and if the aim is to fit that data in a frequentist approach.

According to this study, the key determinants of anaemia in Tanzania and Angola are child age (in months), child gender, the highest education obtained by the mother or guardian and the level of stunting. The effect of the standard of living was also found to be significantly associated with childhood anaemia,in Tanania, whereas in Angola it was found to have no effect throughout all the models fitted. All the models show higher odds for children aged 0-23 months compared to all other age groups, with the odds of a child being anaemic decreasing with an increase in age. The likelihood of a child being anaemic was found to be high for severely stunted children in comparison to non-stunting children in both countries. The models further reveal that children with illiterate mothers had higher odds of being anaemic compared to those with literate mothers or guardians. Furthermore, male children had higher odds of being anaemic compared to female children. These findings are

similar to studies of Habyarimana et al. (2017); Ewusie et al. (2014); Villamor et al. (2000); Oliveira et al. (2015) and Foote et al. (2013). According to the results there was no significant effect found for lack of vitamin supplementation, size of the child at birth, currently breastfeeding and availability of a television in a household.

Assuming that children aged less than 2 years have much higher odds of being anaemic compared to all the other age groups would be hihly reasonable. Children after birth, especially those aged 0-11 months, have a very weak immune system that cannot fight viruses or bacteria, and thus they would be easily affected in an outbreak of a certain virus or bacteria at a particular time or place. Baring in mind that the immune system in a human body plays a crucial role, with a non-functional immune system even minor infections can hold and prove fatal (Parham, 2014). Many studies exist in the literature showing that stunting is one of the common causes of childhood anaemia (Habyarimana et al., 2017; Oliveira et al., 2015; Foote et al., 2013). Stunting is a long-term indicator of poor nutrition, and in severely stunted child can suffer from severe irreversible physical and cognitive damage. In this study, severely stunted children were more likely to develop anaemia compared to the non-stunted and moderalately stunted children. Since anaemia and malnutrition share common causes; thus, stunting is associated with anaemia (Dey et al., 2013; Khan et al., 2016). Another important determinant of anaemia, which was found to be a key determinant was the level of education of the mother or a guardian. To fight infectious diseases, education is a weapon that highly highly recommended. Children with mothers who lack education are more likely to suffer from infectious diseases. This can be further related to the UN sustainable development goals, specificall goal 2 and goal 3.

Giving a quality education to the parents or guardians could slow down the prevalence of childhood anaemia and it can also help the UN in attaining it long term goals which states as follows, goal 2: End hunger by achievin good food security and improve nutrition and promote sustainable agriculture; goal 3: Ensre healthy lives and promote well-being for all at all ages. This can be attained by the government giving quality education about the importance of food security and agriculture. The lack of good nutrient intake causes malnutrition, which further causes diseases like anaemia. Thus, it is of high importance that the government give education to caretakers or parents about nutritional interventions, infecious diseases and how they can be treated at early stages and that could ensure healthy lives and promote well being for all people at all ages, which is one ultimate UN sustanable developments goals to be attained by 2030. People should be given education about food preservation , for it to last longer thus that will reduce hunger promote sustain-

able agriculture. Furthermore, the government should consider distributing food and medicine with high nutrients to help strengthening the immune systems, thus fight anaemia.

In conclusion, the main objective of this study was successfully achieved. The three models fitted into our data sets did show consistency in the investigation of the factors associated with childhood anaemia in both Tanzania and Angola. Furthermore, since we were working with Demographic and Health Survey data sets, as cross-sectional data sets, they do not have information over a period of time, but are instead once-off observational studies. Thus, we could not determine the cause and effect; hence, a longitudinal data set would be recommended for future studies.

References

Agresti, A. (2003). *Categorical data analysis*, vol. 482. John Wiley & Sons.

Agresti, A. (2018). *An introduction to categorical data analysis*. John Wiley & Sons.

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

Allali, S., Brousse, V., Sacri, A.-S., Chalumeau, M., & de Montalembert, M. (2017). Anemia in children: prevalence, causes, diagnostic work-up, and long-term consequences. *Expert review of hematology*, *10*(11), 1023–1028.

Allard, D. (2013). J.-p. chilès, p. delfiner: Geostatistics: Modeling spatial uncertainty.

Allison, P. D. (2012). *Logistic Regression Using SAS: Theory and Application, Second Edition*. Cary, NC, USA: SAS Institute Inc., 2nd ed.

An, A. B., et al. (2002). 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.

Angola (2016). Multiple Indicator and Health Survey. `https://dhsprogram.com/pubs/pdf/SR238/SR238.pdf`. [Online; last modified: 2016].

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

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

Aydın, D., Memmedli, M., & Omay, R. E. (2013). Smoothing parameter selection for nonparametric regression using smoothing spline. *European Journal of Pure and applied mathematics*, *6*(2), 222–238.

Ayele, D. G. (2013). *Use of statistical modelling and analyses of malaria rapid diagnostic test outcome in Ethiopia.*. Ph.D. thesis.

Ayele, D. G., Zewotir, T. T., & Mwambi, H. G. (2013). Spatial distribution of malaria problem in three regions of ethiopia. *Malaria journal*, *12*(1), 207.

Balarajan, Y., Ramakrishnan, U., Özaltin, E., Shankar, A. H., & Subramanian, S. (2011a). Anaemia in low-income and middle-income countries. *The lancet*, *378*(9809), 2123–2135.

Balarajan, Y., Ramakrishnan, U., Özaltin, E., Shankar, A. H., & Subramanian, S. (2011b). Anaemia in low-income and middle-income countries. *The lancet*, *378*(9809), 2123–2135.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review/Revue Internationale de Statistique*, (pp. 279–292).

Bingham, N. H., & Fry, J. M. (2010). *Regression: Linear models in statistics*. Springer Science & Business Media.

Breiman, L., & Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, *80*(391), 580–598.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, *88*(421), 9–25.

Buja, A., Hastie, T., Tibshirani, R., et al. (1989). Linear smoothers and additive models. *The Annals of Statistics*, *17*(2), 453–510.

Calis, J. C., Phiri, K. S., Faragher, E. B., Brabin, B. J., Bates, I., Cuevas, L. E., de Haan, R. J., Phiri, A. I., Malange, P., Khoka, M., et al. (2016). Research article (new england journal of medicine) severe anemia in malawian children. *Malawi Medical Journal*, *28*(3), 99–107.

Chen, C. (2000). Generalized additive mixed models. *Communications in Statistics - Theory and Methods*, *29*(5-6), 1257–1271.
URL https://doi.org/10.1080/03610920008832543

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, *74*(368), 829–836.

Craven, P., & Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische mathematik*, *31*(4), 377–403.

Crawley, J. (2004). Reducing the burden of anemia in infants and young children in malaria-endemic countries of africa: from evidence to action. *The American journal of tropical medicine and hygiene*, *71*(2_suppl), 25–34.

Cressie, N. (1985). Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology*, *17*(5), 563–586.

Cressie, N., & Hawkins, D. M. (1980). Robust estimation of the variogram: I. *Journal of the International Association for Mathematical Geology*, *12*(2), 115–125.

Dallman, P. R., Barr, G. D., Allen, C. M., & Shinefield, H. R. (1978). Hemoglobin concentration in white, black, and oriental children: is there a need for separate criteria in screening for anemia? *The American Journal of Clinical Nutrition*, *31*(3), 377–380.
URL https://doi.org/10.1093/ajcn/31.3.377

De Benoist, B., Cogswell, M., Egli, I., & McLean, E. (2008). Worldwide prevalence of anaemia 1993-2005; who global database of anaemia.

De Pee, S., Bloem, M. W., Sari, M., Kiess, L., Yip, R., & Kosen, S. (2002). The high prevalence of low hemoglobin concentration among indonesian infants aged 3–5 months is related to maternal anemia. *The Journal of nutrition*, *132*(8), 2215–2221.

de Savigny, D., Schellenberg, D., Armstrong Schellenberg, J., Mushi, A., & Mgalula, L. (2003). Silent burden of anaemia in tanzanian children: a community-based study. *Bulletin of the World Health Organization, 2003, v. 81, no. 8*.

Demnati, A., & Rao, J. (2004). Linearization variance estimators for survey data. *Survey Methodology*, *30*(1), 17–26.

Dey, S., Goswami, S., & Dey, T. (2013). Identifying predictors of childhood anaemia in north-east india. *Journal of health, population, and nutrition*, *31*(4), 462.

Dobson, A. (2002). *An introduction to generalized linear models*. CRC press.

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

Dunteman, G. H., & Ho, M.-H. R. (2005). *An introduction to generalized linear models*, vol. 145. Sage Publications.

Eilers, P. H., & Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical science*, (pp. 89–102).

Ewusie, J. E., Ahiadeke, C., Beyene, J., & Hamid, J. S. (2014). Prevalence of anemia among under-5 children in the ghanaian population: estimates from the ghana demographic and health survey. *BMC public health*, *14*(1), 626.

Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall/CRC.

Foote, E. M., Sullivan, K. M., Ruth, L. J., Oremo, J., Sadumah, I., Williams, T. N., & Suchdev, P. S. (2013). Determinants of anemia among preschool children in rural, western kenya. *The American journal of tropical medicine and hygiene*, *88*(4), 757–764.

Friedman, J. H., & Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American statistical Association*, *76*(376), 817–823.

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The incorporated statistician*, *5*(3), 115–146.

Genton, M. G. (1998). Highly robust variogram estimation. *Mathematical Geology*, *30*(2), 213–221.

Genton, M. G. (2001). Robustness problems in the analysis of spatial data. In *Spatial statistics: Methodological aspects and applications*, (pp. 21–37). Springer.

Getaneh, Z., Enawgaw, B., Engidaye, G., Seyoum, M., Berhane, M., Abebe, Z., Asrie, F., & Melku, M. (2017). Prevalence of anemia and associated factors among school children in gondar town public primary schools, northwest ethiopia: A school-based cross-sectional study. *PloS one*, *12*(12), e0190151.

Goovaerts, P., Jacquez, G. M., & Greiling, D. (2005). Exploring scale-dependent correlations between cancer mortality rates using factorial kriging and population-weighted semivariograms. *Geographical Analysis*, *37*(2), 152–182.

Goovaerts, P., et al. (1997). *Geostatistics for natural resources evaluation*. Oxford University Press on Demand.

Graubard, B., Korn, E., & Midthune, D. (1997). Testing goodness-of-fit for logistic regression with survey data. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, (pp. 170–174).

Green, P. J., & Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: a roughness penalty approach*. Chapman and Hall/CRC.

Grizzle, J. E., Starmer, C. F., & Koch, G. G. (1969). Analysis of categorical data by linear models. *Biometrics*, (pp. 489–504).

Habyarimana, F. (2016). *Measuring poverty and child malnutrition with their determinants from household survey data.*. Ph.D. thesis.

Habyarimana, F., Zewotir, T., & Ramroop, S. (2017). Structured additive quantile regression for assessing the determinants of childhood anemia in rwanda. *International journal of environmental research and public health*, *14*(6), 652.

Hall, A., Bobrow, E., Brooker, S., Jukes, M., Nokes, K., Lambo, J., Guyatt, H., Bundy, D., Adjei, S., Wen, S.-T., et al. (2001). Anaemia in schoolchildren in eight countries in africa and asia. *Public health nutrition*, *4*(3), 749–756.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (2011). *Robust statistics: the approach based on influence functions*, vol. 196. John Wiley & Sons.

Handcock, M. S., & Stein, M. L. (1993). A bayesian analysis of kriging. *Technometrics*, *35*(4), 403–410.

Handcock, M. S., & Wallis, J. R. (1994). An approach to statistical spatial-temporal modeling of meteorological fields. *Journal of the American Statistical Association*, *89*(426), 368–378.

Härdle, W. (1990). *Applied nonparametric regression*. 19. Cambridge university press.

Härdle, W., & Kneip, A. (1999). Testing a regression model when we have smooth alternatives in mind. *Scandinavian journal of statistics*, *26*(2), 221–238.

Härdle, W. K., Müller, M., Sperlich, S., & Werwatz, A. (2012). *Nonparametric and semiparametric models*. Springer Science & Business Media.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, *61*(2), 383–385.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American statistical association*, *72*(358), 320–338.

Hastie, T., & Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, *82*(398), 371–386.

Hastie, T. J. (1990). Generalized additive models. In *Statistical models in S*, (pp. 249–307). Routledge.

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

Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in medicine*, *16*(9), 965–980.

Hosmer, D. W., & Lemesbow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in statistics-Theory and Methods*, *9*(10), 1043–1069.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*. Wiley New York.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*, vol. 398. John Wiley & Sons.

Hosseini Shojaei, S. R., Waghei, Y., & Mohammadzadeh, M. (2018). Parameter estimation in spatial generalized linear mixed models with skew gaussian random effects using laplace approximation. *Journal of Statistical Research of Iran JSRI*, *14*(2), 157–169.

indexmodi (2016a). Anaemia. `https://www.indexmundi.com/facts/angola/prevalence-of-anemia`. [Online; last modified: 2016].

indexmodi (2016b). Anaemia prevalence. `https://www.indexmundi.com/facts/tanzania/prevalence-of-anemia`. [Online; last modified: 2016].

Izenman, A. J. (1991). Review papers: Recent developments in nonparametric density estimation. *Journal of the American Statistical Association*, *86*(413), 205–224.

Johnson-Spear, M. A., & Yip, R. (1994). Hemoglobin difference between black and white women with comparable iron status: justification for race-specific anemia criteria. *The American journal of clinical nutrition*, *60*(1), 117–121.

Journel, A. G., & Huijbregts, C. J. (1978). *Mining geostatistics*, vol. 600. Academic press London.

Kapil, R., & Kapil, U. (????). Determinants of anemia in south east asian countries.

Kassebaum, N. J. (2016). The global burden of anemia. *Hematology/Oncology Clinics*, *30*(2), 247–308.

Keen, A., & Engel, B. (1997). Analysis of a mixed model for ordinal data by iterative re-weighted reml. *Statistica Neerlandica*, *51*(2), 129–144.

Kejo, D., Petrucka, P. M., Martin, H., Kimanya, M. E., & Mosha, T. C. (2018). Prevalence and predictors of anemia among children under 5 years of age in arusha district, tanzania. *Pediatric health, medicine and therapeutics*, *9*, 9.

Khan, J. R., Awan, N., & Misu, F. (2016). Determinants of anemia among 6–59 months aged children in bangladesh: evidence from nationally representative data. *BMC pediatrics*, *16*(1), 3.

Khuri, A. I. (2009). *Linear model methodology*. Chapman and Hall/CRC.

Kitanidis, P. K. (1983). Statistical estimation of polynomial generalized covariance functions and hydrologic applications. *Water Resources Research*, *19*(4), 909–921.

Kohn, R., Ansley, C. F., & Tharm, D. (1991). The performance of cross-validation and maximum likelihood estimators of spline smoothing parameters. *Journal of the american statistical association*, *86*(416), 1042–1050.

Koram, K. A., Owusu-Agyei, S., Utz, G., Binka, F. N., Baird, J. K., Hoffman, S. L., & Nkrumah, F. K. (2000). Severe anemia in young children after high and low malaria transmission seasons in the kassena-nankana district of northern ghana. *The American journal of tropical medicine and hygiene*, *62*(6), 670–674.

Kutner, M. H., Nachtsheim, C. J., Neter, J., Li, W., et al. (2005). *Applied linear statistical models*, vol. 5. McGraw-Hill Irwin Boston.

Leal, L. P., Batista Filho, M., Lira, P. I. C. d., Figueiroa, J. N., & Osório, M. M. (2011). Prevalence of anemia and associated factors in children aged 6-59 months in pernambuco, northeastern brazil. *Revista de saude publica*, *45*(3), 457–466.

Lee, H.-S., Cho, Y.-H., Park, J., Shin, H.-R., & Sung, M.-K. (2013). Dietary intake of phytonutrients in relation to fruit and vegetable consumption in korea. *Journal of the Academy of Nutrition and Dietetics*, *113*(9), 1194–1199.

Lin, X., & Zhang, D. (1999). Inference in generalized additive mixed modelsby using smoothing splines. *Journal of the royal statistical society: Series b (statistical methodology)*, *61*(2), 381–400.

Lindsey, J. K. (2000). *Applying generalized linear models*. Springer Science & Business Media.

Linton, O., & Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, (pp. 93–100).

Magalhaes, R. J. S., & Clements, A. C. (2011). Mapping the risk of anaemia in preschool-age children: the contribution of malnutrition, malaria, and helminth infections in west africa. *PLoS medicine*, *8*(6).

Magalhães, R. J. S., Langa, A., Pedro, J. M., Sousa-Figueiredo, J. C., Clements, A. C., & Nery, S. V. (2013). Role of malnutrition and parasite infections in the spatial

variation in children's anaemia risk in northern angola. *Geospatial health*, (pp. 341–354).

Matérn, B. (1960). *Spatial variation*, vol. 36. Springer Science & Business Media.

McCullagh, P. (1989). *Generalized linear models*. Routledge.

McCullagh, P. (2019). *Generalized linear models*. Routledge.

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia medica: Biochemia medica*, *23*(2), 143–149.

McLean, E., Cogswell, M., Egli, I., Wojdyla, D., & De Benoist, B. (2009). Worldwide prevalence of anaemia, who vitamin and mineral nutrition information system, 1993–2005. *Public health nutrition*, *12*(4), 444–454.

Mehta, V. K. (2004). Anemia in urban and rural school girls aged 12-16 years, shimla-a comparative study.

Mildenberger, T. (2012). André i. khuri: Linear model methodology. *Statistical Papers*, *53*(3), 809.

Milman, N. (2011). Anemia—still a major health problem in many parts of the world! *Annals of hematology*, *90*, 369–77.

Ministry of Health, G. E., Community Development, Children (MoHCDGEC)[Tanzania Mainland], N. B. o. S. N. O. o. t. C. G. S. O., Ministry of Health (MoH)[Zanzibar], & ICF (2016). Tanzania demographic and health survey and malaria indicator survey (tdhs-mis) 2015-16.

Mitas, L., & Mitasova, H. (1999). Spatial interpolation. *Geographical information systems: principles, techniques, management and applications*, *1*(2).

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis*, vol. 821. John Wiley & Sons.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, *37*(1/2), 17–23.

Morel, J. G. (1989). Logistic regression under complex survey designs. *Survey Methodology*, *15*(2), 203–223.

Murphy, S. C., & Breman, J. G. (2001). Gaps in the childhood malaria burden in africa: cerebral malaria, neurological sequelae, anemia, respiratory distress, hypoglycemia, and complications of pregnancy. *The American journal of tropical medicine and hygiene*, *64*(1_suppl), 57–67.

Ncogo, P., Romay-Barja, M., Benito, A., Aparicio, P., Nseng, G., Berzosa, P., Santana-Morales, M. A., Riloha, M., Valladares, B., & Herrador, Z. (2017). Prevalence of anemia and associated factors in children living in urban and rural settings from bata district, equatorial guinea, 2013. *PloS one*, *12*(5), e0176613.

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

Newton, C., Warn, P., Winstanley, P., Peshu, N., Snow, R., Pasvol, G., & Marsh, K. (1997). Severe anaemia in children living in a malaria endemic area of kenya. *Tropical medicine & international health*, *2*(2), 165–178.

Ngesa, O., & Mwambi, H. (2014). Prevalence and risk factors of anaemia among children aged between 6 months and 14 years in kenya. *PLoS One*, *9*(11), e113756.

Ngwira, A., & Kazembe, L. N. (2015). Bayesian random effects modelling with application to childhood anaemia in malawi. *BMC public health*, *15*(1), 161.

Oliveira, D., Ferreira, F. S., Atouguia, J., Fortes, F., Guerra, A., & Centeno-Lima, S. (2015). Infection by intestinal parasites, stunting and anemia in school-aged children from southern angola. *PLoS One*, *10*(9).

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

Organization, W. H., et al. (2008). Worldwide prevalence of anaemia 1993-2005: Who global database on anaemia.

Organization, W. H., et al. (2015). The global prevalence of anaemia in 2011.

O'Sullivan, F., et al. (1986). A statistical perspective on ill-posed inverse problems. *Statistical science*, *1*(4), 502–518.

Osungbade, K. O., & Oladunjoye, A. O. (2012). Anaemia in developing countries: burden and prospects of prevention and control. In *Anemia*. IntechOpen.

Parham, P. (2014). *The immune system*. Garland Science.

Pasricha, S.-R., Black, J., Muthayya, S., Shet, A., Bhat, V., Nagaraj, S., Prashanth, N., Sudarshan, H., Biggs, B.-A., & Shet, A. S. (2010). Determinants of anemia among young children in rural india. *Pediatrics*, *126*(1), e140–e149.

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*(3), 545–554.

Peña-Rosas, D. J. P., & García-Casal, D. M. N. (2014). Anaemia. https://www.who.int/nutrition/topics/globaltargets_anaemia_policybrief.pdf. [Online; last modified: 2014].

Pektaş, E., Aral, Y. Z., & Yenisey, Ç. (2015). The prevalance of anemia and nutritional anemia in primary school children in the city of aydın. *Meandros Medical And Dental Journal*, *16*(3), 97–107.

Rakanita, Y., Sinuraya, R. K., Suradji, E. W., Suwantika, A. A., Syamsunarno, M. R. A., & Abdulah, R. (2020). The challenges in eradication of iron deficiency anemia in developing countries. *Systematic Reviews in Pharmacy*, *11*(5), 383–401.

Rao, C. R. (1979). Separation theorems for singular values of matrices and their applications in multivariate analysis. *Journal of Multivariate Analysis*, *9*(3), 362–377.

Rao, J., Kumar, S., & Roberts, G. (1989). Analysis of sample survey data involving categorical response variables: Methods and software. *Survey Methodology*, *15*, 161–186.

Rice, J. A., & Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society: Series B (Methodological)*, *53*(1), 233–243.

Roberts, D., & Matthews, G. (2016). Risk factors of malaria in children under the age of five years old in uganda. *Malaria journal*, *15*(1), 246.

Roberts, G., Rao, N., & Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, *74*(1), 1–12.

Robinson, G. K., et al. (1991). That blup is a good thing: the estimation of random effects. *Statistical science*, *6*(1), 15–32.

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical association*, *88*(424), 1273–1283.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. 12. Cambridge university press.

Sanou, D., & Ngnie-Teta, I. (2012). Risk factors for anemia in preschool children in sub-saharan africa. In *Anemia*. IntechOpen.

Sardy, S., & Tseng, P. (2004). Amlet, ramlet, and gamlet: automatic nonlinear fitting of additive models, robust and generalized, with wavelets. *Journal of Computational and Graphical Statistics*, *13*(2), 283–309.

Schabenberger, O., & Gotway, C. A. (2005). *Statistical methods for spatial data analysis*. Chapman and Hall/CRC.

Schellenberg, D., Schellenberg, J., Mushi, A., Savigny, D. d., Mgalula, L., Mbuya, C., & Victora, C. G. (2003). The silent burden of anaemia in tanzanian children: a community-based study. *Bulletin of the World Health Organization*, *81*, 581–590.

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

Searle, S. R., Casella, G., & McCulloch, C. E. (2009). *Variance components*, vol. 391. John Wiley & Sons.

Sedgwick, P. (2013). Logistic regression. *Bmj*, *347*, f4488.

Semedo, R. M., Santos, M. M., Baião, M. R., Luiz, R. R., & da Veiga, G. V. (2014). Prevalence of anaemia and associated factors among children below five years of age in cape verde, west africa. *Journal of health, population, and nutrition*, *32*(4), 646.

Sharmanov, A. (1998). Anaemia in central asia: demographic and health survey experience. *Food and nutrition bulletin*, *19*(4), 307–317.

Sherman, M. (2011). *Spatial statistics and spatio-temporal data: covariance functions and directional properties*. John Wiley & Sons.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society: Series B (Methodological)*, *47*(1), 1–21.

Simbauranga, R. H., Kamugisha, E., Hokororo, A., Kidenya, B. R., & Makani, J. (2015). Prevalence and factors associated with severe anaemia amongst under-five children hospitalized at bugando medical centre, mwanza, tanzania. *BMC hematology*, *15*(1), 13.

Sloane, D., & Morgan, S. P. (1996). An introduction to categorical data analysis. *Annual review of sociology*, *22*(1), 351–375.

Stein, M. L. (1987). Minimum norm quadratic estimation of spatial variograms. *Journal of the American Statistical Association*, *82*(399), 765–772.

Stevens, G. A., Finucane, M. M., De-Regil, L. M., Paciorek, C. J., Flaxman, S. R., Branca, F., Peña-Rosas, J. P., Bhutta, Z. A., Ezzati, M., Group, N. I. M. S., et al. (2013). Global, regional, and national trends in haemoglobin concentration and prevalence of total and severe anaemia in children and pregnant

and non-pregnant women for 1995–2011: a systematic analysis of population-representative data. *The Lancet Global Health*, *1*(1), e16–e25.

Stoltzfus, R. J., Mullany, L., & Black, R. E. (2004). Iron deficiency anaemia. *Comparative quantification of health risks: global and regional burden of disease attributable to selected major risk factors*, *1*, 163–209.

Tim Sutton, M. S. L. N., Otto Dassau, & Mthombeni, S. (2017). Spatial analysis (interpolation).

Tobler, W. R. (1979). Cellular geography. In *Philosophy in geography*, (pp. 379–386). Springer.

Turner, H. (2008). Introduction to generalized linear models. *Rapport technique, Vienna University of Economics and Business*.

Villamor, E., Mbise, R., Spiegelman, D., Ndossi, G., & Fawzi, W. W. (2000). Vitamin a supplementation and other predictors of anemia among children from dar es salaam, tanzania. *The American journal of tropical medicine and hygiene*, *62*(5), 590–597.

Vogiatzi, M. G., Macklin, E. A., Fung, E. B., Cheung, A. M., Vichinsky, E., Olivieri, N., Kirby, M., Kwiatkowski, J. L., Cunningham, M., Holm, I. A., et al. (2009). Bone disease in thalassemia: a frequent and still unresolved problem. *Journal of Bone and Mineral Research*, *24*(3), 543–557.

Wahba, G., et al. (1985). A comparison of gcv and gml for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statistics*, *13*(4), 1378–1402.

Weisburg, S. (2005). Applied linear regression . hoboken.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.

Wood, S. N., Scheipl, F., & Faraway, J. J. (2013). Straightforward intermediate rank tensor product smoothing in mixed models. *Statistics and Computing*, *23*(3), 341–360.

Worldbank (2016). Angola ao: Prevalence of anemia among children: 5.
URL `https://www.ceicdata.com/en/angola/health-statistics/ao-prevalence-of-anemia-among-children--of-children-under-5`

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

Young, M. F. (2018). Maternal anaemia and risk of mortality: a call for action. *The Lancet Global Health*, *6*(5), e479–e480.

Zeger, S. L., & Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. *Biometrics*, (pp. 689–699).

Zhang, D., & Lin, X. (2003). Hypothesis testing in semiparametric additive mixed models. *Biostatistics*, *4*(1), 57–74.

Zhang, D., Lin, X., Raz, J., & Sowers, M. (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association*, *93*(442), 710–719.

Zurnila, K., Saiful, M., & Selvi, M. (2019). Identifying determinants of child malnutrition using spatial regression analysis. In *IOP Conference Series: Materials Science and Engineering*, vol. 506, (p. 012051). IOP Publishing.

Appendix A

Taylor Series Expansion of $Var(\hat{\beta})$.

Demnati & Rao (2004), Binder's linearised variance estimator,$Var(\hat{\beta})$ is the default variance estimator in most software packages for complex survey data analysis, up to date. Jackknife repeated replication (JRR) or Balanced repeated replication method are also used to estimate the variance covariance matrix,$Var(\hat{\beta})$, for the estimated model coefficients (Heeringa et al., 2017).

The J matrix of the second derivative is used in the computation of the variance estimators for the pseudo-maximum likelihood estimates of finite population parameters in the logistic regression model. The J matrix defined as follows,

$$\begin{aligned} J = \left[\frac{\partial^2 \beta PL(\beta)}{\partial^2 \beta}\right] |\beta = \hat{\beta} \\ = \sum_h \sum_\alpha \sum_i X'_{h\alpha i} X_{h\alpha i} W_{h\alpha i} \hat{\pi}_{h\alpha i}(\beta) \left(1 - \hat{\pi}_{h\alpha i}(\beta)\right) \end{aligned} \tag{6.1}$$

Where, h, is a stratum index. α, is a cluster(SECU) index within stratum h and i is an index for individual observations within cluster α.
Because of the complex survey sample data analysis, J^{-1} is not equivalent to variance-covariance matrix of the pseudo-maximum likelihood parameter estimates as in the case of simple random sample setting. Instead a different matrix for J is used to incorporate for the variance estimator, that matrix is said to be a Sandwich-type variance, expressed as,

$$Var(\hat{\beta}) = J^{-1} var\left[S(\hat{\beta})\right] J^{-1} \tag{6.2}$$

$var(S(\hat{\beta}))$ is a symmetric variance-covariance of $p+1$ estimating equations. Whereby each of these $p+1$ estimating equations is a summation over strata, clusters and elements of the individual scores for the n survey respondents.

Thus, variance and covariance of the estimating $p+1$ is given by (Heeringa et al., 2017),

$$Var(S(\hat{\beta})) = \frac{n-1}{n-(p+1)} \sum_{h=1}^{H} \frac{a_h}{a_{h-1}} \sum_{a=1}^{a_h} (S_{h\alpha} - \bar{S}_h)'(S_{h\alpha} - \bar{S}_h) \tag{6.3}$$

A standard formula for stratified sampling of ultimate clusters was applied because each estimating equation is a sample of total respondent.

For a very large sample sizes (large n),the $var(S(\hat{\beta}))$

$$Var(S(\hat{\beta})) = \sum_{h=1}^{H} \frac{a_h}{a_{h-1}} \sum_{a=1}^{a_h} (S_{h\alpha} - \bar{S}_h)'(S_{h\alpha} - \bar{S}_h) \tag{6.4}$$

where,

$$S_{h\alpha} = \sum_{i=1}^{n_\alpha} S_{h\alpha i} = \sum_{i=1}^{n_a} W_{h\alpha i} \left(Y_{h\alpha i} - \hat{\pi}_{h\alpha i}(\beta)\right) X'_{h\alpha i}$$

And

$$\bar{S}_h = \frac{1}{a_n} \sum_{a=1}^{a_h} S_{h\alpha}$$

For interested readers, a detailed theory about the Taylor series estimation of $var(\hat{\beta})$can be found on the book by (Heeringa et al., 2017)

Appendix B

The method of Iterative Reweighted Least Squares, Newton Raphson and Fisher Scoring are discussed below. As discussed, maximum likelihood estimates can be obtained by applying the IRLS, NR or a FS method to the score function

$$\frac{\partial l(\beta_j)}{\partial \beta_j} = \sum_{i=1}^{n} (y_i - \mu_i) W_i \frac{\partial \eta_i}{\partial \mu_i} x_{ij} \tag{6.5}$$

The Taylor expansion is generally given written as follows

$$f(x_0) + (x_1 - x_0) f'(x_0) + \frac{(x_1 - x_0)^2}{2!} f''(x_0) + \frac{(x_1 - x_0)^3}{3!} f'''(x_0)$$

The Newton Raphson method is an iterative method whose derivation is based on the second term of the Taylor series expansion of the log likelihood function.
Assuming that the law of higher order terms is negligible and considering the first two terms of the Taylor expansion, we have

$$f(x_0) + (x_1 - x_0) f'(x_0)$$

That is equivalent to,

$$f(x_0) = -(x_1 - x_0) f'(x_0)$$

Simplifying for x_1,

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

That is the basis for the iterative updating equation in the Newton Raphson estimation algorithm.

Following the above process, using the score function of the log-likelihood as the basis for parameter estimation, the Newton Raphson method yield the following

$$\beta_r = \beta_{r-1} - \left(\frac{\partial l(\beta_{r-1})}{\partial \beta} \right) \left(\frac{\partial^2 \beta}{\partial^2 l(\beta_{r-1})} \right) \tag{6.6}$$

If we let $S(\beta_{r-1}) = \frac{\partial l(\beta_{r-1})}{\partial \beta}$ and $S'(\beta_{r-1}) = \frac{\partial^2 \beta}{\partial^2 l(\beta_{r-1})}$, Therefore 6.6 becomes

$$\beta_r = \beta_{r-1} - S(\beta_{r-1}) \left[S(\beta_{r-1})\right]^{-1} \tag{6.7}$$

$S(\beta_{r-1})$ is the partial derivative of the score equation with respect to β, evaluated at (β_{r-1}) and is referred to as Hessian matrix.

The Fisher-Scoring Method and Iterative Re-weighted Least Squares method

Fisher's Scoring method is used as an alternative method to solve for the unknown parameters in the log-likelihood estimating equation. It is similar to method by Newton Raphson, the only difference FS method uses the expected value of the Hessian matrix based on the information matrix. By some complicated procedures it can be shown that the inverse $I(\beta)$ is

$$I(\beta) = E\left(\frac{\partial^2 l}{\partial \beta_i \beta_j}\right) = E\left(\frac{\partial l}{\partial \beta_i}\right)\left(\frac{\partial l}{\partial \beta_j}\right) \tag{6.8}$$

There are advantages of using the expected Hessian rather than Hessian itself (Heeringa et al., 2017)

The Iterative Re-weighted Least Squares method make use of the Fisher Scoring method to find the unknown maximum likelihood estimates. Fishers-Scoring method at each step can be regarded as kind of the weighted least squares procedure. In the context of generalized linear models, Fishers coring method is also called an Iterative re-weighted least squares method.

Appendix C

The following are the SAS and R codes used in this study.

The final Survey logistic regression model :

Tanzania DHS

```
proc surveylogistic data=tdhsdata;
cluster V001;
weight V0055;
strata V023/list;
class b4 b8 V404A stunting v025 v101 v106(ref="No education") v190ab v121 /param=glm;
model HW577A(event='anemic')= b4 b8 V404A stunting v025 v101 v106 v152 v190ab
v121;
output out=pred p=phat lower=lcl upper=ucl predprob=(individual crossvalidate);
ods output Association=Association;
run;
```

ANGOLA DHS

```
proc surveylogistic data=adhsdata;
cluster V001;
weight V0055;
strata V023/list;
class b4 b8 V404 stunting v025 v101 v106(ref="No education") v190bb v121 /param=glm;
model HW577b(event='anemic')=b4 b8 V404 stunting v025 v101 v106 v190bb v121
v152; run;
```

GAMM model Fitted in both Angola and Tanzania data sets

```
Anemia∼ Gender+Residence Type+Television+Breastfeeding+Education
Level+Gender Household head age+stunting+V024+Wealth index+s(Child
```

```
age,bs="cr")+s(V012,bs="cr"),random= (1|V001),weights=NULL,na.action=na.omit,
family = binomial(link=logit), data = datas1,subset=NULL,knots=NULL,
REML=TRUE,verbose=0L,drop.unused.levels=TRUE)
```

SGLMM model Fitted in both Angola and Tanzania data sets using SAS

```
proc glimmix data=gpst ;
class v024 B4 V025 b8 V106 V151 Stunting V190;
model hw577a (event='anemic') = v024 B4 V025 b8 V106 V151 Stunting V190 /solu-
tion dist=binary link=logit oddsratio (at msesc = .5 unit msesc =.1);
random int/type=SP(GAU)(latnum longnum) sub=v001;
run;
```