

# Prevalence and risk factors associated with malaria infection in children under the age of fourteen years in Kenya

By

Dona Namukowa Ongoma

*A thesis submitted in fulfillment of the requirements  
for the degree of Masters in Statistics  
in the School of Mathematics, Statistics and Computer Science  
at the University of KwaZulu-Natal  
September 2017*



# Declaration

I, Dona Ongoma declare that, the work presented under this thesis titled ‘Prevalence and risk factors associated with malaria in children under the age of fourteen years in Kenya’ is my original research work. I confirm that:

- The work was undertaken for the candidature of a masters degree at the university.
- The thesis has not been submitted for any degree or qualification at any other university.
- I have duly acknowledged the published work of other researchers and referenced the sources of pictures, graphs, and data used in the research work.

---

Ms. Dona N. Ongoma

---

Date

---

Prof Henry G. Mwambi

---

Date

# Acknowledgments

First and foremost, I would like to thank my supervisor Prof. Henry Mwambi for his guidance, patience, encouragement and support during the period of my study. I am truly grateful to Dr. Oscar Ngesa for providing the data, laying the foundation for the study and for his guidance during the study.

I am deeply honored by the support of the entire School of Mathematics, Statistics and Computer science department staff and postgraduate students. In particular I would like to appreciate the entire G14 office members. Thank you for the good moments away from research.

To my friends Maggie, Justine, Albert, Penia, Vuyo, Davies and Edna, for being the “best”. I could always count on you for support, love, cheers and prayer.

Lastly to my family. Thank you for never giving up.

To God almighty, thank you for your immense love and care.

# Abstract

Despite various efforts by multilateral agencies and governments to prevent, control and eliminate malaria, it continues to be a major plague with more than 300 million at risk of infection worldwide. In Kenya, it is still a leading cause of morbidity and mortality, affecting more than 70% of the population. Malaria is endemic in most parts of the country with either high to moderate transmission patterns or seasonal epidemic patterns. Statistics from the ministry of health records show that it accounts for about 30% of outpatient care and 20% of the admissions in hospitals nationwide. Therefore, it is important to constantly review and understand the epidemiology, and the risk factors associated with malaria infection. Such efforts will help the government and the multilateral agencies in their planning, monitoring and evaluation efforts to control and eventually eradicate malaria.

The main objective of the study was to identify the risk factors associated with malaria infection in children under the age of fourteen years. To achieve this, three different statistical methods for analysing complex survey data with a binary outcome, with both linear and non-linear covariates were used. The data used was obtained from a household survey conducted by the government of Kenya in the year 2010 during the peak malaria transmission period. A total of 240 clusters with 30 households in each cluster was sampled from highland epidemic, lake endemic, coastal endemic, seasonal risk and low risk epidemiological regions of Kenya. Probability weights were assigned at each stage of sampling to provide accurate estimates. A total of 11,310 children between 3 months and 14 years were the identified study subjects.

To account for the complexity in the sampling design, survey logistic regression (SLR), a special model under the generalized linear models (GLM) framework was used to identify the risk factors associated with childhood malaria infections. However, the SLR model fails to account for variability arising from correlation between subjects from the same household and clusters. Therefore, the generalized linear mixed effects model (GLMM) was also applied to the data. To relax the assumptions of normality and linearity in the two parametric models, the semi-parametric generalized additive mixed effects model (GAMM), was finally applied to the data.

The findings of the study showed that age of the child, cluster altitude in metres, region, place of residence, type of housing structure, availability of toilet facilities, use of insecticide treated bed nets and mother's level of education were the key determinants of the risk malaria. In the fight to control and eliminate malaria, the results of the study can aid in policy formulation.

# Contents

- 1 Introduction** **1**
  - 1.1 Background . . . . . 2
  - 1.2 Malaria in Kenya . . . . . 3
  - 1.3 Risk factors associated with malaria . . . . . 6
    - 1.3.1 Demographic factors . . . . . 6
    - 1.3.2 Social-economic factors . . . . . 8
    - 1.3.3 Geographic and environmental factors . . . . . 10
  - 1.4 Objectives of the study . . . . . 11
  - 1.5 Thesis outline . . . . . 12
  
- 2 Data presentation and description** **13**
  - 2.1 Introduction . . . . . 13
  - 2.2 Background . . . . . 13
  - 2.3 The Data sets . . . . . 15
  - 2.4 Exploratory data analysis . . . . . 16
    - 2.4.1 Tests of association . . . . . 26
    - 2.4.2 Summary . . . . . 28
  
- 3 The Generalized linear model** **29**

3.1	Introduction . . . . .	29
3.1.1	The model structure . . . . .	30
3.1.2	Exponential family . . . . .	31
3.1.3	Maximum likelihood estimation . . . . .	34
3.1.4	Assessing the fit of a model . . . . .	39
3.1.5	Model selection . . . . .	40
3.2	Logistic regression model . . . . .	40
3.2.1	Introduction . . . . .	40
3.2.2	Binary logistic regression . . . . .	41
3.2.3	Fitting the logistic regression model . . . . .	42
3.2.4	Model selection and model fit . . . . .	43
3.2.5	Odds ratio . . . . .	44
3.3	Survey logistic regression model . . . . .	45
3.3.1	Estimation of parameters . . . . .	46
3.3.2	Variance estimation . . . . .	47
3.3.3	Model selection and fit . . . . .	54
3.4	Analysis of data using survey logistic regression procedure . . . . .	55
3.5	Summary and discussion . . . . .	64
<b>4</b>	<b>The Generalized Linear Mixed Models</b>	<b>68</b>
4.1	Introduction . . . . .	68
4.1.1	The model structure . . . . .	69
4.1.2	Estimation of the parameters . . . . .	70
4.1.3	Inference for GLMM's . . . . .	76
4.1.4	Advantages and disadvantages of GLMMs . . . . .	81
4.1.5	GLMMs for binary response data . . . . .	81

4.1.6	Data analysis . . . . .	82
4.1.7	Summary and discussion . . . . .	91
<b>5</b>	<b>Semi parametric regression approach</b>	<b>94</b>
5.1	Introduction . . . . .	94
5.2	Generalized additive model . . . . .	96
5.2.1	Additive model . . . . .	97
5.2.2	Smoothing Techniques . . . . .	98
5.2.3	Generalized additive models . . . . .	105
5.2.4	Generalized additive mixed effects models . . . . .	107
5.3	Application of GAMM to the dataset . . . . .	110
5.3.1	Summary and discussion . . . . .	117
<b>6</b>	<b>Discussion and conclusion</b>	<b>119</b>
6.0.2	Discussion . . . . .	119
6.0.3	Conclusion . . . . .	122
<b>A</b>	<b>SAS and R codes for the models</b>	<b>124</b>

# List of Tables

2.1	Prevalence of malaria by selected covariates . . . . .	17
2.2	Cross-tabulation of malaria status with selected covariates . . . . .	27
3.1	Type 3 analysis results of the univariate survey logistic regression analysis . . . . .	57
3.2	Type 3 analysis of the final multivariate survey logistic regression . . . . .	59
3.3	Parameter estimates, adjusted Odds ratio (aOR) and 95 % Confidence intervals for the final survey logistic model . . . . .	60
4.1	Covariance parameter estimates . . . . .	83
4.2	Tests of covariance parameters based on the likelihood . . . . .	83
4.3	Type 3 analysis of the fixed effects for the final multivariate GLMM . . . . .	85
4.4	Parameter estimates, odds ratio and 95 % confidence intervals for the final GLMM with household random effect . . . . .	87
4.5	Parameter estimates, odds ratio and 95 % confidence intervals for the final GLMM with household random effect . . . . .	88
4.6	Covariance parameter estimates for household nested within cluster random effects . . . . .	90
4.7	Comparison of the AIC estimates for the models . . . . .	90
5.1	Anova results for model 1 . . . . .	113

5.2	Anova results for model 2 . . . . .	113
5.3	Parameter estimates, odds ratio and 95 % Confidence intervals for the parametric covariate in the final GAMM . . . . .	114
5.4	Approximate significance of the smooth terms for GAMM . . . . .	115

# List of Figures

1.1	Map of the malaria endemic regions of Kenya . . . . .	4
2.1	Prevalence of malaria by age . . . . .	19
2.2	Prevalence of malaria by region . . . . .	20
2.3	Prevalence of malaria by place of residence . . . . .	21
2.4	Prevalence of malaria by cluster altitude in metres . . . . .	21
2.5	Prevalence of malaria by wealth quantile . . . . .	22
2.6	Distribution of malaria status by source of drinking water and type of toilet facility . . . . .	23
2.7	Distribution of malaria by ownership of household assets . . . . .	23
2.8	Distribution of malaria status by wall material, floor material and roof material used in household construction . . . . .	25
2.9	Prevalence of malaria by mother's highest education level . . . . .	26
3.1	Interaction of mother's highest education and floor material use in household construction . . . . .	62
4.1	Diffogram of floor material and Number of nets interaction effects . . . . .	88
5.1	Smoothing components for malaria status with age and cluster altitude . . . . .	116

# Chapter 1

## Introduction

Malaria is an infectious disease that continues to cause morbidity and mortality in children. According to the WHO (2015) statistics, there were 214 million reported malaria cases with 438,000 deaths of with the African region accounting for 88%. This region has temperate climatic conditions that provides a conducive environment for the *Plasmodium* parasite that is responsible for malaria, to breed and mature rapidly. The region is also characterized by abject poverty and under-development (Gilles, 1981; Gallup and Sachs, 2001). In 2013, malaria was the leading cause of death in Kenya resulting in 12.2% of the total recorded deaths and accounting for 20% of the malaria incidences countrywide (Ministry of Health, 2015).

Malaria has a great impact on the livelihoods of individuals as well as the government due to its effects on the economic growth. Several governmental and multilateral programs have been established worldwide with the aim of combating and eliminating malaria. Some of the measures that have been implemented include vector control through the use of insecticide treated mosquito nets, indoor residual spraying, and basic care of the environment such as clearing bushes and stagnated waters near homes. Prompt and accurate diagnosis and

treatment of malaria using artemisinin based combination therapy (ACT), could also help control and manage infections. As a result of these efforts, malaria cases and mortality due to malaria declined by 18% and 48% respectively between the years 2000 and 2015 (WHO, 2015). There is still a need to study its epidemiology, and determine its prevalence and risk factors particularly in Kenya and the sub-Saharan region. This will aid in identifying challenges in intervention programs and policy reformulation.

## 1.1 Background

### Lifecycle of the parasite

Malaria is a vector borne disease, transmitted from one person to another by the female *Anopheles* mosquito. It is caused by protozoan parasites of the genus *Plasmodium*. The most common species within the African region is the *Plasmodium falciparum* (Greenwood et al., 2005). During a blood meal, the infected female mosquito bites the human host injecting in to their blood stream the sporozoite form of the parasite. These reproduce asexually in the liver cells to merozoites that invade the red blood cells and again multiply asexually releasing more merozoites (Bray and Garnham, 1982; Cox, 2010). This cycle continues resulting in the invasion of many more uninfected red blood cells and may lead to severe malaria complications. Some of the merozoites mature sexually to become gametocytes that are taken up by the female anopheles mosquito during a blood meal.

In the mosquito, the ingested gametocytes differentiates, and matures into sporozite form of the parasite that invades the salivary glands, ready to be injected on the human host during feeding (Ghosh et al., 2000).

Environmental conditions such as temperature, rainfall and humidity play an important role

in the survival of the mosquito, in the development of the malaria parasite in the mosquito, and in the breeding and feeding habits of the vector (Hunter, 2003; Gemperli, 2013; Beck-Johnson et al., 2013).

Some of the general symptoms of malaria include fever, general body weakness, vomiting, shivers and chills, and joint aches. Severe malaria often caused by *plasmodium falciparum* may lead to impaired consciousness, multiple convulsions, respiratory distress, acute pulmonary oedema, shock, kidney failure, clinical jaundice and vital organ dysfunction (Ministry of public health and sanitation, 2010). Prompt diagnosis, treatment and vector control measures are important to prevent and reduce severe malaria cases that may result in death.

## 1.2 Malaria in Kenya

Kenya is located in the Eastern parts of Africa, bordered by Ethiopia to the North, Sudan to the North West, Somalia to the East, Tanzania to the South and Uganda to the West. Administratively, it used to be divided into 8 provinces with 158 districts which are in turn divided in to divisions, locations and sub-locations. With a new constitution promulgated in 2010, 47 counties were introduced, as administrative units, headed by governors. The 2009 population and housing census estimates the population to be 38.6 million people, with 32 percent living in urban areas and 43 percent of the population being under 15 years (KNBS, 2010).

It enjoys tropical climate, with four ecological zones namely; hot and humid climate at the coast, temperate at the inland and higher altitudes, very dry at the north and north

eastern parts of the country, and cool at the highland areas. It experiences two rainy seasons, with long rains during the months of April to June and the short rains from October to December. The temperatures, altitude, rainfall patterns and proximity to Lake Victoria and the Indian Ocean impacts greatly on the malaria epidemiological zones.

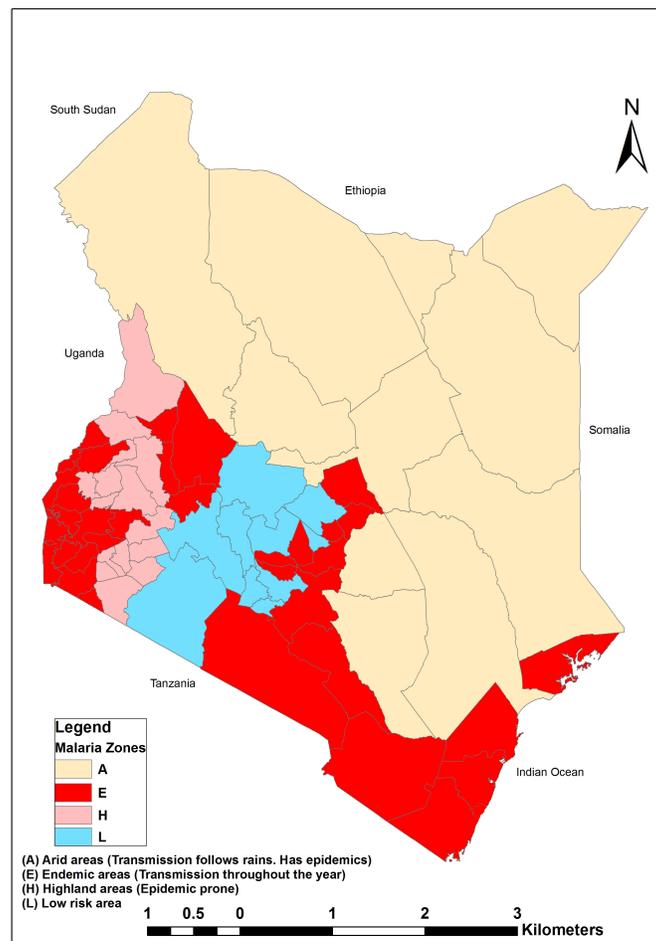


Figure 1.1: Map of the malaria endemic regions of Kenya

The four malaria epidemiological zones as shown in Figure 1.1 are:

**Endemic:** These are areas around Lake Victoria in the western parts of Kenya and area within the coastal region surrounded by the Indian Ocean. They lie within an altitude ranging between 0 and 1,300 metres. Malaria transmission is perennial due to temperature,

humidity and rainfall effects. The life cycle of the vector is short with high survival rates due to the favorable climatic conditions.

**Seasonal transmission:** These are the arid and semi-arid areas around the northern and south-eastern parts of the country. Transmission of malaria occurs during the rainy season, where water pools form breeding grounds for the vector and the high temperatures ensure survival of the mosquito.

**Highland epidemic prone areas of Kenya:** These are western highlands of Kenya where transmission is seasonal. These areas are characterized by low temperatures, that are not suitable for vector breeding. In the rainy season, temperatures tend to increase providing suitable conditions for vector breeding.

**Low risk malaria areas:** This region include Nairobi and the central highlands of Kenya where transmission is relatively low due to low temperatures.

About 80% of the population is at risk for malaria with 27% living in the epidemic and seasonal malaria prevalence (Ministry of Health, 2015). The Ministry of Health reports that it accounts for 30% of the outpatient cases in health facilities and 19% of hospital admissions (Division of Malaria control, 2009). In recognition of the health burden of malaria, the Kenyan government has supported and implemented various programs to help control malaria. The national malaria control program (NMCP) and the president's malaria initiative are some initiatives that have been set-up between the government and other donor organizations. The NMCP in particular set up a 10 year strategic plan, involving all sectors of government, whose objective is to have a malaria free country achieved through internationally approved intervention measures. Routine household surveys such as the malaria indicator survey and demographic and health surveys are conducted to evaluate whether the objectives of such programs have been achieved. The Kenya malaria indicator survey of 2010, was carried out to determine the progress towards reduction in malaria cases and

deaths articulated in the national malaria strategy. This survey provided the data for this study.

A number of studies have been carried out in Kenya to determine prevalence and risk factors for malaria. Most of which are hospital based investigating clinical malaria, community based or undertaken in a particular endemic region (Atieli et al., 2009, 2011; Ernst et al., 2009; Nevill et al., 1996; Njau et al., 2014; O'Meara et al., 2008; Ter Kuile et al., 2003). Unlike other surveys, this is a nationally representative survey carried out during malaria transmission peak period and included all children below the age of 14 years. The previous surveys were limited to children under 5 years.

### **1.3 Risk factors associated with malaria**

There are several risk factors that have been associated with malaria infections and they include; age, gender, housing type/structure, proximity to vector breeding sites, ecological location, household crowding, room size, use of vector control measures such as antimalarial spraying, use of ITN bed nets, gender and wealth. These factors can be generalized in to demographic, geographical factors, socio-economic factors and environmental factors.

#### **1.3.1 Demographic factors**

The demographic factors associated with malaria status in children are: age, gender, and family size.

##### **Age**

Malaria is an infection that affects people of all ages but its severity differs from one individual to another depending on the immunity, proximity to the vector breeding sites, geographic

and ecological factors. Genetic factors may also influence an individual's susceptibility to the disease, progression of the infection in the individual and ultimately the outcome of the infection (Fortin et al., 2002).

One of the major causes of mortality and morbidity in children worldwide is malaria. Children, particularly those under five years are susceptible due to their weak immune system that is still developing. According to Unicef (2007), 1 in 10 deaths worldwide was as a result of malaria while in sub-Saharan Africa malaria accounts for 1 in 5 deaths. Several research studies have shown that during the malaria peak periods, the burden is higher in younger age groups (Carneiro et al., 2010; Molineaux et al., 1980). This may be attributable to the acquired immunity by the older populations as the malaria intensity progresses. Mortality as a result of severe malaria differs with the age of the patients, but in high transmission areas, the intensity is again higher among young children (Olliaro, 2008). However, due to intensive intervention methods, studies have shown a shift in malaria morbidity from the young children (under 5 years) to older children (between 5 – 9 years) (Ceesay et al., 2008; Greenwood et al., 1987; O'Meara et al., 2008; Peterson et al., 2009; Schellenberg et al., 2004).

### **Gender**

Evidence from literature is inconclusive about the effect of gender on malaria risk. Most studies find no association between gender and malaria infection (Deressa et al., 2007). However a study by Clark et al. (2008) finds that the human host genetic makeup influences the rate of malaria incidence. They find that female children with glucose-6-phosphate dehydrogenase (G6PD) deficiency had lower risk of malaria infection. Haque et al. (2011) and Kateera et al. (2015) find that male children had a higher risk of malaria infection compared to their female counterparts due to behavioral differences.

### 1.3.2 Social-economic factors

Some of the socio-economic factors that have been studied as indicators of the socio-economic status of a household are; household income/wealth (Noor et al., 2006), type of housing construction, and ownership of household assets such as radio, bicycle, and mobile phone.

#### **Structure of the house**

The risk of malaria infection has been closely linked to the type of housing structure (Ayele et al., 2012; Chirebvu et al., 2014; Gamage-Mendis et al., 1991). The structure of the house consists of the type of wall material, floor material and roofing material used in their construction. The more traditional houses usually with large open eaves, thatched roofs without ceilings, earth floors and mud walls, provide conducive environments for the mosquitoes to rest and move easily thus increasing the risk of malaria (Snyman et al., 2015). Associated with the housing construction was the socio-economic status of the household. Usually poor households with low incomes lived in poorly constructed houses, thus had a greater risk of malaria, while the more wealthy households lived in better constructed houses (Ayele et al., 2013; Chirebvu et al., 2014).

#### **Use of preventive and control measures**

Some of the preventive and control measures for the prevention of malaria that have proved to be significant are: use of insecticide treated mosquito nets (ITNs) and indoor residual spraying (IRS). Mortality rates as a result of malaria infection decreased significantly by 7%, between 2000 and 2013, due to implementation of vector control measures, the use of diagnostic testing and ACTs (WHO, 2014).

The most widely used measure is the ITNs due to its availability and affordability. They are distributed freely to vulnerable groups particularly pregnant mothers and infants during ante-natal and post natal care in the WHO endemic regions. However, the distribution of

ITNs to the whole population at risk is usually hindered by lack of sufficient funding to support the program. Evidence from various studies shows a link in reduction of malaria mortality and morbidity in both adults and children due to the use of ITNs (Atieli et al., 2011; Ter Kuile et al., 2003; Nevill et al., 1996; Nyarango et al., 2006). Its use within households and amongst communities is affected by various factors such as the attitudes, gender, and education level of the household head (Atieli et al., 2011); the number of mosquito nets within households; and household income (Okrah et al., 2002).

The use of indoor residual spraying (IRS) has also been seen to reduce incidences of malaria (Nyarango et al., 2006; Shiff, 2002; Snow, 2015). The eradication of malaria in Europe and North America has been linked to the use of IRS particularly DDT (Carter and Mendis, 2002). In their study in Western Kenya Gimnig et al. (2016) find that the use of both ITN and IRS resulted in the reduction in the prevalence of malaria infection.

### **Maternal education**

One of the key determinants of proper care, treatment and control of tropical diseases is the human attitudes and behavior (Mwenesi et al., 1995). Mothers are usually the first caregivers to children because they spend more time with them and hence are able to detect any changes in their children. Various studies have linked maternal education to improvement in the health of household members particularly children (Caldwell and McDonald, 1982; Medrano et al., 2008; Siri, 2014). In a cross sectional study, in three countries (Angola, Tanzania and Uganda) Njau et al. (2014) confirms the analogy and finds that children with educated mothers were less likely to have malaria infections.

Educated caregivers also provided a protection effect on the household members against malaria infection due to knowledge on the malaria intervention methods such as case management, use of ITNs, vector control, child immunization and intermittent preventative treatment (ITP) for pregnant women (Keating et al., 2005; Noor et al., 2006; Siri, 2014).

The socio-economic status of a household plays an important role on the health status of the family members. Malaria has often been linked to poverty (Ayele et al., 2012) and occurs more often in endemic regions that are characterized by poverty. Household income, place of residence (either urban or rural setting), household structure and ownership of household assets such television, radios, mobile phone, bicycles often characterize the socio-economic status of families. Studies show that the lower the socio-economic status the greater the risk for malaria (Ayele et al., 2012; Deressa et al., 2007; Yadav et al., 2014). It influences the ability of the family to take up treatment due to costs, live in clean and hygienic environments, and acquire preventative paraphernalia such as ITNs.

### 1.3.3 Geographic and environmental factors

The geographic factors that have often been associated with malaria are the malaria endemic region and altitude. These regions are often located in low altitudes, and ecological factors such as temperature, rainfall and humidity play a role in determining the risk of malaria. Temperature, rainfall and humidity influence the mosquito's survival, the lifecycle of the parasite in the mosquito, and the breeding and feeding habits of the vector (Hunter, 2003; Gemperli, 2013). Many studies found the risk for malaria to decline with increasing altitude (Ayele et al., 2012; Peterson et al., 2009). In Kenya malaria infection is high in the western parts of the country and the coastal region, influenced by high rainfall, proximity to large water bodies (Lake Victoria and Indian Ocean), low altitudes and high temperature (Chaves et al., 2012). In the highland region of the country, at higher altitude with cooler temperatures, malaria has been associated with seasonal rainfall, vegetation cover, and distance from swampy environments (Ernst et al., 2006, 2009).

## 1.4 Objectives of the study

The main objective of this research is to identify the risk factors associated with malaria in children below the age of 14 years. This will be achieved through applying different statistical methods to our data. The data used in this study was collected through sampling survey that has the following characteristics:

- Sampling weights due to unequal probability of an observation being selected
- Cluster sampling whereby individuals or households are selected as groups making up the cluster.
- Stratification where homogeneous groups of clusters are sampled making up the strata.

This often results in correlation of observations between and within the clusters, non-response by some subjects yielding biased results and the use of unequal sampling weights. In order to account for the survey design and achieve our objective, the traditional logistic regression method, would not be used for analysis of our data. Instead, we shall use survey logistic regression method. The method is an example of the parametric generalized linear model for fitting non-normal data and includes the survey attributes in making inference about the parameter estimates.

Further, to cater for the subject specific random effects from the primary sampling unit, generalized linear mixed effects model (GLMM) method was used to fit the data. The GLMM is an extension of the GLM for modeling non-normal data by including both the fixed effects and random effects in the linear predictor. Both the SLR method and the GLMM are parametric methods that assume that the functional relationship between the response and the covariates are known a priori. In order to model the non-linear relationship between the response and effects of some of the covariates, a semi-parametric model, the

generalized additive mixed effects model (GAMM), was also used. Semi-parametric models are useful for modeling non-normal and non-linear data and account for over-dispersion and correlation by adding random effects to the additive predictor in the data.

## 1.5 Thesis outline

The first chapter of this thesis gives the introduction and the background to the study. In Chapter 2, a brief description of the data is provided through exploratory analysis of the data set, and the prevalence of malaria by various factors such as age and region is determined. Chapter 3 focuses on the class of generalized linear models (GLMs), particularly logistic regression methods for modeling data with a binary outcome. Attention is drawn to the survey logistic regression approach that is most applicable to our data set. A review of the generalized linear mixed effects models (GLMMs) is provided in Chapter 4. The data set is then fitted using GLMM to determine the factors associated with malaria incidence. Chapter 5, examines the use of semi-parametric methods in analysis of binary outcome data sets with non-normal and non-linear effects. Finally, Chapter 6 discusses and compares the results from the various statistical methods applied to the data. Conclusions are drawn from the discussions providing various limitations and possibilities for further research.

# Chapter 2

## Data presentation and description

### 2.1 Introduction

In this chapter, we describe the data set used in this study: The Kenya Malaria indicator survey (2010). Descriptive data analysis was performed to determine the relationships between the variables of interest. A chi-square test of independence analysis was equally performed to determine the association between the covariates and malaria status.

### 2.2 Background

This study uses data obtained from the Kenya malaria indicator survey (KMIS), that was carried out by the Government of Kenya, between June and August in 2010. This period is identified as the malaria transmission peak period.

One of the objectives of the study was to assess the prevalence of malaria in children between the ages of 3 months to 14 years. This would provide estimates on the prevalence of malaria on a nation wide scale for urban and rural areas in the five malaria endemic regions of Kenya

namely: Highland epidemic areas, lake endemic zone, coast endemic zone, seasonal risk areas and low risk areas.

The survey used a two stage stratified cluster sampling, adopted from the national sample survey and evaluation programme (NASSEP) IV sampling framework, that was developed by the Kenya National Bureau of Statistics (KNBS). The first stage sampling involved the selection of enumeration areas. A total of 1,800 clusters were created, with probability proportional to measure of size of the design frame and districts as strata. The second stage cluster sampling involved the selection of households within each cluster for the survey through simple random sampling.

From the sampling frame, 240 clusters were selected for the survey, and 30 households allocated to each cluster, making a total of 7,200 households to be used in the sample. From this sample, children under the age of 14 years in each household, were to be tested for malaria and women aged between 15 - 49 years were to be selected to participate in the individual survey.

Two types of questionnaires were used in the survey; a household questionnaire and an individual questionnaire. The household questionnaire captured information on the household membership, age, gender and relations, household dwellings and characteristics such as source of drinking water, type of toilet used, wall material for the house, roofing material and floor material; household possessions, such as, ownership of television, radio, mobile telephone, watches/clocks, bicycles e.t.c; net ownership and anti malarial spraying.

The individual questionnaire was given to consenting women, between the ages of 15 - 49 years of age, identified in the household questionnaire. The questionnaire captured details on their background characteristics such as age, religion, education, reproductive health in-

formation such as number of children, ante natal care, intermittent pregnancy treatment during pregnancy and attitude towards malaria treatment and child survival.

Malaria testing was performed on children between 3 months and 14 years whose parent/guardian provided a written consent during the household interview. Blood samples were taken from children through a finger blood prick. On spot malaria testing was carried out through rapid diagnostic malaria tests (RDT) using the CareStart® kit. Further, thick and thin blood smear samples were obtained for each tested child for microscopic analysis at the KEMRI/Walter Reed Project Malaria Diagnostic center Laboratory in Kisumu.

## 2.3 The Data sets

The KMIS 2010 was the second such survey ever conducted in Kenya after the first survey in 2007. The 2007 survey, only covered children under the age of 5 years. On the other hand, the KMIS 2010 survey assessed prevalence in children between 3 months and 15 years. It was also a national representative survey as it included low to no transmission endemic areas in its survey.

Out of the 7,200 households selected for the survey, 6,538 household heads were interviewed, and 11,310 children between 3 months and 14 years tested for malaria. A total of 5,749 women aged 15 - 49 years took part in the individual questionnaire. The main variable of interest in this survey is malaria status in children. A child is said to have malaria if their blood samples tested through RDT, had malaria parasites. Therefore, the response variable is binary, indicating whether a child has malaria or not.

The independent variables of interest, are categorized as either demographic, geographical or socio-economic variables and they are:

- Demographic variables including age and gender.
- Geographical variables include malaria endemic region and cluster altitude in metres.
- Socio-economic variables include: type of place of residence, wealth quantile, mothers highest education level, source of drinking water, type of toilet facility, wall material, floor material, roofing material, rooms used per person, mosquito nets used per person, mosquito nets used for sleeping and anti malarial spraying.

Malaria status and mother's highest education level were collected at an individual level. While age, household possession, type of place of residence, wealth quantile, source of drinking water, type of toilet facility, wall material, floor material, roofing material, rooms used per person, mosquito nets used per person, mosquito nets used for sleeping and anti malarial spraying were collected at household level.

## 2.4 Exploratory data analysis

Exploratory data analysis (EDA) is a fundamental step in data analysis and is used to determine the relationships and associations between the variables of interest in a data set. Cross tabulation was used to estimate associations between the response variable which is malaria status and the predictor variables.

Table 2.1 summarizes the descriptive statistics, showing malaria prevalence in children by the selected covariates.

Table 2.1: Prevalence of malaria by selected covariates

<b>Variable</b>	<b>Percent</b>	<b>N</b>
<b>Age</b>		
<1	8.1	530
1	7.4	876
2	10.3	957
3	9.9	935
4	11.7	934
5	13.1	890
6	14.4	849
7	15.7	773
8	15.0	765
9	14.0	649
10	13.4	826
11	13.9	590
12	13.6	618
13	11.9	603
14	12.2	515
<b>Gender</b>		
Male	12.7	5620
Female	11.8	5690
<b>Region</b>		
Highland endemic	3.0	2559
Lake endemic	41.7	2883
Moderate area	5.0	1732
Seasonal risk	0.5	2153
Low risk	0.6	1983
<b>Type of place of residence</b>		
Urban	4.0	1224
Rural	13.3	10086
<b>Wealth quantile</b>		
Richest	5.4	1807
Richer	11.6	2085
Middle	13.8	2572
Poorer	14.9	2349
Poorest	13.7	2467
<b>Mother's highest education</b>		

No education	7.8	1300
Primary Incomplete	15.6	2244
Primary complete	10.9	1562
Secondary Incomplete	13.1	451
Secondary complete	5.1	611
Higher	5.8	241
<hr/>		
<b>Toilet facility</b>		
Toilet with flush	1.5	468
Pit latrine	12.4	8682
No facility	14.5	2046
<hr/>		
<b>Source of Drinking water</b>		
Piped	5.7	2907
Borehole/well	16.8	3921
Springs/Rivers/lakes/dams	12.8	4366
<hr/>		
<b>Floor material</b>		
Earth/sand	5.7	5288
Dung	29.5	3105
Cement/tiles	5.8	2847
<hr/>		
<b>Roof Material</b>		
Thatch	15.2	2725
Sticks/mud	13.5	104
Wood/plastic	4.0	101
Corrugated iron	11.5	8208
Cement	4.3	94
<hr/>		
<b>Wall Material</b>		
Plastic/paper	20.8	24
Mud	17.1	6501
Wood/bamboo planks	1.8	1528
Cement	4.9	1644
<hr/>		
<b>Anti-malarial spraying</b>		
Yes	17.2	1654
No	11.4	9561
<hr/>		
<b>Use of Mosquito nets</b>		
Yes	13.6	7532
No	9.8	3918
<hr/>		

The data analyzed in the study consisted of 6,538 households from the 240 clusters, and the number of children eligible for malaria testing was 11,711 children. Out of these, only 11,310 children were tested for malaria. The overall prevalence of malaria in children was found to be 12.3%. From Figure 2.1, children aged 7 years had the highest prevalence rate for malaria at 15.7%. At lower ages between 3 months and 2 years, the prevalence rate is lower but seems to increase at each subsequent age group and then starts to decline for children in the 8 year to 14 year age groups.

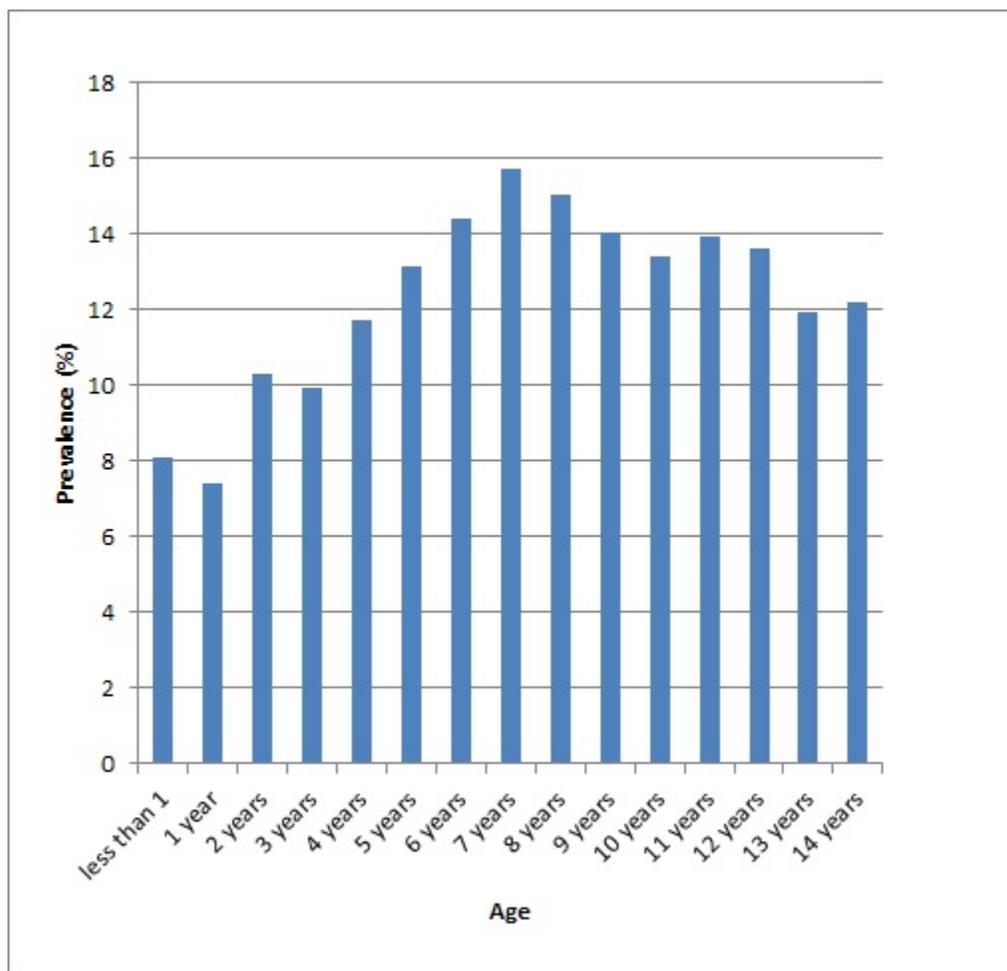


Figure 2.1: Prevalence of malaria by age

Figure 2.2 shows the prevalence of malaria in each of the different malaria endemic regions. It is evident that, there is a huge disparity in malaria prevalence across the endemicity regions. Out of all households living in the lake endemic region, 41.7% of them have children with malaria infection. In comparison, only 3% and 5% of households located in the highland and moderate endemic areas respectively, and less than one per cent of households within the seasonal risk and low risk regions, have children with malaria infection. Moreover, the prevalence of malaria in the rural area is more than thrice, that in the urban area.

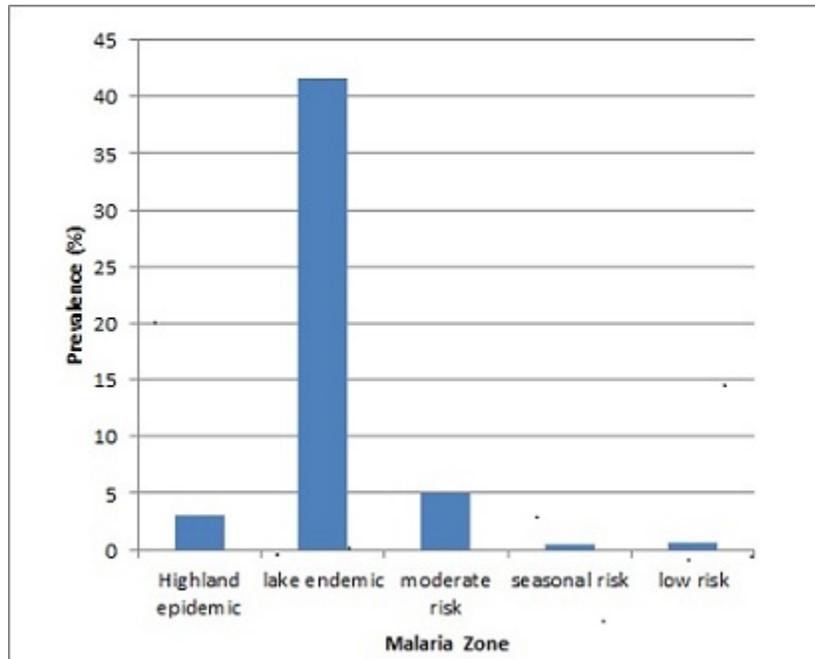


Figure 2.2: Prevalence of malaria by region

Figure 2.3 shows the prevalence of malaria by place of residence. It indicates that 13.3% of rural based households and 4% urban situated households had children with malaria. The prevalence of malaria by cluster altitude in metres is shown in Figure 2.4.

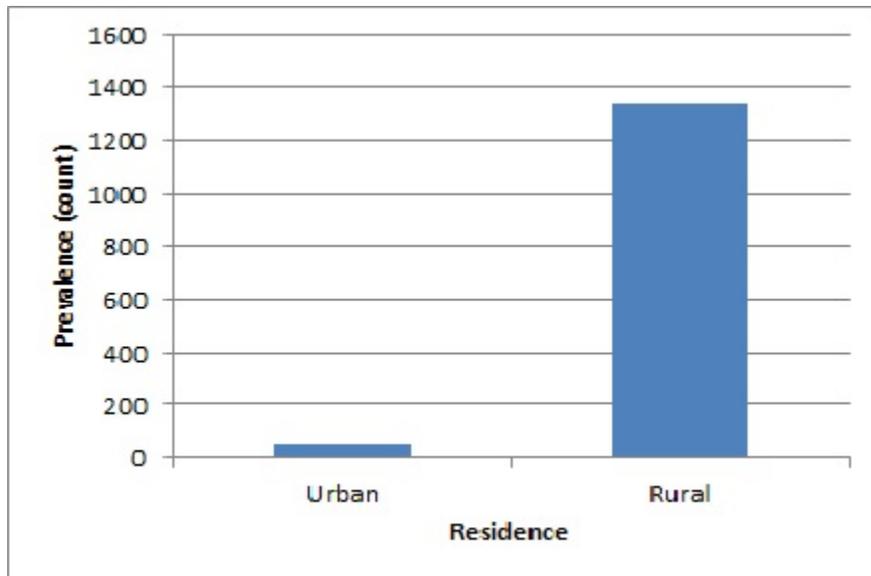


Figure 2.3: Prevalence of malaria by place of residence

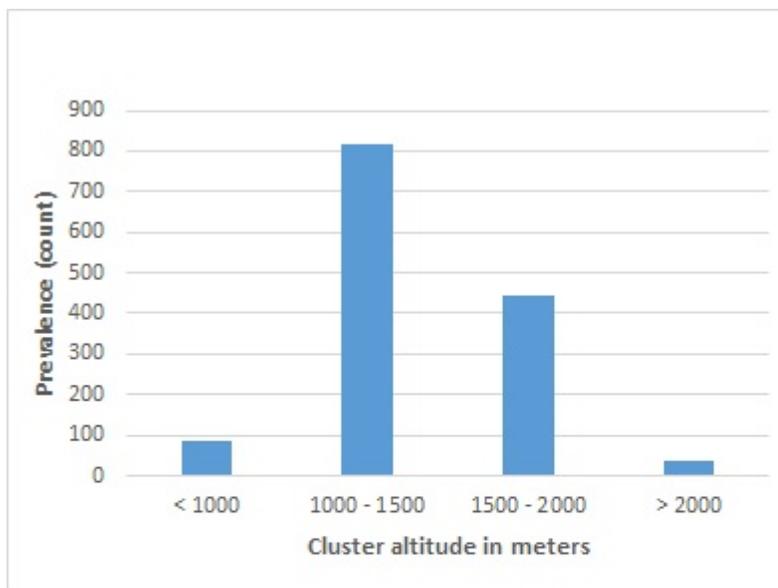


Figure 2.4: Prevalence of malaria by cluster altitude in metres

Malaria is highly prevalent in households located in cluster altitudes of between 1000 metres and 2000 metres in altitude. Of the households with children with positive malaria results, 59% and 32% were located in clusters within 1000 – 1500 metres and 1500 – 2000

metres in altitude respectively.

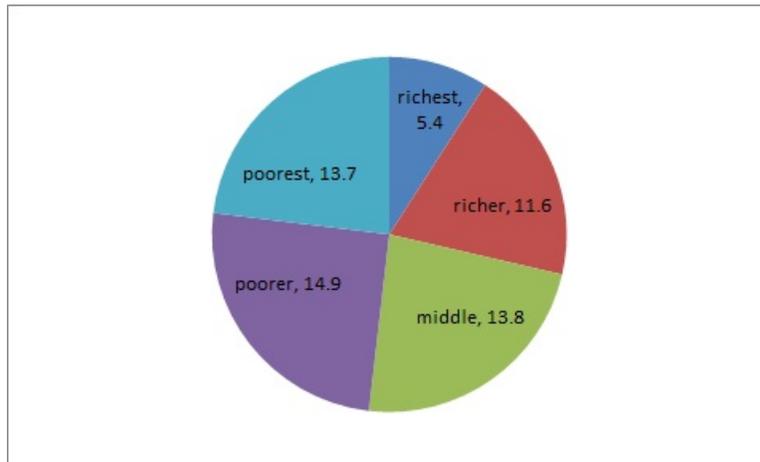


Figure 2.5: Prevalence of malaria by wealth quintile

From Figure 2.5, we see little difference in malaria prevalence among children in the lower wealth quintiles and it declines amongst children in the wealthier quintiles. The prevalence rate for the richest households, is 5.4%. This is much lower than the prevalence rates of households categorized in the middle, poorer and poorest category at 13.8%, 14.9% and 13.7% respectively.

Households that were characterized by either of the following: Use of toilets with flush system, had piped source for drinking water, and had well constructed homes with cemented walls, floors and roof, had lower prevalences of malaria. Figure 2.6 shows the distribution of malaria status by source of drinking water and type of toilet used. From the data, majority of the respondents, 38.2%, get their drinking water from springs/rivers/lakes/dams, while 76.3% of the households used pit latrines. The figure shows that, households who obtained their drinking water from springs/rivers/lakes/dams had higher malaria prevalence. Households with no toilet facility also had the highest malaria infection in their children.

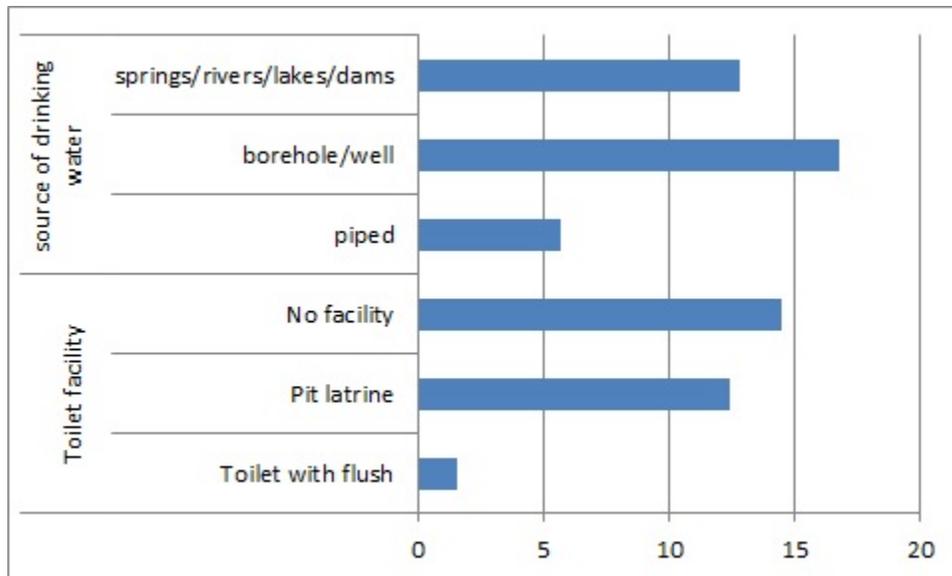


Figure 2.6: Distribution of malaria status by source of drinking water and type of toilet facility

Ownership of household assets such as television sets, radios, mobile phones, bicycles and availability of electricity are indicators of household wealth. Figure 2.7 gives the prevalence and distribution of malaria in children in households with these assets.

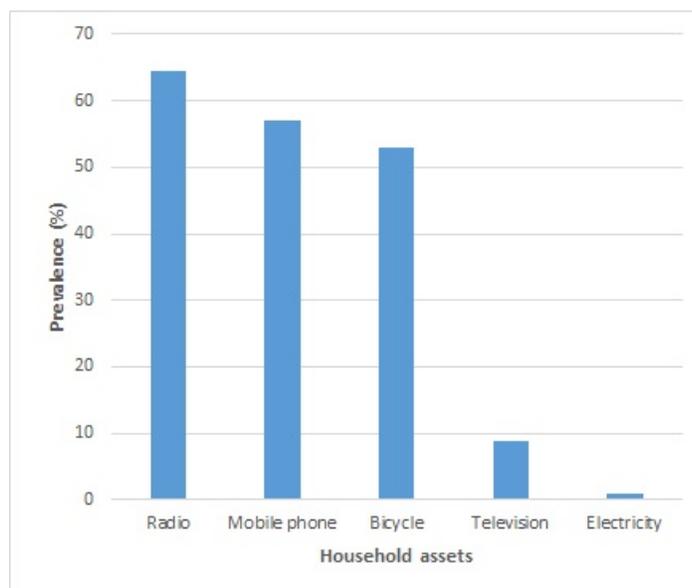


Figure 2.7: Distribution of malaria by ownership of household assets

The results of the survey show that 90% of the households do not have electricity, 84% do not own television sets whereas only 33%, 68% and 64% own bicycles, radios and mobile phones respectively. Interestingly, 64% of the children with positive malaria results hailed from households with radios. A similar result can be deduced for households with mobile phones and bicycles, reporting 57% and 53% respectively of positive malaria cases.

The distribution of malaria status by wall, floor and roofing materials of the households is presented in Figure 2.8. Households whose floors were made of earth/sand and cement/tiles had low percentages of children with malaria, 5.7% and 5.8% respectively. The higher percentage of malaria infection was in households with dung floors, at 29.5%. Households with dwellings made of plastic/paper and mud walls had higher prevalences of malaria, at 2.8% and 17.1% respectively. Dwellings with wood/bamboo planked walls and cemented walls had lower percentages of infected children. Finally, malaria infection within households with thatched roofing was 15.2%, 13.5% with sticks/mud roofing, 4% with wood/plastic roofing, 11.5% with corrugated roofs and 4.3% with cemented roofs.

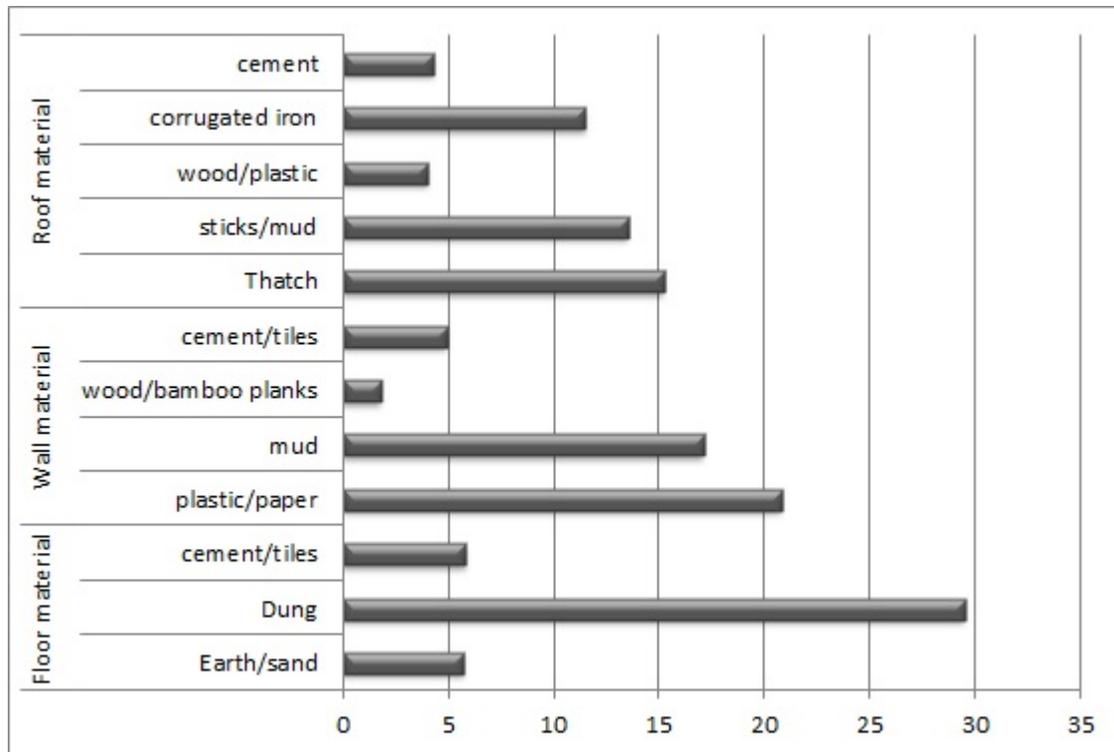


Figure 2.8: Distribution of malaria status by wall material, floor material and roof material used in household construction

The effect of mother's education on malaria prevalence is illustrated in Figure 2.9. Mothers with completed secondary education and higher education, had lower malaria infections in their children, with prevalence rates of 5.1% and 5.8% respectively. Education therefore seemed to have a positive and protective effect since educated mothers could enforce preventive measures to control malaria infection.

Most households in the survey, 85.3%, do not spray their houses to control the malaria vector. The prevalence rate for malaria for households that used anti-malarial spraying was 17.2% and 11.4% for those who did not. On the contrary, most households, 65.2%, used mosquito nets while sleeping. In total, 13.6% of those who used mosquito nets had children with malaria while 9.8% of those who did not use mosquito nets had children with malaria.

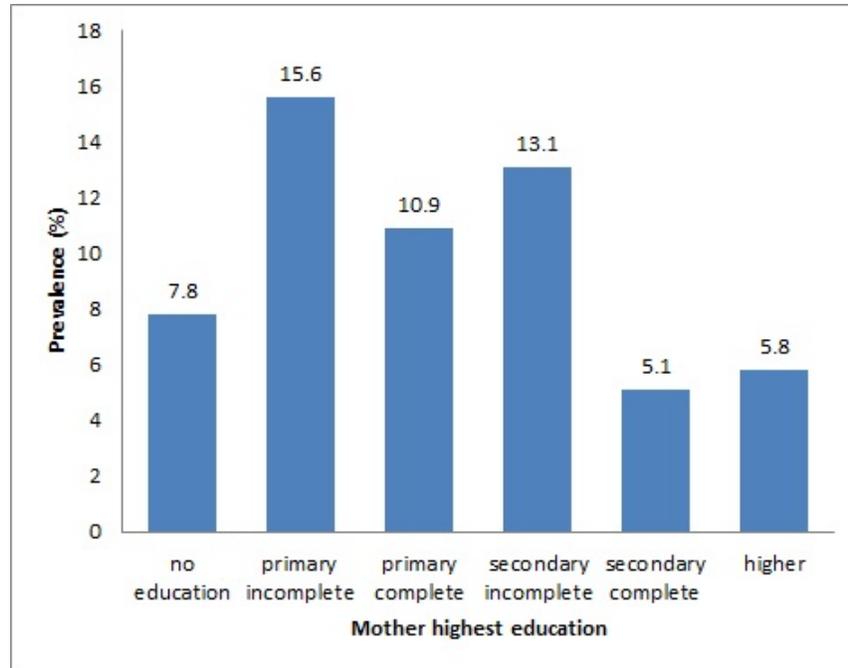


Figure 2.9: Prevalence of malaria by mother's highest education level

### 2.4.1 Tests of association

A  $\chi^2$  test of independence was performed on the various covariates and the results are presented on Table 2.2. From the analysis, there was a significant association between malaria status in children and all the selected covariates except the sex of the child. There is a positive association between malaria and age ( $\chi^2$ : 59.144, p-value: <0.0001). Malaria status is also significantly associated with the endemicity region, within which a household is located ( $\chi^2$ : 3139.598, p-value: <0.0001). The type of place of residence is also similarly associated with malaria infections in children ( $\chi^2$ : 87.039, p-value: <0.0001).

Malaria infection in children is also strongly related with the wealth quintile of each household, ( $\chi^2$ : 104.811, p-value: <0.0001). Other variables, also found to be positively associated with malaria are: wall material, roof material and floor material used in a household's

dwelling construction; source of drinking water and type of toilet facility. Malaria intervention covariates, including: anti-malaria spraying and use of mosquito nets, were also associated with malaria status. However, there was no association between malaria status and the variable, sex of the child ( $\chi^2$ : 1.861, p-value: <0.0001). Therefore, this variable may not be used in the final model.

Table 2.2: Cross-tabulation of malaria status with selected covariates

Variable	$\chi^2$ Statistic	df	P-value
Age	59.144	14	<0.0001
Sex	1.861	1	0.173
Region	3139.598	4	<0.0001
Type of residence	87.039	1	<0.0001
Mothers highest education	90.753	5	<0.0001
Wealth quantile	104.811	4	<0.0001
Wall material	393.260	3	<0.0001
Roof material	38.488	4	<0.0001
Floor material	1179.188	2	<0.0001
Anti-malarial spraying	45.126	1	<0.0001
Source of drinking water	190.288	2	<0.0001
Toilet facility	59.843	2	<0.0001
Use of mosquito nets	34.052	1	<0.0001
Mosquito nets/person	242.088	55	<0.0001
Rooms/person	233.266	62	<0.0001

### 2.4.2 Summary

From the analyses, age of child in years, malaria zone, type of place of residence, and cluster altitude in metres are the important demographic and geographical factors associated with malaria infection. Malaria seemed to increase with age of child but declining after age 8. Households located in rural areas, high endemicity regions and located in low altitude areas had higher malaria prevalences. The socio-economic status of a household played an important role in determining susceptibility to malaria infections. Some of the socio-economic variables that contributed to higher risks of malaria were: wealth quantile; type of housing structure that includes the material used in roof, wall, and floor construction; toilet facility; source of drinking water; ownership of household assets such as bicycles, radio, television and mobile phones; and rooms per person sharing. Use of insecticide treated mosquito nets, number of nets, and antimalarial spraying were important factors for preventing and controlling malaria. An important factor for consideration is the level of mother's education, which seemed to provide a protective cover to children against malaria risk. Households with mother's with primary and higher education had lower prevalence rates.

To refine the findings of this analysis, three different statistical methods are used in the next chapters, to determine the important risk factors for malaria infection in children.

# Chapter 3

## The Generalized linear model

### 3.1 Introduction

The Generalized linear model (GLM) developed by Nelder and Wedderburn (1972) is a statistical technique used in the analysis of data whose outcome may not usually be normally distributed. The response variable belonging to the exponential family of distributions, can be modeled by relating the linear predictor of the predictor covariates to the response variable via a function of the mean response called a link function. GLMs therefore is a generalization of the general linear model that includes regression analysis, analysis of the variance and analysis of the covariance, that explain the variation in the response variable as a linear combination of the explanatory terms and the residual errors based on the normality assumption to more inclusive non-normal distribution. Some examples of models that belong to the class of GLMs include: logistic regression for binary response data, Poisson regression models for count data, multiple regression models for normal response, log-linear categorical data analysis models and exponential models for survival data analysis (Nelder and Wedderburn, 1972).

### 3.1.1 The model structure

A general linear model can be written as:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \epsilon_i$$

where it is assumed a response  $Y$  is independently observed on  $N$  units along with  $p$  explanatory variables denoted by  $Y_i, X_{1i}, \dots, X_{pi}$  for  $i = 1, 2, \dots, N$ . The error term  $\epsilon_i$  are usually assumed to be iid  $N(0, \sigma^2)$ . However this is a strict assumption to be satisfied in reality because the errors can be correlated.

In matrix form, the model is expressed as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where  $\mathbf{Y}$  is a vector of the response variables,  $\mathbf{X}$  is an  $n \times (p + 1)$  design matrix of the independent variables,  $\boldsymbol{\beta}$  is a vector of the  $(p + 1)$  regression parameters including the intercept and  $\boldsymbol{\epsilon}$  is a vector of the error terms. The assumptions of the model based on Kutner et al. (2005):

- The expectation of the error terms  $E(\boldsymbol{\epsilon}) = 0$ , therefore the mean of the response variable  $Y_i$  is  $E(Y) = E(\mathbf{X}\boldsymbol{\beta}) + E(\boldsymbol{\epsilon}) = \mathbf{X}\boldsymbol{\beta}$
- The error terms have a constant variance  $\sigma^2$ . Therefore the response variable  $\mathbf{Y}$  also has a constant variance.
- It is assumed that the error terms say  $\epsilon_i$  and  $\epsilon_j, i \neq j$  are uncorrelated, therefore the response variables  $Y_i$  and  $Y_j$  are also uncorrelated.

These assumptions are somehow strict for data whose distribution is not Gaussian. Therefore, the GLM can be used to model such data whose distributions are from the exponential family of distributions.

### 3.1.2 Exponential family

The exponential family comprises of a set of distributions from discrete, continuous or a mix of both discrete and continuous random variables, and includes distributions such as: Normal, Binomial, Bernoulli, Poisson, Gamma, Multinomial and Weibull distributions. The natural form of the exponential family as defined by McCullagh and Nelder (1989) can be written as:

$$f(y_i|\theta_i; \phi) = \exp \left[ \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right] \quad (3.1)$$

where  $\theta_i$  is known as the canonical or natural parameter,  $a(\phi)$  is the dispersion parameter,  $b(\cdot)$  and  $c(\cdot)$  are some known functions while  $\theta$  and  $\phi$  are unknown parameters.

The function  $a(\phi)$ , the exponential dispersion function (Jorgensen, 1987), and has the form  $a(\phi) = \phi/w_i$  where  $w_i$  is an observation specific weight (Agresti, 1990).

Let  $Y$  be the response variable with  $y_i, i = 1, 2, \dots, n$  independent observations. Using the property  $\int f(y|\theta, \phi)dy = 1$ , the mean and the variance of  $Y$  can be derived through taking first and second derivatives of the function, with respect to  $\theta$ .

$$\int (y - b'(\theta)) f(y)dy = 0$$

$$\int [a(\phi)^{-1}(y - b'(\theta))^2 - b''(\theta)]f(y)dy = 0$$

Therefore, the mean  $E(Y) = b'(\theta)$  and the variance  $Var(Y) = a(\phi)b''(\theta) = a(\phi)v(\mu)$  can be obtained through the following procedure described in Dobson (1990):

Let  $U = \frac{dl(\theta; y)}{d\theta}$  be the score function obtained from taking the derivative of the log-likelihood function  $l(\theta; y)$  of the probability distribution function of  $Y$ ,  $f(y; \theta)$  with respect to  $\theta$ . To find the first and the second moments of  $U$ , we shall make use of the identity,

$$\frac{d \log f(y; \theta)}{d\theta} = \frac{1}{f(y; \theta)} \frac{df(y; \theta)}{d\theta} \quad (3.2)$$

Taking the expectations on both sides of the equation yields

$$\begin{aligned} E(U) &= \int \frac{d \log f(y; \theta)}{d\theta} f(y; \theta) dy \\ &= \int \frac{df(y; \theta)}{d\theta} dy \\ &= \frac{d}{d\theta} \int f(y; \theta) dy = \frac{d}{d\theta} = 0 \end{aligned}$$

Under certain regularity conditions. Since  $\int f(y; \theta) dy = 1$ , it follows that  $E(U) = 0$ , hence it follows  $E(Y) = b'(\theta)$ .

Differentiating equation 3.2 again with respect to  $\theta$  gives:

$$\begin{aligned} \frac{d}{d\theta} \int \frac{d \log f(y; \theta)}{d\theta} f(y; \theta) dy &= \frac{d^2}{d\theta^2} \int f(y; \theta) dy \\ &= 0 \end{aligned}$$

Therefore the left hand side of this equation can be written as:

$$\frac{d^2 \log f(y; \theta)}{d\theta^2} f(y; \theta) dy + \int \frac{d \log f(y; \theta)}{d\theta} \frac{df(y; \theta)}{d\theta} dy$$

substituting equation 3.2 in the second term yields

$$\frac{d^2 \log f(y; \theta)}{d\theta^2} f(y; \theta) dy + \int \left[ \frac{d \log f(y; \theta)}{d\theta} \right]^2 f(y; \theta) dy = 0$$

Thus

$$E \left[ -\frac{d^2 \log f(y; \theta)}{d\theta^2} \right] = E \left[ \left( \frac{d \log f(y; \theta)}{d\theta} \right)^2 \right]$$

The variance of the score function is  $Var(U) = E(U^2) = E(U')$ , where  $U'$  is the derivative of the score with respect to  $\theta$ . Therefore the variance of  $Y$  becomes  $Var(Y) = a(\phi)b''(\theta)$ .

The GLM has three main components (Agresti, 1990), they are:

### ***The random Component***

A GLM consists of a response variable,  $Y$  from the exponential family of distributions with  $N$  independent observations of the form  $Y_1, \dots, Y_N$ . The first and second moments of  $Y_i$  gives the mean,  $\mu_i = E(Y_i) = b'(\theta)$  and variance,  $Var(Y_i) = a(\phi)b''(\theta)$  respectively.

### ***The Systematic component***

This component relates a vector  $\eta = (\eta_1, \dots, \eta_N)'$  to a set of explanatory variables through a link function. Let  $X_i = [1, x_{1i}, \dots, x_{pi}]$  be a  $p$ -dimensional vector of covariates and  $\beta = (\beta_0, \dots, \beta_p)$ , be a vector of the regression coefficients. The distribution of  $Y_i$  depends on  $X_i$ , through the linear predictor,  $\eta_i$ , such that:

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi}$$

### ***The Link Function***

The link function, given by  $g(\mu_i)$ , is a monotonic and differentiable function that describes how the mean  $E(Y_i) = \mu_i$ , depends on the linear predictor. Thus the GLM is generally defined as:

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi}$$

The inverse of the link function,  $g^{-1}(\eta) = \mu$  is referred to as the mean function. Special types of link functions are those obtained directly from the natural parameter  $\theta$  of the exponential family. Such link functions are called the canonical link functions such as the logit link for

binary data, log link for count data and identity link for normal data.

### 3.1.3 Maximum likelihood estimation

The regression parameters in a GLM are estimated using maximum likelihood estimation (MLE) method (Nelder and Wedderburn, 1972). The estimates are the values of the parameters that maximize the log-likelihood function (Olsson, 2002). The likelihood function for  $N$  independent observations of the parameter  $Y$  with a p.d.f from the exponential family is :

$$L(\theta, \phi; Y) = \prod_{i=1}^N f(Y_i | \theta_i, \phi)$$

The log-likelihood equation is:

$$\begin{aligned} l(\beta, Y) &= \sum_{i=1}^N \ln \left[ \exp \left[ \frac{Y_i \theta_i - b(\theta_i)}{a(\phi)} + c(Y_i, \phi) \right] \right] \\ &= \sum_{i=1}^N a(\phi)^{-1} (Y_i \theta_i - b(\theta_i)) + \sum_{i=1}^N c(Y_i, \phi) \end{aligned} \quad (3.3)$$

The log-likelihood equation for a single observation  $i$  is given by:

$$l_i(\beta) = a(\phi)^{-1} (Y_i \theta_i - b(\theta_i)) + c(Y_i, \phi)$$

We partially differentiate the log likelihood equation for observation  $i$ , with respect to the regression coefficients  $\beta_j$ , using chain rule;

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

Since  $\frac{\partial l_i}{\partial \theta_i} = [Y_i - b'(\theta_i)]/a(\phi)$  and  $E(Y_i) = \mu_i = b'(\theta)$  and  $\text{var}(Y_i) = a(\phi)b''(\theta_i)$ , then,

$$\frac{\partial l_i}{\partial \theta_i} = \frac{(Y_i - \mu_i)}{a(\phi)}$$

and

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \frac{\text{var}(Y_i)}{a(\phi)}$$

Again, since  $\eta_i = \sum_j \beta_j x_{ij}$ , then,  $\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$

Substituting these in the score equation, for N independent observations, gives:

$$\begin{aligned} \frac{\partial l(\beta)}{\partial \beta_j} &= \sum_{i=1}^N \frac{Y_i - \mu_i}{a(\phi)} \frac{a(\phi)}{\text{var}(Y_i)} x_{ij} \frac{\partial \mu_i}{\partial \eta_i} \\ &= \sum_{i=1}^N \frac{Y_i - \mu_i}{\text{var}(Y_i)} x_{ij} \frac{\partial \mu_i}{\partial \eta_i} \end{aligned} \quad (3.4)$$

The estimating function can also be used to determine the asymptotic covariance matrix of  $\hat{\beta}$ , the inverse of the Information matrix,  $I(\beta)$  (Agresti, 2002).

$$\begin{aligned} I(\beta) &= -E \left( \frac{\partial^2 l}{\partial \beta_i \partial \beta_j} \right) = E \left[ \left( \frac{\partial l}{\partial \beta_i} \right) \left( \frac{\partial l}{\partial \beta_j} \right) \right] \\ &= E \left[ \frac{Y_i - \mu_i}{\text{var}(Y_i)} x_{ih} \frac{\partial \mu_i}{\partial \eta_i} \frac{Y_i - \mu_i}{\text{var}(Y_i)} x_{ij} \frac{\partial \mu_i}{\partial \eta_i} \right] \\ &= \sum_{i=1}^N \frac{x_{ih} x_{ij}}{\text{var}(Y_i)} \left( \frac{\partial \mu_i}{\partial \eta_i} \right)^2 \end{aligned} \quad (3.5)$$

Let  $W$  be the diagonal matrix with main diagonal elements  $w_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 / \text{var}(Y_i)$ . Then the Fisher information is given as:

$$I(\beta) = (X'WX)$$

The asymptotic covariance matrix becomes:

$$\text{cov}(\hat{\beta}) = (X'WX)^{-1}$$

The score equation then reduces to:

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^n (y_i - \mu_i) w_i \frac{\partial \eta_i}{\partial \mu_i} x_{ij} \quad (3.6)$$

By equating the score equations to zero, the ML estimates,  $\hat{\beta}$  can be obtained using any of these three methods: Iterative re-weighted least squares, Newton Raphson and Fisher scoring (Agresti, 2002).

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. The Taylor series expansion is generally given by:

$$f(x_o) + (x_1 - x_o)f'(x_o) + \frac{(x_1 - x_o)^2}{2!}f''(x_o) + \frac{(x_1 - x_o)^3}{3!}f'''(x_o) + \dots = 0$$

Using the first and second terms, assuming higher order terms are negligible, we have :

$$f(x_o) + (x_1 - x_o)f'(x_o) = 0$$

which yields the equation,

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

that is the basis for the iterative updating equation in the Newton Raphson estimation algorithm.

The Newton Raphson method adopts the above by using the score of the log-likelihood as the basis for parameter estimation:

$$\beta_r = \beta_{r-1} - \left( \frac{\partial l(\beta_{r-1})}{\partial \beta} \right) \left( \frac{\partial^2 l(\beta_{r-1})}{\partial^2 \beta} \right)^{-1}$$

implying,

$$\beta_r = \beta_{r-1} - S(\beta_{r-1})[S'(\beta_{r-1})]^{-1} \quad (3.7)$$

Where  $S'(\beta_{r-1})$  is the partial derivative of the score equation with respect to  $\beta$ , evaluated at  $\beta_{r-1}$  and is referred to as the Hessian matrix .

The fisher scoring is an alternative method for solving the log-likelihood estimating equations. It resembles the Newton Raphson method but the difference being in the use of the expected value of the Hessian matrix based on the information matrix. By some complicated procedures, it can be shown that

$$I(\beta) = E \left( \frac{\partial^2 l}{\partial \beta_j \partial \beta_k} \right) = -E \left( \frac{\partial l}{\partial \beta_j} \right) \left( \frac{\partial l}{\partial \beta_k} \right)$$

Therefore updating under Fisher scoring is:

$$\beta_r = \beta_{r-1} + S(\beta_{r-1})[I(\beta_{r-1})]^{-1} \quad (3.8)$$

The iterative re-weighted least squares method makes use of the fisher scoring method to find the ML estimates. Multiplying both sides of equation 3.8 by  $I(\beta_{r-1})^{-1}$  we get,

$$\beta_r [I(\beta_{r-1})]^{-1} = (\beta_{r-1}) [I(\beta_{r-1})]^{-1} + S(\beta_{r-1})$$

Expressing the fisher information and score equation in terms of its covariates at the  $(r-1)^{th}$  iterate gives:

$$\left( \frac{x_{ih}x_{ij}}{\text{var}(Y_i)} \left( \frac{\partial \mu_i}{\partial \eta_i} \right)^2 \right) \beta_r = \left( \frac{x_{ih}x_{ij}}{\text{var}(Y_i)} \left( \frac{\partial \mu_i}{\partial \eta_i} \right)^2 \right) \beta_{r-1} + \left( \frac{(Y_i - \mu_i)}{\text{var}(Y_i)} x_{ij} \frac{\partial \mu_i}{\partial \eta_i} \right)$$

Let  $\eta_{ih} = x_{ih}\beta_{r-1}$  and  $\text{var}(Y_i) = \left( \frac{\partial \mu_i}{\partial \eta_i} \right)^2 w^{-1}$ . The fisher scoring equation takes the form:

$$[X'WX]\beta_r = X'W\eta_i + x_{ij}W(Y_i - \mu_i) \frac{\partial \mu_i}{\partial \eta_i}$$

Let  $Z_i = \eta_i + (Y_i - \mu_i) \frac{\partial \mu_i}{\partial \eta_i}$ , our equation reduces to

$$[X'WX]\beta_r = X'WZ$$

Therefore

$$\beta_r = [X'WX]^{-1}[X'WZ] \tag{3.9}$$

To obtain the new estimate of  $\beta_r$ , the working dependent variable  $z_{r-1}$  is regressed on X, with weight  $W_{r-1}$ . A new linear predictor is obtained,  $\eta_r = X\beta_r$  and a new working dependent variable  $z_r$  for the next  $(r+1)$   $\hat{\beta}$  estimate. The ML estimator is the limit of  $\beta_r$  as  $r \rightarrow \infty$ .

### 3.1.4 Assessing the fit of a model

Given a data set, a statistical model is of good fit, if it fits the set of observations well. A good model should minimize the discrepancy between the expected values under the model and the observed values. The two statistical methods that are used in assessing this fit are: the deviance and the Pearson's chi square statistic.

The **deviance** measures the discrepancy of fit between the maximum log-likelihood of the saturated model and the log-likelihood of the fitted model, and hence we can define the deviance as:

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

where,  $l(y, \phi; y)$  is the log-likelihood function of the saturated model and  $l(\hat{\mu}, \phi; y)$  is the log-likelihood of the observed model.  $\hat{\mu}$  is the maximum likelihood estimator of the model of interest, and  $\phi = 1$ . The scaled deviance, in the case  $\phi \neq 1$ , is defined as:

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

The **Pearson's goodness of fit statistic** is a score statistic for testing the fitted model against the saturated model defined by Smyth (2003) as:

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

where;  $y_i$  are the response variables,  $w_i$ , the weights,  $\hat{\mu}_i$ , the fitted means evaluated at the MLE  $\hat{\beta}$  and  $v(\hat{\mu})$  is the variance function. The Pearson's score statistic is approximately  $\chi^2$  distributed with the residual degrees of freedom for the fitted model.

### 3.1.5 Model selection

Model selection is an important process in statistical analysis and involves the selection of the best model amongst several competing models. The two main criteria for model selection used in GLMs are; Akaike information criterion (AIC) (Akaike, 1974) and the Bayesian information criterion (BIC) (Schwarz, 1978). The AIC is defined as:

$$AIC = -2l(\beta) + 2p \quad (3.10)$$

where  $l(\beta)$  is the maximum log-likelihood and  $p$  is the number of parameters for the model. The AIC is important in model comparisons, and a model with a smaller AIC is preferred (Lindsey, 1997). The BIC method looks at the asymptotic behavior of the Bayes estimators and takes into consideration the sample size (Schwarz, 1978). The best model is one that minimizes the following equation, given the log-likelihood function  $l(\beta)$ , with  $k$  parameters and a sample size of  $n$ ;

$$BIC = -2l(\beta) + k\log(n) \quad (3.11)$$

## 3.2 Logistic regression model

### 3.2.1 Introduction

Logistic regression is a statistical method for analyzing a data set, where the dependent variable is dichotomous or binary and the independent variables may be categorical or a mix of continuous and categorical variables (Peng et al., 2002). It employs the maximum likelihood method to get the best fit equation and assumes that the relationship between the dependent and independent variables is not necessarily linear and that the residuals are not necessarily normally distributed. It is a special case of the GLM Nelder and Wedderburn

(1972) that is applied in various studies including health science, epidemiology, demographic studies and education (Park, 2013; Peng et al., 2002; Hosmer and Lemeshow, 1989). In this study, we shall limit ourselves to the use of binary logistic regression, where our response variable is strictly dichotomous stating whether a child has malaria or not.

### 3.2.2 Binary logistic regression

Suppose we have a binary outcome variable denoted by  $Y$ , representing the presence or absence of an event such that  $Y = 1$  if the event occurs and  $Y = 0$  if the event does not occur and a set of  $p$  independent variables denoted by the vector  $\mathbf{x}' = (x_1, x_2, \dots, x_p)$ . It is assumed that  $Y$  has a Bernoulli distribution represented as:

$$Y = \begin{cases} 1, & \text{if the event occurs (with probability } \pi) \\ 0, & \text{If the event does not occur (with probability } 1 - \pi) \end{cases} \quad (3.12)$$

Here,  $\pi(x)$  denotes the conditional probability of the event occurring given the independent variables,  $\pi(x) = P(Y = 1|x_1, x_2, \dots, x_p)$ . Therefore,  $1 - \pi(x) = P(Y = 0|x_1, x_2, \dots, x_p)$ . The logistic regression model is defined as:

$$\ln \left( \frac{\pi(x)}{1 - \pi(x)} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots + \beta_p x_p \quad (3.13)$$

Solving for  $\pi(x)$  gives

$$\pi(x) = \frac{e^{x'\beta}}{1 + e^{x'\beta}}$$

The link is the logit link thus, in principle, the model is the log odds as a function of a set of predictions  $x_1, x_2, \dots, x_p$ .

### 3.2.3 Fitting the logistic regression model

Fitting a logistic regression model involves the estimation of the unknown parameters through the maximum likelihood estimation method. The likelihood function for a sample of observations of the pair  $(x_i, y_i)$  for  $i = 1, 2, \dots, n$ , with  $\pi_i$  probabilities is

$$L(\beta) = \prod_{i=1}^n P(y_i | x_1, \dots, x_p) = \prod_{i=1}^n \left( \frac{e^{x'_i \beta}}{1 + e^{x'_i \beta}} \right)^{y_i} \left( \frac{1}{1 + e^{x'_i \beta}} \right)^{1-y_i} = \prod_{i=1}^n \pi_i^{y_i} [1 - \pi_i]^{1-y_i} \quad (3.14)$$

The maximum likelihood estimates for the regression parameters denoted by the vector  $\beta = (\beta_0, \beta_1, \dots, \beta_p)'$ , are the values that maximize the log-likelihood function. Expressing equation 3.14 in terms of its log yields

$$\begin{aligned} l(\beta) &= \sum_{i=1}^n y_i \log[\pi_i] + (1 - y_i) \log[1 - \pi_i] \\ &= \sum_{i=1}^n \log(1 - \pi_i) + \sum_{i=1}^n y_i \log \left( \frac{\pi_i}{1 - \pi_i} \right) \\ &= \sum_{i=1}^n \log \left( \frac{1}{1 + e^{x'_i \beta}} \right) + \sum_{i=1}^n y_i (x'_i \beta) \\ &= \sum_{i=1}^n y_i (x'_i \beta) - \sum_{i=1}^n \log(1 + e^{x'_i \beta}) \end{aligned} \quad (3.15)$$

Partially differentiating the log-likelihood function with respect to the regression parameters and equating the result to 0, yields the likelihood equations. The likelihood equations obtained can be solved iteratively to obtain the maximum likelihood estimates for  $\beta$ . These likelihood equations are defined as:

$$\sum_{i=1}^n [y_i - \pi_i] = 0 \quad (3.16)$$

and

$$\sum_{i=1}^n x_{ij}[y_i - \pi_i] = 0 \quad (3.17)$$

for  $j = 1, 2, \dots, p$

### 3.2.4 Model selection and model fit

The methods commonly used for selecting variables in logistic regression are forward selection, backward elimination and stepwise regression (Chen and Dipak, 2003). Backward elimination method starts with a model with all variables included, and makes use of the results of wald tests for each parameter. A parameter with the least significant effect is eliminated and cannot be returned to the model. Forward selection method, on the other hand starts with an empty or null model. It examines the score chi-square statistic for each parameter not in the model, and if significant, it is added to the model and cannot be eliminated (Bursac et al., 2008). Stepwise regression varies from forward selection method in that variables added to the model, may be eliminated. The BIC and AIC methods discussed in section 3.1.5 may be used for selecting a suitable model that best describes the data.

The goodness of fit for a logistic regression model as defined by Hosmer and Lemeshow (1989) assesses the effectiveness of the model in describing the outcome variable. The fitted model's residual variation is expected to be small, displaying no systematic tendency and follows the model's variability (Hosmer et al., 1997). It can be measured using the Hosmer - Lemeshow (H - L) tests, Pearson chi-square statistic and the deviance statistic.

The H - L statistic is obtained through the calculation of the Pearson chi-square statistic from a  $2 \times g$  table of observed and estimated expected frequencies. Here  $g$  represents the numbers of groups obtained from the estimated probabilities. The H - L statistic is denoted

$\hat{C}$  given as

$$\hat{C} = \sum_{r=1}^g \frac{(o_r - n_r \pi_r)^2}{n_r \pi_r (1 - \pi_r)} \sim \chi_{g-2}^2 \quad (3.18)$$

where  $n_r$  is the total frequency of subjects in the  $r^{th}$  group and  $o_r$  is the total frequency of event outcome in the  $r^{th}$  group while  $\pi_r$  is the average estimated probability of an event outcome in the  $r^{th}$  group. If the logistic regression model is the correct model, the statistic  $\hat{C}$  can be approximated by the chi-square distribution with  $g-2$  degrees of freedom.

The Pearson's chi-square statistic,  $X^2$  and the deviance,  $D$  compares the observed values of the logistic regression model to the predicted values in a  $2 \times n$  table (Hosmer et al., 1997). In a  $2 \times n$  table, 2 rows define the values of the binary dependent variable  $y$  and  $n$  columns defines the number of values the  $p$  covariate variables might take in the model.

$$D = -2 \sum_{i=1}^n y_i \log\left(\frac{y_i}{\pi_i}\right) - 2 \sum_{i=1}^n (1 - y_i) \log\left[\frac{(1 - y_i)}{(1 - \pi_i)}\right] \quad (3.19)$$

and

$$X^2 = \sum_{i=1}^n \frac{(y_i - \pi_i)^2}{\pi_i (1 - \pi_i)} \quad (3.20)$$

### 3.2.5 Odds ratio

The parameter estimates from a logistic regression analysis are reported in terms of odds ratios. To define the odds ratios it is important to understand the concept of odds. Given that  $\pi(x)$  is the probability of an event occurring as a function of the covariate  $x$ , it also follows that  $1 - \pi(x)$  is the probability of an event not occurring. The odds of an event occurring therefore is the ratio of the probability of the event occurring to the probability

of the event not occurring, defined as,

$$O(x) = \frac{\pi(x)}{1 - \pi(x)}$$

It is a measure of association, that estimates the relationship between the risk factor and the outcome while adjusting for other variable (Wilber and Fu, 2010). They are used to compare the odds that the outcome of interest will occur given the exposure, to the odds that the outcome will not occur given no exposure to the variable of interest (Agresti, 1990). Suppose  $x$ , is a categorical independent variable coded 0 and 1. The odds ratio is defined as the ratio of the odds for  $x = 1$  to the odds for  $x = 0$  (Hosmer and Lemeshow, 1989);

$$OR = \frac{\pi(1)/1 - \pi(1)}{\pi(0)/1 - \pi(0)} \quad (3.21)$$

An odds ratio equivalent to 1 i.e.  $OR = 1$  implies no association between the exposure and outcome,  $OR > 1$  implies that the exposure is associated with greater odds of the outcome and  $OR < 1$  implies that the exposure is associated with lesser odds of the outcome.

### 3.3 Survey logistic regression model

Although logistic regression is useful in modelling data with a dichotomous outcome, it is not suitable for modelling data obtained through a complex survey that incorporates weights, stratification and clustering. Survey logistic regression is instead used to model the relationship between binary dependent variables and the set of explanatory variables by making use of the sampling design information (Lu and Yang, 2012). The inclusion of the effects of sampling design in the analysis of data leads to accurate estimation of the standard errors and variabilities (Kish, 1965; Skinner et al., 1989). The advantages of sample surveys are that: they are cost effective, speedy and timely, produce quality and accurate population

estimates and are feasible (Cochran, 1964; Kish, 1965).

The survey logistic regression model is given by:

$$\text{logit}(\pi_{ijh}) = \sum_{h=1}^H \sum_{j=1}^{m_h} \sum_{i=1}^{n_{hj}} x_{ijh}^T \beta \quad (3.22)$$

where  $i = 1, 2, \dots, n_{hj}$ ,  $j = 1, 2, \dots, m_h$  and  $h = 1, 2, \dots, H$ ,  $\beta$  is a vector of unknown regression parameters and  $x_{ijh}^T$  is a vector of the independent variables corresponding to the  $i^{\text{th}}$  individual, from the  $j^{\text{th}}$  cluster within stratum  $h$ .

### 3.3.1 Estimation of parameters

The traditional logistic regression analysis employs the maximum likelihood method to obtain parameter estimates for its model. Survey logistic regression deals with complex survey data that involves stratification, cluster sampling and the use of probability weights, therefore estimation of the parameters and their standard errors is more complex. According to Kish and Frankel (1974), stratification in survey designs creates negative correlation between variables thereby reducing the variance, while clustering increases the correlation between elements. Failure to account for the sample design effects in the analysis, leads to over-estimation of the standard errors, increase in bias and under-estimation of the variabilities. For more details see Lu and Yang (2012). Under such designs, the pseudo-maximum likelihood function is used to obtain the parameter estimates.

Suppose  $y_{ijh}$ , is a dichotomous dependent variable with  $\pi(x_{ijh}) = Pr(y_{ijh} = 1|x_{ijh})$ , and  $w_{ijh}$  is the sampling weight for observation  $y_{ijh}$ , then the pseudo-maximum likelihood func-

tion as defined by Archer et al. (2007), is given by:

$$l_p(\beta) = \prod_{h=1}^H \prod_{j=1}^{m_h} \prod_{i=1}^{n_{hj}} \pi(x_{ijh})^{w_{ijh} * y_{ijh}} [1 - \pi(x_{ijh})]^{w_{ijh} * (1 - y_{ijh})}$$

Suppose  $\beta$  is the unknown  $p \times 1$  parameter vector, the pseudo-maximum likelihood estimator,  $\hat{\beta}$ , is the value that maximizes the pseudo log-likelihood function:

$$\ln[l_p(\beta)] = \sum_{h=1}^H \sum_{j=1}^{m_h} \sum_{i=1}^{n_{hj}} \left\{ (w_{ijh} * y_{ijh}) \ln[\pi(x_{ijh})] + w_{ijh}(1 - y_{ijh}) \ln[1 - \pi(x_{ijh})] \right\} \quad (3.23)$$

The ML estimates,  $\hat{\beta}$ , is obtained by equating the score equation to 0 and solving for  $\beta$  using the iterative methods of Newton Rhapsion and Fisher scoring (SAS Institute Inc., 2015).

### 3.3.2 Variance estimation

In order to make valid inference about the population parameters in survey sampling, it is important to include the sample design in data analysis. The linearization method and the replication methods discussed extensively in Binder (1983); Cochran (1964); Efron (1980); Lu (2004); Skinner et al. (1989); Rao and Wu (1988); Rust (1986); Wolter (1985); Woodruff (1971), can be used to obtain the variance estimators of the population parameters.

#### Taylor expansion approximation

The Taylor expansion approximation method also known as the delta ( $\delta$ ) method is a linearization method for obtaining the variance of the estimators. The main idea behind this method is to reduce the non-linear forms of the estimator to a linearized quantity obtained by using the linear terms of the Taylor series expansion.

Given that  $g(\cdot)$  is a link function such that  $\pi = g(x, \theta)$  and  $\theta$  is a column vector for regression

coefficients, the pseudo-estimator  $\hat{\theta}$ , is obtained by solving the estimating equation,

$$\hat{G}(\hat{\theta}) = \sum_{h=1}^H \sum_{j=1}^{m_h} \sum_{i=1}^{n_{hj}} w_{ijh} (\text{diag}(\pi_{ijh}) - \pi_{ijh}(\pi_{ijh})')^{-1} (y_{ijh} - \pi_{ijh}) = 0 \quad (3.24)$$

The variance of  $\hat{\theta}$ , is obtained by taking a Taylor series expansion of  $\hat{G}(\hat{\theta})$  at  $\hat{\theta} = \theta_0$ , the population parameter, to obtain:

$$\begin{aligned} 0 = \hat{G}(\hat{\theta}) &\simeq \hat{G}(\theta_0) + \frac{\partial \hat{G}(\theta_0)}{\partial \theta_0} (\hat{\theta} - \theta_0) \\ \hat{G}(\theta_0) &\simeq -\frac{\partial \hat{G}(\theta_0)}{\partial \theta_0} (\hat{\theta} - \theta_0) \end{aligned} \quad (3.25)$$

Taking variances on both sides yields the limit,

$$\text{Var}[G(\hat{\theta}_0)] = \left[ \frac{\partial G(\hat{\theta}_0)}{\partial \theta_0} \right] \text{Var}(\hat{\theta}) \left[ \frac{\partial G(\hat{\theta}_0)}{\partial \theta_0} \right]^T \quad (3.26)$$

The variance of  $\hat{\theta}$  can be obtained by reversing the order of equation 3.26 above yielding;

$$\text{Var}(\hat{\theta}) = \left( \left[ \frac{\partial G(\hat{\theta}_0)}{\partial \theta_0} \right] \right)^{-1} \text{Var}(G(\hat{\theta})) \left( \left[ \frac{\partial G(\hat{\theta}_0)}{\partial \theta_0} \right]^T \right)^{-1} \quad (3.27)$$

Which in matrix form is simply,

$$\text{Var}(\hat{\theta}) = [I(\hat{\theta})]^{-1} \text{Var}[\hat{G}(\theta)] [I(\hat{\theta})]^{-1} \quad (3.28)$$

Note that  $\frac{\partial G(\hat{\theta}_0)}{\partial \theta_0}$  is the information matrix evaluated at  $\theta = \hat{\theta}$  and  $\text{var}(\hat{G}(\theta))$  is the variance covariance matrix of the  $p + 1$  estimating equations (SAS Institute Inc., 2015), and can be

estimated as:

$$\text{var}(\hat{G}(\theta)) = \frac{n-1}{n-p} \sum_{h=1}^H \frac{m_h(1-f_h)}{m_h-1} \sum_{j=1}^{m_h} (c_{hj} - \bar{c}_h)' (c_{hj} - \bar{c}_h) \quad (3.29)$$

where

$$c_{hj} = \sum_{i=1}^{n_{hj}} w_{ijh} \left( \text{diag}(\hat{\pi}_{ijh}) - \hat{\pi}_{ijh} \hat{\pi}'_{ijh} \right)^{-1} (y_{ijh} - \hat{\pi}_{ijh}) \quad (3.30)$$

and

$$\bar{c}_h = \frac{1}{m_h} \sum_{j=1}^{m_h} c_{hj} \quad (3.31)$$

The disadvantages of the linearization method is that it requires that a different variance formula be derived for each statistic, and its calculation is cumbersome in post-stratification and non-response adjustments for the estimator  $\hat{\theta}$  (Rao, 1997).

### Jackknife estimator

The Jackknife technique is a resampling method for estimating the bias and the variance of the population statistic of interest. Quenouille (1949) introduced the jackknife as a non-parametric estimate of bias and Tukey advanced Quenouille's method and established an estimate of the variance of the estimators. The jackknife estimator works through dividing the sample into disjoint but equal sized sub-samples, and obtaining the parameter estimates from each sub-sample. Each of the sub-samples is removed one at a time, while recalculating the estimates of the parameter of interest of the remaining sub-samples. The variance of the original sample is estimated from the variability amongst the sub-sample parameter estimates.

Assume we have a sample of independent and identical random quantities  $X_1, X_2, \dots, X_n \sim F$ , where  $F$  is an unknown probability distribution. Let  $\theta$ , be some unknown parameter of interest that can be approximated by  $\hat{\theta} = \theta(\hat{F})$ , where  $\hat{F}$  is the empirical probability distribution

of the sample  $x_1, x_2, \dots, x_n$ . Quenouille's method of obtaining bias involves sequentially deleting the points  $x_i$  and subsequently recomputing the value of  $\hat{\theta}$ . By deleting the point  $x_i$ , a different empirical probability distribution  $\hat{F}_i$  is obtained and the recomputed value of the parameter of interest

$$\hat{\theta}_{(i)} = \theta(\hat{F}_{(i)}) = \hat{\theta}(x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$$

Suppose that

$$\hat{\theta}_{(\cdot)} = \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{(i)}$$

Quenouille's estimate of bias is thus

$$Bias = (n - 1)(\hat{\theta}_{(\cdot)} - \hat{\theta})$$

The non-parametric estimator of the variance derived from the recomputed statistic  $\hat{\theta}_{(i)}$  by Tukey (1958), is defined as:

$$\hat{Var} = \frac{n-1}{n} \sum_{i=1}^n [\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)}]^2 \quad (3.32)$$

For stratified cluster sampling with  $H$  strata and  $n_h$  PSUs sampled from each stratum, the jackknife estimator of the variance is defined by:

$$v(\hat{\theta}^1) = \sum_{h=1}^H \frac{n_h - 1}{n_h} \sum_{j=1}^{n_h} \left( \hat{\theta}_{(hj)} - \hat{\theta}_{h(\cdot)} \right)^2 \quad (3.33)$$

where the estimator  $\hat{\theta}_{(hj)}$  is the estimator  $\theta$  obtained from the original sample after deleting the  $j^{th}$  cluster from the  $h^{th}$  stratum. The remaining clusters in the stratum are then assigned new weights called the jackknife weights, that are used in place of the sampling weights to

obtain the solutions to the estimating equations for each sub-sample. The sampling weights of the remaining strata remain unchanged. One advantage of the jackknife method over the linearization method is that, it makes use of a single variance for each nonlinear statistic. In order to cater for the asymptotic inconsistency of the resampling jackknife, Yung and Rao (2000) developed the linearized jackknife estimator of the variance. One of its advantages is that it is computationally easier to calculate.

### Bootstrap estimator

The bootstrap method introduced by Efron (1979) is another resampling procedure, for estimating the standard error, variance and confidence intervals in sample survey data. Let  $\mathbf{X} = (X_1, X_2, \dots, X_n)$  be a random sample with unknown probability distribution  $F$ , and observed outcome  $\mathbf{x} = (x_1, X_2, \dots, x_n)$ . Suppose  $R(X, F)$  is the random variable of interest. Its sampling distribution for a one sample bootstrap case can be estimated based on the sample  $\mathbf{x}$  through following these steps:

- Construct a sample probability distribution  $\hat{F}$
- Draw a sample of size  $n$  from fixed  $\hat{F}$ ,  $X_i^* = x_i^*$ ;  $X_i^* \sim \hat{F}$ ,  $i = 1, 2, \dots, n$ .
- The sampling distribution of  $R(X, F)$  is approximated by  $R^*(X^*, F^*)$ .

Suppose we are interested in estimating the bias of a functional statistic, such that

$$R(X, F) = \theta(\hat{F}) - \theta(F)$$

Then, its estimate is given by

$$R(X^*, \hat{F}) = \theta(\hat{F}^*) - \theta(\hat{F})$$

where  $\hat{F}^*$  is the empirical probability distribution of the bootstrap sample. The bootstrap estimate of bias is therefore approximated from the  $B$  bootstrap resamples:

$$BIAS = \frac{1}{B} \sum_{b=1}^n \hat{\theta}^{*b} - \hat{\theta}$$

Various bootstrap methods have been proposed in literature, they are: the rescaling method (Rao and Wu, 1988), the without replacement bootstrap (Gross, 1980) and the mirror match method (Sitter, 1992b). Their comparisons have been discussed extensively by Sitter (1992a). The rescaling bootstrap procedure draws a resample vector with replacement from the original sample, rescales each of the resampled unit, and then applies the original estimator to the rescaled vector. The procedure for a stratified cluster sampling may be described as follows:

- For stratum  $h$ , randomly select  $c_h$  clusters from the original  $n_h$  sample clusters, with replacement.
- Let  $c_{hj(d)}$  be the number of times the  $j^{th}$  cluster from the  $h^{th}$  stratum is resampled for replicate  $d$ . Here, “D” represents the number of times a draw is made,  $d = 1, 2, \dots, D$  and  $\sum_j c_{hj(d)} = c_h$ . The bootstrap weights for replicate  $d$  is defined as:

$$w_{hij(d)} = \left[ \left\{ 1 - \left( \frac{c_h}{n_h - 1} \right)^{\frac{1}{2}} \right\} + \left\{ \left( \frac{c_h}{n_h - 1} \right)^{\frac{1}{2}} \frac{n_h}{c_h} c_{hi(d)} \right\} \right] w_{hij} \quad (3.34)$$

The  $d^{th}$  bootstrap estimator  $\hat{\theta}^{(d)}$  is calculated with the sampling weights replaced by the bootstrap weights. The step is repeated  $D$  number of times, to obtain  $\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(D)}$  estimators that are used to obtain the Monte Carlo approximation of the variance of  $\hat{\theta}$  given by

$$V_{Boot}(\hat{\theta}) = \frac{1}{D-1} \sum_{d=1}^D (\hat{\theta}^{(d)} - \hat{\theta})^2 \quad (3.35)$$

where

$$\theta^{(\cdot)} = \frac{1}{D} \hat{\theta}^{(d)}$$

The bootstrap method is computationally easy, it's ideal for arbitrary samples and gives valid inferences for all forms of smooth and non-smooth statistics (Rao, 1997; Efron, 1979).

### Balanced repeated replication method

The balanced repeated replication (BRR) method is a half sampling technique for estimating the statistics of interest. Assume we have a stratified sample design from which only 2 primary sampling units are selected from each stratum,  $n_h = 2$ . Let  $H$  denote the total number of strata and  $R$  denote the total half sample replicate estimates satisfying  $H \leq R \leq H + 3$ . There are  $2^H$  possible half samples, and the estimates  $\hat{\theta}^{(r)}$ , can be calculated for each half sample. Evaluating all possible  $\hat{\theta}^{(r)}$  on the  $2^H$  half samples may be computationally expensive and intensive (Skinner et al., 1989), therefore a balanced set of  $k$  half samples may be selected. The variance of  $\theta$  can hence be estimated by the formula

$$V_{BRR}(\hat{\theta}) = \frac{1}{k} \sum_{r=1}^k (\hat{\theta}^{(r)} - \hat{\theta})^2 \quad (3.36)$$

The advantage of the BRR over the jackknife estimator is that it provides asymptotically valid inferences for both smooth and non-smooth statistics but may not be used for arbitrary sample sizes (Rao, 1997).

### 3.3.3 Model selection and fit

Logistic regression utilizes forward selection, backward elimination and stepwise selection procedures to select the variables that best fit the data set. However, these selection criteria are not implemented in SAS PROC SURVEYLOGISTIC method for analyzing complex survey data. However, the following steps, suggested by Hosmer and Lemeshow (1989), may be used for model selection:

First, perform univariate analysis between the dependent variable and the independent variables one at a time. This can be through a contingency table of the outcome and the nominal or ordinal independent variable or through fitting a univariate survey logistic regression.

Secondly, the variables that are found to be significant in the univariate analysis, and the variables known to be important to the outcome are selected for the multivariate analysis. In the third step, the relevant explanatory variables are included in to the multivariate survey logistic model one at a time. The importance of each variable is confirmed through observing the Wald statistic and also comparing its estimated coefficient with that from the univariate model. Its contribution towards reducing the deviance is also noted. This step is repeated until only the significant main effects are left in the model. Thereafter, one may consider including interaction terms amongst the variables in the model.

The AIC and the BIC discussed in section 3.1.5 are also important measures that can be used to compare two nested models when determining the better model that describes the data set. The goodness of fit tests for survey logistic models are similar to those for the logistic regression, and these are the deviance, the pearson's  $\chi^2$  and the H - L goodness of fit tests. These statistics are however based on independently and identically distributed assumptions, and therefore may give biased results for complex survey data.

### 3.4 Analysis of data using survey logistic regression procedure

Survey logistic regression procedure was applied to our study to investigate malaria infection in children under fourteen years and to determine the risk factors associated with its manifestation. It allows us to fit the best model that explains the risk factors of malaria, while catering for the design effects of the survey on the data. The data used was obtained from a two-stage cluster sample survey (Division of Malaria control [Ministry of public health and sanitation], Kenya National Bureau of Statistics, and ICF Macro, 2011), where 240 clusters and 30 households from each cluster were selected for the analysis. The outcome variable  $Y_{ijh}$  is assumed to be binary with a Bernoulli distribution of the form  $Y_{ijh}$ , representing the presence or absence of malaria in a child. That is  $Y_{ijh} \sim \text{Bernoulli}(\pi_{ijh})$ , where  $\pi_{ijh} = P(Y_{ijh} = 1)$ . The survey logistic regression model is given by:

$$\text{logit}(\pi_{ijh}) = \sum_{h=1}^H \sum_{j=1}^{m_h} \sum_{i=1}^{n_{hj}} \mathbf{x}_{ijh}^T \boldsymbol{\beta} \quad (3.37)$$

where  $i = 1, 2, \dots, n_{hj}$ ,  $j = 1, 2, \dots, m_h$  and  $h = 1, 2, \dots, H$ .  $\boldsymbol{\beta}$  is a vector of unknown regression parameters and  $\mathbf{x}_{ijh}^T$  is a vector of the independent variables corresponding to child  $i$  in household  $j$  within stratum  $h$ .

Data analysis was done using PROC SURVEYLOGISTIC command available on SAS version 9.3 to determine the risk factors associated with malaria in children under fourteen. The response variable was malaria status and the individual and household covariates were age, gender, cluster altitude in metres, malaria zone, type of place of residence, mother's highest education, wealth quartile; ownership of car, bicycle, radio and mobile phone; availability of electricity, wall material, floor material and roofing material used in household construction, anti-malarial spray, nets used for sleeping, nets per person, toilet facilities available for a

household and water source.

A univariate analysis was first performed to determine the significant predictor variables associated with the outcome of interest. Table 3.1 below displays the results of the Type 3 analysis of the univariate survey logistic regression analysis.

Table 3.1: Type 3 analysis results of the univariate survey logistic regression analysis

<b>Effect</b>	<b>df</b>	<b>Wald <math>\chi^2</math></b>	<b>P &gt; Chisq</b>
Age	1	10.1941	0.0014
Gender	1	0.0565	0.8122
Malaria zone	4	192.4842	0.0001
Type of residence	1	6.1037	0.0135
Mothers highest education	3	13.5094	0.0037
Wealth quartile	2	7.8226	0.0200
Wall material	2	33.8150	0.0001
Roof material	2	6.8683	0.0323
Floor material	2	88.1520	0.0001
Anti-malarial spraying	1	0.9374	0.3329
Source of drinking water	2	10.4118	0.0055
Toilet facility	2	13.6840	0.0011
Use of mosquito nets	1	13.4186	0.0002
Mosquito nets/person	1	1.5605	0.2116
Cluster altitude in metres	1	6.3094	0.0120
Electricity	1	13.1973	0.0003
Ownership of radio	1	0.1930	0.6604
Ownership of television	1	3.6579	0.0558
Ownership of mobile phone	1	0.1174	0.7319
Ownership of Bicycle	1	42.6218	0.0001

The variables found to be significant in the univariate analysis, (with a p-value  $<0.05$ ), were selected for the multivariate model. These were: Age, malaria zone, type of place of residence, Mother's highest education, wealth quartile, wall material, roofing material, floor material, source of drinking water toilet facility, use of mosquito nets, cluster altitude in metres, availability of electricity and ownership of bicycle. Each significant independent covariate was fitted one at a time, dropping the variables that contributed to no change in the deviance. Possible two-way and three-way interaction terms were also investigated by comparing the fit of the models with interaction terms and the main effects model. This would entail graphing these interactions to identify important factors that contribute to a change in the response, and also observing the changes to the deviance and AIC. Only the significant ones were added to the model. The final model contained the main effects and only one two way interaction term. This model was also the one with the least deviance as measured by  $-2\log L$ , compared to all the other possible models.

Table 3.2 displays the Type 3 analysis of the final survey logistic regression model.

Table 3.2: Type 3 analysis of the final multivariate survey logistic regression

<b>Effect</b>	<b>df</b>	<b>Wald <math>\chi^2</math></b>	<b>P &gt; Chisq</b>
Age	1	22.8450	0.0001
Floor Material	2	8.0485	0.0179
Altitude	1	10.6497	0.0011
Wall Material	2	8.2430	0.0162
Ownership of Bicycle	1	26.5745	0.0001
Toilet facility	2	12.2290	0.0022
Mother's Highest Education level	3	5.7829	0.1227
Floor Material*Mothers highest Education	6	23.1095	0.0008

The main effects were: age, toilet facility, cluster altitude in metres, wall material, and ownership of bicycle. The only significant 2 way interaction was between type of floor material used in household construction and mother's highest education level. Table 3.3 displays the parameter estimates, adjusted odd ratios, and their respective 95% confidence intervals for the final multivariate survey logistic regression model.

Table 3.3: Parameter estimates, adjusted Odds ratio (aOR) and 95 % Confidence intervals for the final survey logistic model

Variable	Estimate	aOR	95% C. I	S.E(SLR)	S.E(SRS)	P- value	Deff
<b>Intercept</b>	-3.1622			0.6422	0.3319	0.0001	
<b>Age</b>	0.0697	1.072	(1.042,1.103)	0.0146	0.0117	0.0001	1.25
<b>Toilet facility (Ref = No facility)</b>							
Pit Latrine	-0.9297	0.395	(0.192,0.811)	0.3677	0.1302	0.0115	2.82
Toilet with flush	-2.6953	0.068	(0.014,0.333)	0.8143	0.6089	0.0009	1.34
<b>Floor Material (Ref = Dung)</b>							
Earth/Sand	-1.3280	0.529	(0.215, 1.299)	0.4708	0.2518	0.0048	1.87
Cement	-0.9172	0.551	(0.148,2.094)	0.6830	0.4439	0.1793	1.54
<b>cluster altitude in metres (Ref - &gt;2000)</b>							
<2000	1.9541	7.057	(2.182,22.823)	0.5988	0.2472	0.0011	2.42
<b>Wall Material (Ref = Mud)</b>							
Plastic/Wood/Bamboo planks	-1.4171	0.242	(0.091,0.644)	0.4982	0.3158	0.0044	1.58
Cement	-0.3601	0.698	(0.314,1.549)	0.4072	0.2130	0.3765	1.91
<b>Ownership of Bicycle (Ref = No)</b>							
Yes	0.9476	2.580	(1.799,3.698)	0.1838	0.0934	0.0001	1.97
<b>Mother's Education ( Ref = No Education)</b>							
Primary	0.6725	1.959	(0.857,4.481)	0.4221	0.2202	0.1111	1.92
Secondary	0.1011	1.106	(0.497,2.463)	0.4083	0.2662	0.8045	1.53
Higher	0.4956	1.641	(0.270,9.992)	0.9215	0.5207	0.5907	1.77
<b>Floor Material*Education (Ref = No education/ Dung)</b>							
Earth/Sand and Primary	-0.6149	0.541	(0.185,1.584)	0.5485	0.2803	0.2623	1.96
Earth/Sand and Secondary	-0.9305	0.394	(0.114,1.370)	0.6353	0.5442	0.1431	1.17
Earth/Sand and Higher	1.5912	4.910	(0.552,43.684)	1.1152	0.7108	0.1537	1.57
Cement and Primary	-0.3371	0.714	(0.178,2.863)	0.7087	0.4627	0.6343	1.53
Cement and Secondary	-0.0555	0.946	(0.226,3.952)	0.7295	0.5058	0.9394	1.44
Cement and Higher	-4.2967	0.014	(0.001,0.268)	1.5213	1.1939	0.0047	1.27

The effects of the variables on positive malaria infection in children under fourteen years can

### 3.4. ANALYSIS OF DATA USING SURVEY LOGISTIC REGRESSION PROCEDURE61

be interpreted through odds ratios while confounding for the effects of the other variables. A unit increase in age in years of the children, implies an increase in the odds for malaria infection (OR = 1.072, 95% C.I: 1.042, 10103, p - value = 0.0001). The variable cluster altitude in metres was treated as a categorical variable with two different levels ( <2000m, and >2000m). In comparison to households in greater than 2000 metres in altitude, households within clusters located in altitudes less than 2000metres had children with increased odds of testing positive for malaria (OR = 7.057, 95% C.I: 2.182, 22.820, p-value = 0.0011).

Compared to houses with no toilet facility, households using pit latrines and those with flush toilets, had decreased odds for malaria infection (OR = 0.395, 95% C.I: 0.192, 0.811, p-value=0.0015) and (OR=0.068, 95% C.I: 0.014, 0.333, p-value=0.0010) respectively. Ironically, households who own a bicycle had greater odds for positive malaria infection in children in comparison to those without a bicycle (OR=2.580, 95% C.I: 1.799, 3.698, p-value=0.0001).

The type of wall material used in the house construction was also a significant factor related with malaria infection. Houses that had either cemented walls or walls made from bamboo planks and wood, were 30% and 76% less likely to have a child testing positive for malaria respectively. The result was more so significant for houses with bamboo planks and wooden walls (OR = 0.242, 95% C.I :0.088, 0.650, P-value = 0.0050).

#### **Interaction terms**

The relationship between floor material used in household construction and the mother's highest education level is presented in Figure 3.1.

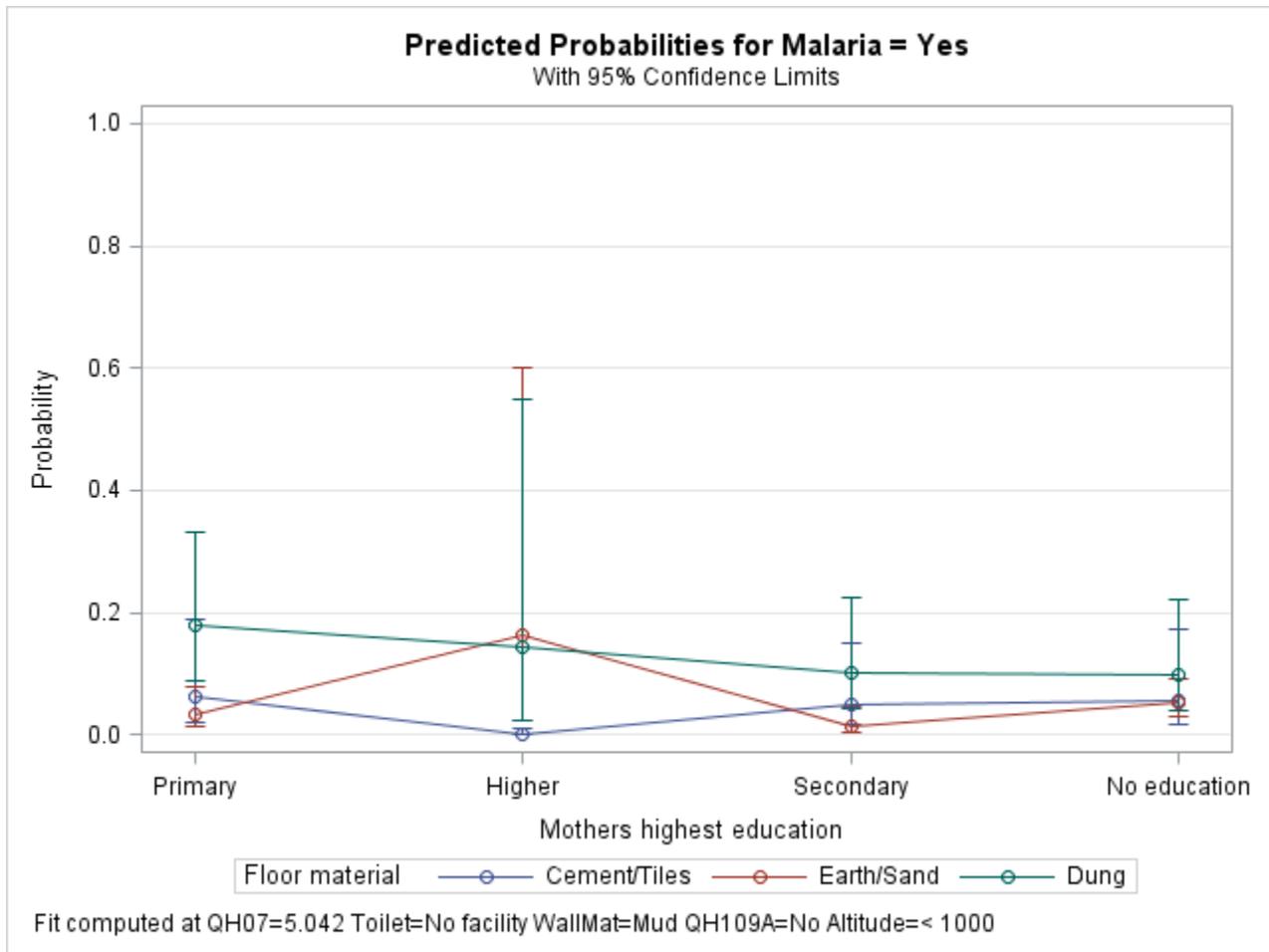


Figure 3.1: Interaction of mother's highest education and floor material use in household construction

The risk of testing positive for malaria was highest amongst households with dung floors and a mother with primary education. In the same span, the risk was lowest in households with cemented floors and mother's with higher education. Generally, households with cemented floors had lower odds for malaria across all levels of mother's education. The odds were higher for households with dung floors across all levels of mother's education. The risk was relatively low for households with mother's with no education across all floor types. This could be because the frequency of uneducated women from the exploratory data analysis was quite low. Most women had attained at least primary level of education.

One other measure reported on Table 3.3 is the design effect (Deff), defined as:

$$def f(\hat{\theta}) = \frac{Var_{SLR}(\hat{\theta})}{Var_{SRS}(\hat{\theta})} \quad (3.38)$$

which explains the large variability in the estimates obtained through survey logistic regression (SLR) modelling compared to the same estimates obtained through simple random sample (SRS) methods such as logistic regression model. Its value is usually greater than 1, indicating that deviating from SRS leads to increased variability of the estimates (Kish, 1965; Kish and Frankel, 1974; Skinner et al., 1989).

The PROC SURVEYLOGISTIC procedure displays the following four statistics for assessing the predictive accuracy of the model: Concordance index (C), Goodman-Kruskal Gamma (GKC), Somer's D (SD), and the Kendal's Tau-a (KT). These are defined as:

$$\begin{aligned} c &= [n_t - 0.5(t - n_c - n_d)]t^{-1} \\ SD &= (n_c - N_d)t^{-1} \\ GKC &= (n_c - n_d)(n_c + n_d)^{-1} \\ KT &= (n_c - n_d)[0.5N(N - 1)]^{-1} \end{aligned}$$

The meaning of  $N$  is the sum observation frequencies in the data,  $t$  is the total number of pairs with different responses,  $n_c$  are the concordant pairs,  $n_d$  are the discordant pairs and  $t - n_c - n_d$  are the tied pairs. The concordance index ( $c$ ), ranges between 0 and 1, and it is equivalent to the area under the receiver operating characteristic (ROC) curve. A value of 0 indicates no association and a value between 0.5 and 0.6 shows a poor predictive accuracy.

If the value lies between 0.7 and 0.8 the accuracy is said to be moderate and c greater than 0.8 connotes excellent accuracy. The Somers' D is another useful statistic that measures rank correlation between the binary response variable and the predicted probabilities. It ranges between  $-1$ , showing negative association and  $1$  indicating positive association. The Goodman-Kruskal's Gamma measures the difference between the probability of concordance and the probability of discordance (Agresti, 1990).

The concordance index for the final multivariate survey logistic model was 0.823 indicating that the model is excellent in predicting malaria infection in children.

### 3.5 Summary and discussion

Since the data used in the analysis was obtained from a complex survey, the survey logistic regression method under the generalized linear models was relevant in assessing the risk factors associated with positive malaria in children. The findings from the study show that age and the cluster altitude in metres are the important demographic and geographical factors affiliated to malaria in children. Similarly, type of toilet facility, ownership of bicycle and the wall material used in housing construction were important socio-economic factors linked to malaria.

Our results are consistent with other studies showing a link between malaria prevalence and age (Ayele et al., 2012; Gahutu et al., 2011; Siri, 2014). Of concern is the fact that the odds for malaria infection seems to increase with increasing age. In separate studies conducted in Rwanda and Tanzania, Kateera et al. (2015) and Winskill et al. (2011) respectively, find that children aged between 5 and 15 years had higher odds for malaria compared to children under 5 years.

Households located in cluster altitudes below 2000 metres were found to be more at risk for malaria infection while those in cluster altitudes above 2000 metres were less susceptible to malaria. The literature on this correlation upholds these findings, that the risk for malaria decreases with increasing altitude (Brooker et al., 2004; Ernst et al., 2006; Githeko et al., 2006; Graves et al., 2009; Woyessa et al., 2013).

The findings on socio-economic factors associated with malaria prevalence are similar to those from previous studies (Ayele et al., 2012, 2013; Gahutu et al., 2011). The results reveal that households that had toilet facilities, either pit latrines or flush toilets, were less likely to have children infected with the malaria parasite. This implies that households with no toilet facility had higher odds for positive malaria. Ownership of a bicycle is also a symbol of the wealth status of a family. Ironically, households that at least owned a bicycle had higher odds for malaria infection in children. Similarly, a study in Tanzania by de Castro and Fisher (2012) finds no significant association between malaria and the socioeconomic status of a household. The effect of type of housing structure on the prevalence of malaria has been of great interest to many researchers (Chirebvu et al., 2014; Gamage-Mendis et al., 1991; Sintasath et al., 2005). Our results show that poorly constructed houses with mud walls had higher incidences of malaria than those with better wall constructions such as cement and wood/bamboo planks. This result is consistent with the transmission mechanism of malaria. Poorly constructed households create ideal conditions for contact between the mosquito vector and the human host, hence increased probability of transmission from an infested mosquito.

The interaction between mother's highest education level and the type of floor material used in household construction was significant in explaining malaria risk in children. Households with mother's with at least the basic education and living in homes with cemented

floors had lower chances for positive malaria infection. Regardless of the mother's education level, homes with floors constructed of dung had higher risks for malaria infection.

Malaria infection has for long been perceived as a disease of poverty (Worrall et al., 2005). The socioeconomic factors such as toilet facility, ownership of household assets and type of housing structure are indicators of the wealth status of a household. This study suggests that having better toilet facilities reduces the chances of malaria illness. Different housing structures also contributes differently to the risk of malaria. Houses constructed with good materials such as cement, iron sheets, and bricks, on the walls, roof or floors would greatly reduce the risk of its members testing positive for malaria. Literature shows that poorly constructed houses provide easy entry and resting environments for the mosquitoes (Atieli et al., 2009; Howell and Chadee, 2007; Schofield and White, 1984). This in turn increases human exposure to the vector hence increasing incidences of malaria.

The study shows that the ownership of a bicycle did not greatly contribute towards decreasing the susceptibility towards malaria. Bicycles may quite generally be used as a source of transport to the nearest health centre, or a faster means to purchase malarial drugs, whenever an individual manifests its symptoms. However, the results of the study indicate that despite many families owning one, its use may have been quite limited. These poverty related factors are important factors to be considered by policy makers. Improving the living conditions of households through providing proper sanitation, and improved housing structures may prevent poor health amongst household occupants. They can also be used to highlight the use of intervention methods and treatment seeking procedures Worrall et al. (2003, 2005).

Closely linked to these socioeconomic status is the education level of the primary caregivers. Our study reveals that homes with mother's with the basic level of education had

better living quarters and hence less likely to have childhood malaria infections. Therefore, improving the literacy knowledge of mother's indirectly improves the welfare of the household. Siri (2014) alludes that educated mothers are better equipped to improve the living conditions of a family, invest in control measures such as ITNs, have knowledge on malaria and make use of the health systems for treatment.

The results obtained through SLR modelling tends to be unbiased, since it takes into account the complex design of the sample in the analysis. However, variability due to correlation amongst the elements selected from the same household and/or cluster also needs to be incorporated in the analysis. Therefore, the next chapter introduces the Generalized linear mixed effects model (GLMM), an extension of the GLM that fits outcomes with non-normal distributions and includes the random effects in addition to the fixed effects in the analysis.

# Chapter 4

## The Generalized Linear Mixed Models

### 4.1 Introduction

The previous chapter made use of survey logistic regression modelling under generalized linear models to investigate the prevalence and risk factors associated with malaria in children. This chapter provides us with an alternative method for modelling malaria in children, given that our data was collected from a survey that incorporated stratification and cluster sampling, that could lead to variability and correlation amongst subjects from households within the same cluster.

The generalized linear mixed effects model (GLMM) is an extension of the generalized linear model (GLM) by Nelder and Wedderburn (1972), and allows Statisticians to model non-normal and non-linear data that includes both random effects and fixed effects. According to McCulloch et al. (2008) random effects are achieved from factors with infinite levels drawn from a sample in a population, and the main interest is in the variations in the levels. The GLMM's structure is similar to that of a GLM with the only difference being the introduction of the random effects in the linear predictor. The GLMM is an important model in

solving the problems of over-dispersion and also makes inference of the population heterogeneity. It is applicable in various research fields including epidemiology, ecology, actuarial statistics, educational studies, biomedical studies and household surveys (McCulloch et al., 2008; Verbeke and Molenberghs, 2009; Antonio and Beirlant, 2007; Agresti et al., 2000).

This chapter will focus on using the GLMM to investigate the risk factors associated with malaria in children. We will discuss the structure of the GLMM, provide various methods of estimation for both the fixed effects and random effects parameters and finally apply the model to our data.

### 4.1.1 The model structure

GLMMs are an advancement of the linear mixed model (LMM) proposed by (Laird and Ware, 1982), that caters for outcomes that are non - Gaussian in nature (McCulloch et al., 2008). The general structure of the LMM with both fixed and random effects is:

$$Y = X\boldsymbol{\beta} + Z\mathbf{b} + \boldsymbol{\epsilon} \quad (4.1)$$

$$\mathbf{b} \sim N(0, D) \quad (4.2)$$

$$\boldsymbol{\epsilon} \sim N(0, R) \quad (4.3)$$

where:

$Y$  is  $N \times 1$  response vector of observations

$X$  is  $N \times p$  model matrix for fixed effects

$\boldsymbol{\beta}$  is  $p \times 1$  vector for fixed effects coefficients

$Z$  is  $N \times q$  model matrix for random effects

$\mathbf{b}$  is  $q \times 1$  vector of random effects coefficients

$\epsilon$  are the error terms for the observations

$D$  is  $q \times q$  variance and covariance component of the random effects

$R$  is  $N \times N$  matrix for the error terms

Given the random effects parameter  $b_i$ , the response variable,  $y_{ij}$  (the  $j^{\text{th}}$  observation from cluster  $i$  for  $i = 1, 2, \dots, N$  and  $j = 1, 2, \dots, n_i$ ) are assumed to be independent and have a distribution similar to that from the exponential family. The random effect parameter  $b_i$  is drawn independently and has a distribution of  $f(b_i|D)$ .

$$y_{ij}|b_i \sim f_{y_{ij}|b_i}(y_{ij}|b_i, \xi_{ij})$$

$$f_{y_{ij}|b_i}(y_{ij}|b_i) = \exp \left[ \frac{y_{ij}\xi_{ij} - b(\xi_{ij})}{a(\phi)} + c(y_{ij}, \phi) \right] \quad (4.4)$$

The conditional mean  $\mu_{ij} = E(y_{ij}|b_i)$  is modeled as:

$$g(\mu_{ij}) = \eta_{ij} = x'_{ij}\beta + z'_{ij}b_i$$

where  $g(\cdot)$  is the link function and  $\eta(\cdot)$  is the linear predictor.

### 4.1.2 Estimation of the parameters

The maximum likelihood method is the preferred estimation method for parameters in a GLM using iterative methods such as the Newton Raphson method, Fisher scoring and the Iterative weighted least squares procedure (McCullagh and Nelder, 1983). Under certain regularity conditions of the likelihood, quadratic convergence of the iterations is achieved (Gad and El Kholy, 2012). In a GLMM model, the  $i^{\text{th}}$  subject contribution toward the

likelihood is defined by:

$$f_i(y_{ij}|\beta, D, \phi) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|u_i, \beta, \phi) f(b_i|G) db_i$$

Therefore the likelihood for  $\beta, D, \phi$  is given as:

$$L(\beta, D, \phi) = \prod_{i=1}^N f_i(y_{ij}|u_i, \beta, \phi) f(u_i|D) du_i \quad (4.5)$$

$$= \prod_{i=1}^N \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}|b_i, \beta, \phi) f(b_i|D) db_i \quad (4.6)$$

Obtaining the ML estimates for equation 4.5 involves integrating over the  $b_i$  random effects. Neyman and Scott (1948), question the efficiency of the ML method in estimating parameters in partially consistent situations where the set of unknown parameters is infinite. Since integration of the likelihood for a GLMM is intractable, producing inconsistent estimates and is computationally infeasible for high dimensions of the random effects, various other methods have been proposed for parameter estimation. These methods include: Penalized quasi-likelihood, marginal quasi-likelihood, Laplace approximations, Gauss - Hermite quadrature, Markov chain Monte Carlo approximations and the Gibbs sampler.

### Penalized Quasi-likelihood

The penalized quasi - likelihood (PQL) method (Breslow and Clayton, 1993) approximates the integral of the quasi-likelihood given by  $ql(\beta, \xi)$ , by decomposing data in to the mean and the error terms and performing a Taylor series expansion of the mean. The integrated quasi-likelihood function for estimating  $(\beta, \xi)$  is defined as:

$$e^{ql(\beta, \xi)} \propto c|D|^{-1/2} \int e^{-k(b)} db \quad (4.7)$$

where:

$$k(b) = \frac{1}{2\phi} \sum_{i=1}^n d_i(y_i; \mu_i) + \frac{1}{2} b^T D^{-1} b$$

$c$  is a constant term,  $b$  is the random effects parameter with a multivariate normal distribution with mean 0 and covariance matrix  $D = D(\theta)$ ,  $\beta$  are the fixed effects

$$Q_i = d_i(y_i; \mu_i) = \int_{y_i}^{\mu_i} \frac{y_i - t}{\tau^2 V(t)} dt \quad (4.8)$$

is the conditional quasi likelihood, with  $\tau$  as the constant of proportionality that relates  $\text{var}(y_i)$  to  $v(\mu_i)$  (McCulloch et al., 2008). By adding a penalty of  $\frac{1}{2} b^T D^{-1} b$  to the QL one obtains the PQL which when partially differentiated with respect to the fixed effect and random effect parameters yields the maximum quasi-likelihood equations that can be solved iteratively through fisher scoring or the Newton Raphson methods.

$$PQL = \sum_{i=1}^n Q_i - \frac{1}{2} b^T D^{-1} b \quad (4.9)$$

The PQL method is implemented in various statistical software although it has a few disadvantages. McCulloch et al. (2008) postulates that it is not an efficient method for parameter estimation of binary data in small clusters. It also produces asymptotically biased results in the regression coefficients and variance components especially for correlated Poisson and binary data. More details on the bias of the PQL as authored by various researchers can be found in McCulloch et al. (2008); Breslow and Clayton (1993); Lin (2007); Breslow and Lin (1995).

### **Marginalized quasi-likelihood**

The Marginalized quasi likelihood (MQL) is related to the PQL method. Although it is commonly used when interest is on the marginal relationship between the co-variables and the

outcome of interest (Breslow and Clayton, 1993). The MQL method is discussed extensively by (Breslow and Clayton, 1993), who apply the first and second order marginal moments of the response variables to estimate the regression parameters and (Sutradhar and Rao, 2001) who compute the joint moments of the clustered observations up to the fourth order.

The PQL and the MQL methods both provide a linear approximation of the integrand based on the Taylor expansion of the conditional mean. However, the expansion in the MQL method is based on the current estimates for  $\beta$  and the zero vector for the random effects (Tuerlinckx et al., 2006). The estimating equations are again solved iteratively to obtain the parameter estimates.

### Laplace approximation

The Laplace approximation method is widely used in approximating the likelihood functions of the closed form. Breslow and Clayton (1993); Breslow and Lin (1995) use the Laplace approximation method and second and fourth order expansion methods respectively to obtain the likelihood equations useful for obtaining the GLMM parameter estimates. In the Laplace approximation method, the integral of the form  $I = \int e^{-Q(b)} db$  can be approximated through second order Taylor series expansion of its logarithmic integrand. The resulting integral, assumed to be Gaussian distributed is then evaluated normally.

### Gauss-Hermite quadrature

The Gauss-Hermite quadrature is a numerical method of approximating the intractable integral of the marginal likelihood. Liu and Pierce (1994) defines the Gauss-Hermite quadrature as follows:

Given an integral of the form:

$$\int_{-\infty}^{\infty} f(z)\Phi(z)dz \quad (4.10)$$

The Gauss-Hermite quadrature approximation is given as

$$\int_{-\infty}^{\infty} f(z) \exp(-z^2) dz \approx \sum_{q=1}^Q w_q f(z_q) \quad (4.11)$$

where  $z_q$  are nodes, with zero value of the  $r^{th}$  order in the Hermite polynomial,  $Q$  is the order of approximation and  $w_q$  are the weights.

### Markov chain Monte Carlo methods

Markov chain Monte Carlo methods described in Hastings (1970) are more useful in solving high dimensional numerical integrations. They are used in solving integrals of the form:

$$I = \int f(x)p(x)dx$$

Given that  $p(x)$  is a probability distribution function of the random variables. The standard Monte Carlo methods for solving the integral involved drawing  $N$  independent and identical samples of  $X$  from the density  $p(x)$ , and using its estimate  $\hat{I}_1 = \frac{1}{N} \sum f(x_i)$ . However, drawing samples from the density  $p(x)$  is not always feasible, therefore the integral may be approximated through Markov chain processes described in Hastings (1970) and Spall (2003). MCMC methods allow for drawing of samples independently without sampling from  $p(x)$  and producing dependent Markov sequences with density  $q(x)$  approximately equal to the density of interest  $p(x)$ . The MCMC method is especially useful in Bayesian analysis and has been applied in two algorithms, the Gibbs sampler and metropolis hastings methods.

The Gibbs sampler method discussed extensively in Casella and George (1992) and Zeger and Karim (1991) is a Markov chain Monte Carlo method of obtaining valid observations from the joint distribution by invoking a Markov chain through repeated sampling from the conditional distribution (Rodriguez and Goldman, 2001). Suppose we have a vector  $\theta$ , of

three parameters,  $\theta_1, \theta_2, \theta_3$ , we want to draw a sample from the posterior distribution  $p(\theta|y)$  using the following steps: We pick an arbitrary vector  $\theta^{(0)}$ , and start with any parameter  $\theta$ , say  $\theta_1$ . Draw  $\theta_1^{(1)}$  from the conditional distribution,  $p(\theta_1|\theta_2^{(0)}, \theta_3^{(0)}, y)$ . Again, we draw  $\theta_2^{(1)}$  from the conditional distribution,  $p(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, y)$  with the updated value of  $\theta_1^{(1)}$ , and complete the first iteration by drawing  $\theta_3^{(1)}$  from the conditional distribution  $p(\theta_3|\theta_1^{(1)}, \theta_2^{(1)}, y)$  with the updated value from the second draw.  $\theta^{(2)}$  is drawn using the updated values from  $\theta^{(1)}$ . The steps are repeated for  $M$  iterations to obtain the empirical distribution that is used to approximate the joint distribution of the parameters. Monte Carlo integration on the draws is then performed to obtain the quantities for the parameter of interest. The Gibbs sampler method is applicable to multivariate and non-gaussian random effects and is easy to implement (Zeger and Karim, 1991).

The Metropolis Hastings (M - H) algorithm developed by Metropolis et al. (1953) for use in Physics and generalized by Hastings (1970) is described extensively by Chib and Greenberg (1995) and is also useful in Bayesian statistics for numerical integration. The Gibbs sampler is a special case of the M-H algorithm (Zeger and Karim, 1991), and is used when the conditional probabilities are known. Given a likelihood function of the form given in equation 4.6, the joint distribution of  $[\beta, D, b]$  and its marginal distributions  $[\beta, D|y]$  and  $[b_i|y]$  can be obtained from the conditional distribution. The Gibbs sampler algorithm follows these steps to generate the joint distribution for the random variables  $\beta, D$  and  $b$ :

- Given the arbitrary starting values  $\beta^{(0)}, D^{(0)}, b^{(0)}$ , draw  $\beta^{(1)}$  from the conditional distribution  $[\beta|D^{(0)}, b^{(0)}]$ . Then draw  $D^{(1)}$  from the conditional distribution  $[D|\beta^{(1)}, b^{(0)}]$ . The first iteration is completed by drawing  $b^{(1)}$  from the conditional distribution  $[b|\beta^{(1)}, D^{(1)}]$ .
- After  $M$  iterations, we obtain  $(\beta^{(m)}, D^{(m)}, b^{(m)})$ .

- The joint distribution is approximated by the empirical distribution of the  $R$  values  $(\beta^{(k)}, D^{(k)}, b^{(k)})$ , ( $k = M + 1, M + R$ ), for a sufficiently large  $M$ , so that the Gibbs sampler converges.

### 4.1.3 Inference for GLMM's

Once the parameter estimates for a GLMM have been obtained through the methods discussed in section 4.1.2, statistical inference can be performed. This is an important step in statistical analysis since one can accurately make conclusions about the population based on the results obtained from the population sample.

#### Inference for fixed effects

GLMMs are mostly fitted using maximum likelihood methods, therefore the estimates obtained are usually asymptotically normally distributed. The classical tests that are useful in making inference for the fixed effects parameters have been analyzed and discussed by several authors (Verbeke and Molenberghs, 2009; Bolker et al., 2008; Tuerlinckx et al., 2006; McCulloch et al., 2008). They are: Wald tests, F - tests T -tests, score statistics, and the likelihood ratio tests. An approximate Wald test and the corresponding confidence intervals can be obtained through approximating the distribution of  $\frac{(\hat{\beta}_j - \beta_j)}{S.E(\hat{\beta}_j)}$ , by the standard univariate normal distribution, for each parameter  $\beta_j, j = 1, \dots, p$ . More generally to test the hypothesis:

$$H_0 : L\beta = 0 \quad \text{against} \quad H_A : L\beta \neq 0 \quad (4.12)$$

for any matrix  $L$ , the distribution:

$$(\hat{\beta} - \beta)' L' \left[ \left( \sum_{i=1}^N X_i' V_i^{-1}(\hat{\alpha}) X_i \right) \right]^{-1} L(\hat{\beta} - \beta) \quad (4.13)$$

is asymptotically chi square distributed with rank (L) degrees of freedom. The Wald tests are based on estimated standard errors, that may fail to take into account the variability in  $\hat{\beta}$  introduced by estimating  $\alpha$ , the vector of variance components (Verbeke and Molenberghs, 2009). This bias can be corrected by using approximate t and F statistics to test the hypothesis given in equation 4.12. The distribution of  $\frac{(\hat{\beta}_j - \beta_j)}{S.E(\hat{\beta}_j)}$ , can be approximated by t - distribution, in order to obtain the t - tests and confidence intervals for the parameters  $\beta_j$ . The F - statistic is given by

$$F = \frac{(\hat{\beta} - \beta)' L' \left[ \left( \sum_{i=1}^N X_i' V_i^{-1}(\hat{\alpha}) X_i \right)^{-1} L(\hat{\beta} - \beta) \right]}{\text{rank}(L)} \quad (4.14)$$

and it also takes in to account the variability that arises as a result of estimating the dispersion parameter  $\phi$ . It is estimated by taking the Wald statistic and dividing by the degrees of freedom for the test. The Likelihood ratio test (LRT), can also provide inference for the fixed effects especially in nested model with different mean structures. The LRT statistic can be defined as:

$$LR = -2 \ln \left[ \frac{L_{ML}(\hat{\theta}_{ML})}{L_{ML}(\tilde{\theta}_{ML})} \right] \quad (4.15)$$

where  $\hat{\theta}_{ML}$  and  $\tilde{\theta}_{ML}$  is the maximum likelihood estimate of  $\theta$  obtained after maximizing the marginal likelihood function over the restricted model and the unrestricted model respectively. The score statistic is also referred to as the Lagrange multiplier test and is given by:

$$U = s(\tilde{\beta}) F_{\tilde{\beta}}^{-1} s(\hat{\beta}) \quad (4.16)$$

where  $s(\tilde{\beta})$  is the score function under the restricted model while  $F_{\tilde{\beta}}$  is the expected or the observed information matrix. The Wald statistic and the score tests are based on the quadratic approximation of the log-likelihood function, and thus are more advantageous since they only require fitting a single model unlike the LR tests which require fitting of

both the restricted model and the unrestricted model. The Wald and F tests also depend on the dispersion parameter. The Wald statistic and the score test are based on the quadratic approximation of the log-likelihood function. The LR tests can be used in models whether the dispersion parameter is known or unknown. The LR tests provide better results than the Wald, score tests and F tests because they use information from both the null model and the maximal model.

### **Inference for random effects**

The GLMM of the form  $g(\mu_i) = x_i^T \beta + z_i^T b_i$  has subject specific random effects which are usually multivariate normally distributed with mean of zero and a variance covariance matrix  $D(\theta)$ . The vector  $\theta$  is a  $q \times 1$  vector of the unknown variance components. The presence of random effects in mixed models, makes it necessary to model the within subject and between subject variations, in the covariates (Zhang and Lin, 2008). The statisticians therefore tests whether no subject variation exists in the covariates of  $Z$ , therefore the null hypothesis would be to test whether  $\theta = 0$ . This may result in some of the variance components to lie on the boundary of the parameter space, and thus causing tests such as the Wald tests, Score tests and the LR tests not to have a chi square distribution Lin (1997). According to (Verbeke and Molenberghs, 2003) it is paramount for a statistician to determine between one sided and two sided tests whenever one has to make inference about the variance components in mixed models. For the two sided tests of hypothesis, under the null distribution, the traditional inference tests, LR tests, Wald tests and Score tests are used and are asymptotically chi square distributed.

Under certain regularity conditions, the distribution of the MLE  $\hat{\theta}$ , obtained through likelihood theory can be approximated by normal distribution with a mean vector  $\theta$  and covariance matrix given by the inverse of the Fisher information matrix (Verbeke and Molenberghs,

2009). By using the asymptotic normality of the parameter estimates, one is able to obtain the approximate Wald test and corresponding confidence intervals. This procedure fails in restricted hypothesis testing, and may require derivation of one sided test statistics and their null distribution. The calculation of score tests in one sided hypothesis depends on the information matrix and requires computation of mixed chi square distributions Zhang and Lin (2008). The score tests are computationally easy as they only require fitting of the GLMM under the null hypothesis.

Stram and Lee (1994) shows that, in one sided tests of hypothesis, with  $k$  correlated random effects against the alternative  $k + 1$  correlated random effects, the asymptotic null distribution for the LRT statistic, has a mixture of  $\chi_k^2$  and  $\chi_{k+1}^2$ , with equal probability 1/2 (Zhang and Lin, 2008; Verbeke and Molenberghs, 2003, 2009). The random effects  $b_i$  in a GLMM are assumed to be random variables with a multivariate normal marginal distribution. Empirical Bayes estimate for  $b_i$  can be obtained by calculating the mean of its posterior distribution, ( $f(b_i|y_i)$ ), which is conditional on the observed values of the response variable  $y_i$ . Inference on  $b_i$  is obtained from  $var(\hat{b}_i(\theta) - b_i)$ .

The process of model selection is important in statistical analysis and involves comparing and selecting the best model amongst many models with different numbers of parameters. The most commonly used method for model selection especially in linear mixed models is the information criterion (IC), based on the maximized log-likelihood function. The IC is given as:

$$IC = -2\log(\lambda_m) + \varpi k$$

where  $\log(\lambda_m)$  is the maximized log likelihood function,  $k$  is the number of parameters. When the penalty  $\varpi = 2$ , the IC approach is the AIC (Akaike, 1974), and when  $\varpi = \log(n)$ , it is

the BIC (Schwarz, 1978). Lindstrom and Bates (1990) suggests that these IC approaches, can be extended to Normally distributed, non-linear mixed effects models. These methods are based on the likelihood functions whose integrals are numerically intractable in GLMMs thus, it may be difficult to estimate the marginal distribution. Several other methods for assessing the goodness of fit have been proposed. Vonesh et al. (1996) came up with the concordance correlation coefficient, similar to the  $R^2$  for linear models, that measures the level of concordance between the fitted and observed responses. This method does not require the specification of the null model, but is however not useful for discrete data. Pan and Lin (2005) developed procedures for assessing the adequacy of GLMMs by taking the cumulative sums of residuals over the predicted values. Besides not requiring specification of the alternative hypothesis, they give information on the nature of model misspecification and also check the model fit for the random components. (Lavergne et al., 2008) proposes a simple IC approach obtained by computing the log-likelihood corresponding to the LMM for the final working variable, given as

$$IC^S = n \log(2\pi) + \log(|\hat{\Gamma}|) + (Z^f - X\hat{\beta})' \hat{\Gamma}^{-1} (Z^f - X\hat{\beta}) + \varpi k \quad (4.17)$$

$\hat{\Gamma} = \hat{W} + UD'U'$ ,  $U$  and  $D$  are the independent variables and the variance-covariance matrix for the random effects, respectively,  $Z^f$  is the final working data and  $k = p + s$ , the sum of fixed and random effects parameter lengths. The GLMM with the smallest  $IC^S$  is selected. Another alternative to assessing the goodness of fit is the use of graphical procedures such as residual plots (Pan and Lin, 2005). They are mostly useful in assessing model fit for independent outcomes whose residuals are uncorrelated.

#### 4.1.4 Advantages and disadvantages of GLMMs

Generalized mixed effects models (GLMMs) were developed to cater to some of the limiting assumptions of linear mixed models (LMMs), as well as generalized linear models (GLMs). The LMMs assume that the relationship between fixed and random effects and the mean of the outcome variable can be modelled through a linear function. They also presume that the variance function of the mean and random effects follows a normal distribution which is not true for binary outcome random variables. They again deduce that predictions can take on any values from negative infinity to positive infinity, an assumption that does not hold for binary outcome variables bounded in the range  $(0, 1)$ , and count data that only takes positive values. Treating correlations amongst observations as fixed effects has led to overdispersion in generalized linear modelling, and therefore GLMMs prove to be useful in handling non-normal data that also have random effects. The GLMM procedure is computationally easy to use and has been implemented in most statistical packages. It may however be challenging for complex models, where the choice of the significance of the random effect would depend on the researcher's objectives. Most of the assumptions of GLMMs are drawn from LMMs and GLMs, contributing to high risk of model misspecification, producing biased parameter estimates. Some of the model testing and inferential procedures may also not be applicable in these models. They are also limited to linear functions yet some predictor values may be sigmoidal.

#### 4.1.5 GLMMs for binary response data

Several research studies have been conducted for modelling dichotomous data whose covariates have both fixed and random effects, (Pendergast et al., 1996; Zhang et al., 2011; Capanu et al., 2013; Agresti, 1990). Researchers incorporate logistic regression models with mixed effects in such situations which uses the logit link. Let  $Y_{ijh}$  be a binary response variable cor-

responding to the  $i^{th}$  observation in the  $j^{th}$  household within the  $h^{th}$  cluster. The conditional mean of  $Y_{ijh}$  depends on the fixed and random effects,  $\mu_{ijh} = E(Y_{ijh}|b_i) = Pr(Y_{ijh} = 1|b_i)$ . It is related to the linear predictor via a logit link function such that:

$$g(\mu_{ijh}) = \text{logit}(\mu_{ijh}) = \log\left(\frac{\mu_{ijh}}{1 - \mu_{ijh}}\right) = \eta_{ijh}$$

In the previous chapter, we made use of survey logistic regression approach under GLMs to identify the risk factors associated with malaria in children. As much as the method takes in to account the complex survey design utilized in the household survey, it does not take in to account the variability as a result of drawing samples from the same sampling unit. The GLMM has the advantage of modelling these variability and thus will be applied in this chapter.

#### 4.1.6 Data analysis

The PROC GLIMMIX procedure available in SAS version 9.3 was used to analyze our data. The response variable was malaria status indicating whether child  $i$ , from household  $j$  within cluster  $h$  tested positive for malaria. The geographic as well as the demographic covariates were: age, gender, cluster altitude in metres, malaria zone, and type of place of residence. The socio-economic variables were: toilet facility; water source; household structure that included roofing material, wall material and floor materials used in household construction; availability of electricity; ownership of mobile, radio, television and bicycle; and wealth quintile. Associated with the socio-economic factors were the intervention factors such as use of mosquito nets while sleeping, nets per person, number of mosquito nets and antimalarial spraying.

The distribution of the response variable and the link function are specified through the DIST= and LINK= options in the model statement of the PROC GLIMMIX procedure.

The binary distribution option with logit link were used in this model. The parameter and covariance estimation techniques are specified using the `METHOD =` option syntax in the `PROC GLIMMIX` statement. The default estimation technique in the `GLIMMIX` procedure is the residual pseudo likelihood method (RSPL). However, the model did not converge, hence the marginal distribution was approximated by using the Gauss-Hermite quadrature approximation method and the Laplace methods. Both these methods produced similar results, with minimal differences in the parameter estimates and their corresponding standard errors. However, the final model was based on the Gauss-Hermite quadrature with 20 quadrature points.

The `Random` statement specifies the G-side and the R-side random effects and their covariance structures. The clusters and the households in the data set were chosen at random. The random effect was the "household" effect.

Table 4.1: Covariance parameter estimates

Cov Parm	Subject	Estimate	Standard error
Chol (1,1)	HouseID	0.3671	0.08657

Inference about the covariance parameters can be made through likelihood based, tests of significance produced by the `COVTEST` statement. The results of the test is given in Table 4.2. The significance of the random effect was obtained by testing whether the G matrix can be reduced to a zero matrix.

Table 4.2: Tests of covariance parameters based on the likelihood

Label	DF	-2 Log like	Chisq	Pr>Chisq
No. G-side effects	1	2110.44	13.88	0.0001

The results show that the household effect is significant (i.e. there is evidence of heterogeneity in malaria status in children from different household). Further, the variance component

of the random effect was estimated at 0.3671 with a standard error of 0.08657 (see table 4.1) confirming its significance. The model was fitted with different covariance structures for the G matrix. The default covariance structure in the PROC GLIMMIX procedure is the variance component given by the syntax “type= vc “. The model was fit using the “type=chol” option, that is numerically stable, and it specifies an unstructured variance-covariance matrix through the Cholesky root.

Selecting the best fit model was achieved through various processes. All the predictor variables were first fit into the model and a process of backward selection criterion was applied until only the significant effects (with p - value <0.05) remained. Two-way interactions and higher order interactions were also exploited. A comparison of the models was done through the information criteria statistics such as the AIC and BIC.

The results of the Type 3 analysis of the final model selected are displayed on Table 4.3. Statistical inference about the fixed effect parameters are based on Wald tests and are also dependent on the estimated covariance matrix (Gurka et al., 2011).

Table 4.3: Type 3 analysis of the fixed effects for the final multivariate GLMM

<b>Effect</b>	<b>df</b>	<b>F - value</b>	<b>P &gt; F</b>
Age	1	65	0.0001
Malaria zone	4	100.70	0.0001
Floor Material	2	7.57	0.0005
Nets per person	1	6.51	0.0108
Mother's Highest Education level	2	4.75	0.0087
Number of nets	1	0.87	0.3506
Type of place of residence	1	4.22	0.0400
Toilet facility	2	8.12	0.0003
Floor Material*No. of nets	2	13.99	0.0001

The age of the child in years, type of floor material used in household construction and type of toilet facility were once again the significant main effects. Other significant main effects were nets per person, mother's highest education level, malaria zone and type of place of residence. The only two-way interaction under consideration was between number of nets and floor material used in household construction.

Based on the results (see Table 4.4), we again observe that for a unit increase in the age of child in years, the odds for positive malaria results increases by 13%, [OR = 1.131, p-value = 0.0001, 95% C.I(1.098, 1.166)]. The odds for malaria infection decreased by 70.7% as the

number of mosquito nets per person sharing increased by a single unit [OR = 0.293, p-value = 0.0001, 95% C.I.(0.114, 0.752)]. Compared to households located in the seasonal risk zones, the odds for having malaria was greatest for households located in the lake regions [OR = 281.913, p-value = 0.0001, 95% C.I. = (86.718, 916.480)], followed by the households in moderate zones [OR = 12.643, p-value = 0.0001, 95% C.I.(3.826, 41.784)], and lastly the households located in the highland zones [OR = 8.232, p-value = 0.0006, 95% C.I.(2.463, 27.511)]. The risk for malaria also seemed to increase by 75% in households located in the rural areas compared to those within the urban areas [OR = 1.752, p-value = 0.040, 95% C.I.(1.026, 2.992)].

Compared to households with mother's with a primary education, children in households with mother's without an education had a 76% chance for positive malaria outcome [OR = 1.7596, p-value = 0.0079, 95% C.I.(1.1598, 2.6698)]. Toilet facility was once again a significant socio-economic factor in explaining the prevalence of malaria. In comparison to households with flush toilets, households without toilet facilities had a higher likelihood of positive malaria outcome in their children.

Table 4.4: Parameter estimates, odds ratio and 95 % confidence intervals for the final GLMM with household random effect

Variable	Estimate	Odds ratio	95% C. I	Standard error	P- value
<b>Intercept</b>	-6.8517			0.9964	0.0001
Age	0.1235	1.1315	(1.098,1.1659)	0.01532	0.0001
<b>Malaria zone ( Ref = Seasonal)</b>					
Lake region	5.6416	281.9134	(86.7179,916.4798)	0.6015	0.0001
Moderate risk	2.5371	12.6430	(3.8255,41.7836)	0.6099	0.0001
Highland	2.1080	8.2318	(2.4631,27.5107)	0.6156	0.0006
Low risk	0.3878	1.4737	(0.1500,14.4800)	1.1658	0.7394
Nets per Person	-1.2263	0.2934	(0.1145,0.7519)	0.4802	0.0107
<b>Mother's highest education ( Ref = Primary)</b>					
Higher	-0.2035	0.8159	(0.5986,1.1120)	0.1580	0.1980
No education	0.5651	1.7596	(1.1598,2.6698)	0.2127	0.0079
<b>Toilet facility( Ref = Toilet with flush)</b>					
Pit Latrine	0.5919	1.8074	(0.4075,8.0165)	0.7600	0.4362
No facility	1.2620	3.5325	(0.7834,15.9278)	0.7684	0.1006
<b>Type of place of residence ( Ref = Urban)</b>					
Rural	0.5668	1.7626	(1.0322,3.0098)	0.2730	0.0400

### Interaction terms

The two way interaction between number of nets in the household and type of floor material used in household construction was significant. The results are given in Table 4.5 below.

Table 4.5: Parameter estimates, odds ratio and 95 % confidence intervals for the final GLMM with household random effect

Variable	Estimate	Odds ratio	95% C. I	Standard error	P- value
Number of nets	-0.3060	0.7364	(0.5723,0.9475)	0.1286	0.0174
<b>Floor material( Ref = Earth/Sand)</b>					
Cement	-0.5508	0.5765	(0.2547,1.3049)	0.4168	0.1864
Dung	-1.1151	0.3279	(0.1858,0.5786)	0.2898	0.0001
<b>Floor material*Number of Nets( Ref = Earth/Sand)</b>					
Number of nets*Cement	0.050	1.0511	(0.7304,1.5126)	0.1857	0.7884
Number of nets*Dung	0.6273	1.8725	(1.4319,2.4489)	0.1369	0.0001

Figure 4.1 displays the pair comparison of least square means of the number of nets and floor material used in household construction interaction on malaria status.

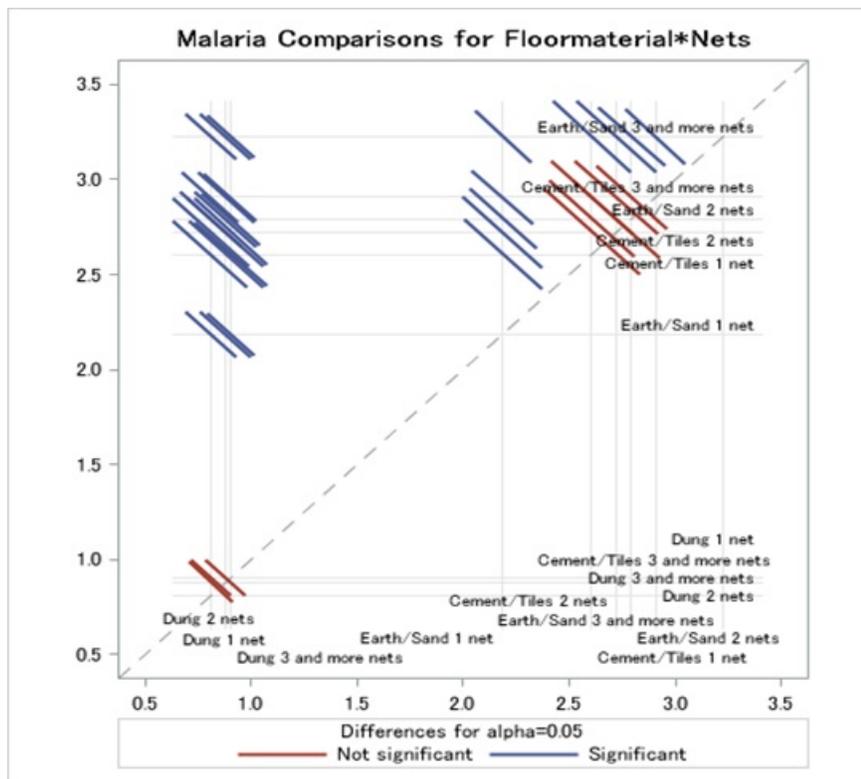


Figure 4.1: Diffogram of floor material and Number of nets interaction effects

The diffogram is useful in providing a visual representation of the differences in the least square means (LS-means) amongst the pairs and levels of classification variables. Both the variables number of nets and floor material used in household construction had three levels. The horizontal and vertical axes of the diffogram have similar lengths. The point of intersection of the line coming from the horizontal and the vertical axes gives the values of the differences in log scale of the LS-means being compared. The length of line segments on both sides of the crossing point shows the width of the confidence intervals for this differences. The  $45^\circ$  reference line determines the significance of a pair of classification. Usually, comparisons whose confidence interval entails zero, cross the reference line and are considered non-significant. Line segments that do not cross the reference line suggests significant LS-means difference.

From figure 4.1, the combination between dung floor with one net per household and dung floor with 2 nets per household; dung floor with one net per household and dung floor with 3 nets per household; dung floor with 2 nets per household and dung floor with 3 nets per household; cemented floor with one net per household and cemented floors with 2 nets per household; cemented floor with one net per household and cemented floors with 3 nets per household; cemented floor with one net per household and earthen floors with 2 nets per household; and cemented floor with 2 nets per household and cemented floors with 3 nets per household, had insignificant LS-means differences. The remaining interactions have statistically significant LS-mean differences at the 5% level of significance. Generally, an increase in number of mosquito nets in earthen floors had a significant effects in decline in malaria risk compared to the other floor materials.

The model with household random effect had a Pearson chi-square over the degrees of freedom statistic of 0.94 where a value of 1 is desirable. This value is close to 1 indicating

that variability in the data was properly modeled. The cluster effect may also be a source of random variation. There may be similarity in incidences of malaria amongst households from the same clusters. However, fitting the same model but with cluster as the random effect produced a less dispersed model with a Pearson chi square over the degrees of freedom statistic shifting from 0.94 to 0.60. The variance component estimate for the cluster effect is however significant at 1.1715, with a standard error of 0.1463, indicating a significant cluster to cluster variation. The parameter estimates are similar to those of the previous model but with slightly higher standard errors.

Further variability can also be induced through nesting the households within the clusters. The dataset included a selection of 30 households in each of the 240 randomly selected clusters. Attempting this approach, we observe that the model is much less dispersed, with a Pearson chi square over the degrees of freedom statistic of 0.36. The G-side random effects were still significant.

Table 4.6: Covariance parameter estimates for household nested within cluster random effects

Cov Parm	Subject	Estimate	Standard error
Intercept	Cluster	1.6542	0.4440
Intercept	Household(Cluster)	0.9852	0.2880

A comparison of the AIC for the model shown in Table 4.3, shows that the GLMM with two random effects had a lower AIC. The model with cluster and household nested within cluster random effects was a better fitting model than the two other models with separate household and cluster random effects.

Table 4.7: Comparison of the AIC estimates for the models

Model 1	Household random effect	2132.50
Model 2	Cluster random effect	2003.75
Model 3	Cluster and Household nested in cluster	1938.97

However the final model selected for the study was model 1 with household random effect. Model 2 and model 3 were less dispersed, and gave insignificant results for type of place of residence and toilet facilities, which are important covariates in explaining malaria.

#### 4.1.7 Summary and discussion

The results from the study show that demographic, socio-economic as well as geographic factors are important in explaining malaria infection in children. These factors are: age, housing structure particularly the type floor material used in household construction, type of toilet facility, nets per person sharing, number of nets in the household, malaria zone, type of place of residence and mother's highest education level.

We observe that the socio-economic factors which are actually poverty related factors played a vital role in determining the risk of malaria in children. Children whose families had no toilet facility, had relatively fewer mosquito nets and lived within a rural setting had higher chances for malaria infection. In addition, having better living conditions through decently structured homesteads reduces the risk for malaria. Homes that had cemented floors and dung floors had lower malaria prevalence than home with earthen floors. Having more mosquito nets within such families could lead to a reduction in malaria prevalence.

The geographical setup within which the household is located also contributed to observed pattern for malaria spread. Children living in the lake regions, highland malaria zones and moderate risk zones had high presence of malaria parasitamea in their blood samples. Geographical factors provide conducive environments for both the parasite and the mosquito to thrive and multiply. Hence encouraging such communities to effectively control the vector through anti-malarial spraying, use of insecticide treated bed nets and environmental hygiene

could help reduce malaria. The study supports this notion since we see that households that equivalently had more bed nets as the household members had lower risks for malaria.

Several multinational and multilateral programs and initiatives have been set-up and supported by the Kenyan government to help control and eventually eliminate malaria. Some of the key strategies employed include control measures such as early malaria diagnosis and treatment distribution of insecticide treated mosquito nets programs and anti-malarial spraying. The study findings have shown that some of these strategies have contributed to a significant decline in the risk of malaria. Despite these, more effort needs to be added especially in regions with high malaria cases. Intense vector control measures such as indoor malaria spraying and clearing of vegetation surrounding the homestead should be encouraged. The number of mosquito nets per household should be commensurate with the number of household members and individuals should use the provided mosquito nets for their intended purpose.

The government can also include poverty alleviation strategies in their social welfare programs for the communities. The study has shown that poverty is a big contributing factor towards the risk for malaria. Households that cannot afford better housing with toilet facilities had higher likelihood for malaria infections. Public awareness campaigns on the importance of proper sanitation and hygiene and prompt diagnosis and treatment could help the poorer households combat infectious diseases.

An important observation from the study is the fact that older children, with a perceived acquired immunity, seem to be more susceptible to malaria than the younger ones. This may be due to immunization plans facilitated by the government that distributes bed nets to children under 5 years during ante-natal visits. The government policy can also be shifted

to encourage distribution of insecticide treated mosquito nets commensurate to the number of household members.

This chapter has also highlighted that there is variability in malaria prevalence as a result of testing children from the same household. This may be due to similarity in living conditions that encourage malaria risk and also, one household may have more children testing for malaria than the next different household. This means that individual household care and interventions may be necessary, although they may be more costly than the population based interventions.

The next chapter introduces the generalized additive mixed effects model (GAMM), which enhances the GLMM framework by modelling non-linearity exhibited in some of the covariates.

# Chapter 5

## Semi parametric regression approach

### 5.1 Introduction

The methods discussed in the previous chapters; survey logistic regression (a special case of the GLM for modelling survey data) and the generalized mixed effects model (GLMM), are both parametric methods used for describing the relationship between the outcome of interest and the covariates. Although the parametric methods are computationally easy to use and interpret, they assume that the functional form of the model is known prior. As a consequence, the results may be biased, and therefore the need for nonparametric modelling that assumes an unknown functional form of the model, prior to modelling.

Nonparametric regression is a flexible approach for modelling nonlinear forms of data that have no predetermined functional forms. Suppose we have a pair of random variables  $\{X_i, Y_i\}$ ,  $i = 1, 2, \dots, n$ , the general form of the nonparametric model is given as;

$$Y_i = g(X_i) + \epsilon_i \tag{5.1}$$

where  $g$  is the unknown regression function of the predictor variable(s) and can be estimated by a roughness penalty method (Green and Silverman, 1994). Other methods of approximating unbiased and consistent estimators of the nonparametric regression model are: kernel estimators, smoothing splines, regression splines, running mean estimator, running line smoothers, bin smoothers, wavelets and locally weighted scatter plot smoothing (LOWESS). One such model is the project pursuit regression suggested by (Friedman and Stuetzle, 1981), that fits the model of the form,

$$Y = \sum_{j=1}^p s_j(\alpha'_j X) + \epsilon$$

where  $\alpha'_j X$  is a one dimensional projection of the vector  $X$ ,  $s_j$  is the arbitrary smooth function and the error term  $\epsilon$ , is an independent random variable with mean 0 and variance  $\sigma^2$ . These models are parsimonious smooth surface estimators but are difficult to interpret for larger  $p$  (Hastie and Tibshirani, 1990).

The alternating conditional expectation discussed by (Breiman and Friedman, 1985), is also another nonparametric approach for estimating nonlinear multiple regression. The model is given by:

$$E(\theta(Y)|X) = \sum_{j=1}^p s_j(x_j)$$

The response variable is estimated as a transformation of the form  $\theta(Y)$ . Extensive literature on the nonparametric regression has been done, among them: (Härdle, 1990), (Faraway, 2006), (Staniswalis and Lee, 1998), (Lin and Ying, 2001), (Izenman, 1991), (Silverman, 1985), (Buja et al., 1989).

Nonparametric model fits, however, suffer from a problem researchers refer to as "the curse of

dimensionality” in cases where the data is of high dimension. These often results in bias and unreliable interpretations of the fitted model. Semiparametric regression models have thus been developed, that combine properties of parametric regression and the nonparametric methods (Lin and Ying, 2001; Carota and Parmigiani, 2002; Zeger and Diggle, 1994; Härdle et al., 1998).

The semiparametric regression models are used to fit models with unknown functional forms and non-linear fits, examples of which can be found in works by (Engle et al., 1986) and (Green et al., 1985). This chapter utilizes this form of regression in particular the generalized additive mixed effects model (GAMM) in determining the risk factors associated with malaria in children. Section 5.2 of this chapter discusses the generalized additive model (GAM) that provides a background towards the generalized additive mixed model discussed in section 5.2.4. Our data is then fit using GAMM procedure, and a summary and discussion of the results follows thereafter.

## 5.2 Generalized additive model

The curse of dimensionality in nonparametric regression modelling led to development of semiparametric models that can fit data with outliers and non-linear covariates. The generalized additive model (GAM) is an example of the semiparametric regression models. It is a generalization of the GLM for modelling non-gaussian data and also an extension of the nonparametric additive model. It is therefore important to give a summary of the additive model.

### 5.2.1 Additive model

Given a sample of  $n$  data points,  $\{(X_i, Y_i) : i = 1, 2, \dots, n\}$ , the additive model (AM), suggested by Friedman and Stuetzle (1981) and developed by (Hastie and Tibshirani, 1990) is of the form:

$$Y_i = \alpha + \sum_{j=1}^p f_j(x_{ij}) + \epsilon_i \quad (5.2)$$

where  $Y_i$  is the outcome variable,  $x_{ij}$  are the predictor variables,  $\alpha$  is the intercept term,  $\epsilon_i \sim N(0, \sigma^2)$ , is the error component and  $f_j(\cdot)$  are the unknown smooth functions with no pre-specified functional form. The additive model, is a special type of the project pursuit regression model (Hastie and Tibshirani, 1987), that uses a one-dimensional smoothers in building the nonparametric multiple regression models. Some of the methods proposed for estimating the additive function are: the marginal integration estimation methods (Linton and Nielsen, 1995); the Fourier series approximation (Amato et al., 2002); the nonlinear wavelet estimation (Sardy and Tseng, 2004); and the backfitting algorithm (Buja et al., 1989). The backfitting algorithm is the most commonly used method and it enables one to fit the additive model using any of the regression fitting mechanisms (Hastie and Tibshirani, 1990).

The main concern in additive regression is obtaining an approximate estimate of the smooth functions in order to obtain the best fit model for our data. Therefore the estimate can be obtained through a process referred to as smoothing. Smoothing is a process of obtaining an approximation of the regression curve  $f_j(\cdot)$ , which is the mean of the response variables near the neighborhood of a point  $x_i$ . The process involves first identifying the smoothing technique and then secondly determining the smoothing parameter that controls the trade-off between under-smoothing and over-smoothing (Wood, 2011). The next section defines some of the smoothing techniques.

## 5.2.2 Smoothing Techniques

### Running mean smoother

The running mean smoother is also known as the moving average smoother and it estimates the smooth at point  $x_i$  by taking the average of the data points in a neighborhood  $N(x_i)$ , with  $n_i$  observations, around  $x_i$  (Buja et al., 1989).

$$f(x_i) = \sum_{i \in N(x_i)} \frac{y_i}{n_i} \quad (5.3)$$

Using a window of  $2k + 1$ , the running mean smoother assumes that the neighborhood of  $x_i$  is the symmetric nearest neighborhood such that

$$N_i = \max(i - k, 1), \dots, i - 1, i + 1, \dots, \min(i + k, n) \quad (5.4)$$

Running line smoothers are associated with functions that are wiggly and biased at the end points.

### Running line smoother

The bias problem in running mean smoothers can be solved through fitting the smooth curve to the data points through least squares, in a symmetric nearest neighborhood  $N_i$  around each  $x_i$ .

$$f(x_i) = \hat{\beta}_0 + \hat{\beta}_i x_i \quad (5.5)$$

where  $\hat{\beta}_0$  and  $\hat{\beta}_i$  are the ordinary least squares (OLS) estimates of  $x_i$ .

### Kernel smoothers

Suppose  $W_i(x)_{i=1}^n$  is a sequence of weights depending on the vector  $\{x_i\}_{i=1}^n$ , then the regression estimate of the smooth function  $f(x)$  is defined as:

$$\hat{f}(x) = n^{-1} \sum W_i(x) Y_i \quad (5.6)$$

The kernel smoother defines the shape of the weight function  $W_i(x)$  by a density function that has a scale parameter that adjusts the size and the form of the weights near  $x$ . This shape function is referred to as the kernel  $K$ , and it is a continuous, bounded, symmetric real function that integrates to unity.

The Nadaraya Watson kernel estimator of  $f(x)$  is defined as:

$$\hat{f}(x) = \frac{\sum_{i=1}^n k\left(\frac{x-x_i}{h}\right) y_i}{\sum_{i=1}^n k\left(\frac{x-x_i}{h}\right)}$$

where  $k$  is an even function that determines the shape of the kernel weights and  $h$ , the bandwidth, is used to parametrize the size of the weights. Generally, the kernel smoother matrix has the elements  $f_{ij} = c_i d_\lambda(x_i, x_j)$ , where  $d$  is the even function,  $\lambda$  is the bandwidth and  $c_i$  is a normalizing constant. To obtain the estimate of  $f(x_i)$ , the weights at the  $j^{th}$  point is assigned

$$w_{ij} = \frac{c_i}{\lambda} d\left(\frac{|x_i - x_j|}{\lambda}\right)$$

Some of the common choices of  $d(\cdot)$  are:

- Epanechnikov kernel described in (Härdle, 1990)

$$d(u) = 0.75(1 - u^2)I(|u| \leq 1)$$

- The Gaussian kernel (Buja et al., 1989)

$$d_\lambda(x_i, x_j) = \exp\left(-\left(\frac{x_i - x_j}{\lambda}\right)^2\right)$$

### Locally weighted running line smoother (LOESS)

Locally weighted regression, implemented by (Cleveland, 1979), makes use of both the running line smoother and kernel smoother methodologies to obtain the smooth. Computation of the smooth is through following these steps:

- Find a symmetric nearest neighborhood ( $N(x_i)$ ) of  $x_i$
- Calculate the distance to the  $k^{th}$  nearest neighbor, denote it by  $d_i$ .
- Assign a tri-cube weight function to each point in  $N(x_i)$ :

$$w_{ij} = \left(1 - \left|\frac{x_j - x_i}{d_i}\right|^3\right)^3$$

The estimate of the regression function  $f(x_i)$  is the fitted value at the point  $x_i$  obtained through fitting a weighted least square line.

Another method that can be used to estimate the smooth function  $f(x)$  is through the use of splines. A spline may be defined as a function that has been joined "piece-wise" from polynomial functions. An example is a sequence of knots defined by  $\zeta_1 \leq \zeta_2 \leq \dots \leq \zeta_k$ , that join smoothly at the knots. Smoothing splines are a flexible approach for estimating the regression curve using a piecewise polynomial in which the knots are the observed values of  $x_i$ . It is also subject to some smoothing constraints at the knots. Extensive literature on the common smoothing splines such as: the cubic smoothing spline, thin plate regression splines, p-splines and the B-spline can be found in works by Wood (2006); Hastie and Tibshirani (1986); Silverman (1985); Green and Silverman (1994); Buja et al. (1989); Wahba (1975). The natural cubic spline is the most common smoothing spline and hence the subject for

our discussion.

### Natural cubic spline smooth

Suppose we have a model of the form:

$$Y_i = g(x_i) + e_i \quad (5.7)$$

where  $\{(x_i, y_i)\}_{i=1}^n$  is a sequence of the response variables and predictor variables,  $e_i$  are the error terms such that  $e_i \sim N(0, v_i)$ . The regression function of  $y$  on  $x$  is defined by  $g(x_i)$ , a nonparametric smooth functions that needs to be estimated. The optimization problem in this setting is to minimize the penalized sum of squares (PSS) defined in equation 5.8

$$s_\lambda(g) = \sum_{i=1}^n (Y_i - g(x_i))^2 + \lambda \int_{-\infty}^{\infty} (g''(x))^2 dx \quad (5.8)$$

where  $\lambda$  is a smoothing parameter that controls the trade-off between the curve smoothness and proximity to the values of  $y$ . Notice that when  $\lambda = 0$ , the solution is an interpolating function and when  $\lambda = \infty$ , the solution is the standard least squares line. The PSS is minimized by a natural cubic spline with knots at each distinct  $x_i$ . The solution to  $s_\lambda(g)$  is a cubic polynomial whose derivatives are continuous at the boundary points say  $x_{(*)}$  and  $x_{(**)}$ , and the second derivative equates to 0.

### Regression splines

Regression splines use fewer knots than the natural cubic splines and they apply parametric regression to the bases functions. Suppose  $\zeta_1, \dots, \zeta_L$  are a set of knots and  $B_1(z), \dots, B_l(z)$

are a set of basis functions, then  $g(x)$  in equation 5.8 can be estimated by

$$g(x) \approx \sum_{i=1}^L B_i(z)\alpha_i \quad (5.9)$$

The vector  $(\alpha_l)^T$  can be estimated by fitting a parametric model via OLS, of the form

$$y_i = \sum_{i=1}^L B_l(z)\alpha_l + \epsilon_l$$

### P-splines

The regression function described in equation 5.7 above is approximated with more  $L$  knots than those used in regression splines, although these knots are smaller than the sample size. An advantage of p-splines is that they are less computer intensive especially for large sample sizes and also less sensitive to knot allocation.

In spline smoothing, it is important to choose the smoothing parameter well. Stone (1985) justifies the use of cross validation, which involves omitting one at a time the data points, and choosing  $\lambda$  at which the missing data points are best predicted by the remainder of the data (Silverman, 1985). Let  $g_\lambda^{-1}$  be the smoothing spline obtained from all data points excluding  $(x_i, y_i)$ , the cross validation choice of  $\lambda$  is the value that minimizes the following cross validation score:

$$cv(\lambda) = n^{-1} \sum_{i=1}^n \{Y_i - g_\lambda(x_i)\}^2 \quad (5.10)$$

(??) propose a suitable alternative to cross validation coined the generalized cross validation (GCV), that minimizes the average squared errors at design points  $x_1, \dots, x_n$ . It is defined as

$$GCV(\lambda) = \frac{n^{-1}RSS(\lambda)}{\{1 - n^{-1}trA(\lambda)\}^2} \quad (5.11)$$

where  $RSS(\lambda)$  is the residual sum of squares,  $\sum_{i=1}^n \{Y_i - g(\hat{x}_i)\}^2$ . The function  $n^{-1}trA(\lambda)$  is the average value of the matrix  $A(\lambda) = n^{-1}G(x_i, x_j)$ , where  $G(\cdot)$  is the weight function that depends on the design points and the smoothing parameter.

The details that follow are borrowed extensively from (Hastie and Tibshirani, 1990). Given any function  $f$  defined on the interval  $[a, b]$ , and a smoothing parameter  $\lambda$ , the optimization problem, is to minimize the equation:

$$\sum_{i=1}^n \left[ y_i - \sum_{j=1}^p f_j(x_{ij}) \right]^2 + \sum_{j=1}^p \lambda_j \int_a^b [f_j''(t)]^2 dt \quad (5.12)$$

If the solution function  $\hat{f}$ , is a cubic spline with knots in each  $x_i$ , we can obtain a smoothing matrix. Hence equation 5.12 may be rewritten as;

$$\left( \mathbf{y} - \sum_{j=1}^p \mathbf{f}_j \right)^T \left( \mathbf{y} - \sum_{j=1}^p \mathbf{f}_j \right) + \sum_{j=1}^p \lambda_j \mathbf{f}_j^T \mathbf{K}_j \mathbf{f}_j \quad (5.13)$$

The  $\mathbf{K}_j^s$  are the penalty matrices. Differentiating equation 5.13 with respect to  $\mathbf{f}_k$  and equating the resulting solution to 0, we obtain the estimating equation

$$\hat{\mathbf{f}}_k = \mathbf{S}_k (\mathbf{y} - \sum_{j \neq k} \hat{\mathbf{f}}_j)$$

where  $\mathbf{S}_k = (\mathbf{I} + \lambda_k \mathbf{K}_k)^{-1}$ , is the smoother matrix. The smoothing parameter  $\lambda$ , can be estimated through cross validation (Hastie and Tibshirani, 1990), or GCV (Wood, 2006; ?).

An  $np \times np$  system of equations can be obtained from equation ... for all  $k = 1, 2, \dots, p$ .

$$\begin{pmatrix} I & S_1 & \cdots & S_1 \\ S_2 & I & \cdots & S_2 \\ \vdots & \vdots & \ddots & \vdots \\ S_p & S_p & \cdots & I \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_p \end{pmatrix} = \begin{pmatrix} S_1 Y \\ S_2 Y \\ \vdots \\ S_p Y \end{pmatrix} \quad (5.14)$$

Notice that these equations can be written in a linear form:

$$\hat{p}f = \hat{Q}y$$

and the solution derived through iterative methods such as the backfitting algorithm (Friedman and Stuetzle, 1981; Hastie and Tibshirani, 1986). This algorithm works by going through each of the predictors, replacing each smooth function with its estimate while controlling for the effects of the others, and then going through the cycle till convergence (Buja et al., 1989). The steps can be summarized as follows:

- Assign the initial values of the intercept term with the average of  $y$ ,  $\alpha = E(y)$ , and the smoothing function as  $f_j = f_j^o$ ,  $j = 1, 2, \dots, p$ , where  $f_j^o$  is the estimate obtained from linear regression.
- Repeat this cycle  $j = 1, \dots, p, 1, \dots, p, \dots$  times until  $\hat{f}_j$  converges for each  $j^{th}$  predictor, such that  $\hat{f}_j = S_j(y - \mu - \sum_{k \neq j} f_k(x_k))$ . The matrix  $S$  is the smoother matrix defined above.

Additive models are rarely affected by the "curse of dimensionality" problem, are easy to interpret and also cost effective. (Hastie and Tibshirani, 1990; Buja et al., 1989).

### 5.2.3 Generalized additive models

The generalized additive models (GAM) (Hastie and Tibshirani, 1986, 1990), are an extension of the GLM (Nelder and Wedderburn, 1972) that includes an additive term in the linear predictor. Suppose that  $y_i$  is the response variable whose distribution is amongst one of the exponential family of distributions, then the GAM has a general form defined by:

$$g(\mu_i) = \eta_i = \mathbf{X}_i^* \boldsymbol{\theta} + \sum_j f_j(x_j) \quad (5.15)$$

where  $g(\mu)$  is the monotonic, invertible and a differentiable link function,  $\mathbf{X}_i^*$  is the  $i^{\text{th}}$  row of the model matrix  $X^*$ , whose parameter estimates defined by the vector  $\boldsymbol{\theta}$ , can be solved parametrically. The function  $f_j(\cdot)$  are the smooth functions of the covariates  $x_j$ .

Estimating the generalized additive model is a two-fold process involving the estimation of the smoothing parameters and also obtaining the model coefficients of the maximum penalized likelihood function. Choosing the basis function and the smoothing parameter is therefore central in GAM estimation. The most common choice of the basis are the penalized regression smoothers, that are based on smoothing splines discussed in the previous section. The smooth terms are now represented as a linear combination of the basis functions,  $b_{jk}$ , and the unknown regression parameters,  $\beta_{jk}$  such that:

$$f_j(x_j) = \sum_{k=1}^{q_k} \beta_{jk} b_{jk}(x_j)$$

Substituting each smooth term  $f_j(x_j)$ , by their bases, equation 5.15 can be written as:

$$g(\mu_i) = \eta_i = \mathbf{X}_i \boldsymbol{\beta} \quad (5.16)$$

where  $\mathbf{X}_i$  contains the columns of  $\mathbf{X}_i^*$  and the columns containing the spline bases evaluated at covariate values. The column vector  $\boldsymbol{\beta}$  contains  $\boldsymbol{\theta}^*$  and all the smooth coefficient vectors  $\beta$ . This equation is similar to a GLM that is fitted using iterative reweighted least squares (IRLS) procedure. Due to the additive structure of GAMs, the penalized likelihood function is maximized by penalized iterative least squares (P-IRLS). The optimization problem is thus minimizing equation 5.17 with respect to  $\beta$ :

$$\|\mathbf{W}^{[k]}(Z^{[k]} - \mathbf{X}\boldsymbol{\beta})\|^2 + \sum_j \lambda_j \beta^T S_j \beta \quad (5.17)$$

The constant  $k$  represents the iteration index, and  $\lambda_j$  is the smoothing parameter. The diagonal matrix of weights,  $\mathbf{W}$  has diagonal elements  $w_i^{[k]} = \omega_i^{\frac{1}{2}} \frac{V(\mu_i^{[k]})^{-\frac{1}{2}}}{g'(\mu_i^{[k]})}$  where  $V(\mu_i) = \phi^{-1} \text{var}(y_i)$ , and  $Z^{[k]} = X\boldsymbol{\beta}^{[k]} + G^{[k]}(y - \mu^{[k]})$  where  $G^{[k]}$  is a diagonal matrix such that the diagonal element,  $G_{rr}^{[k]} = g'(\mu_i^{[k]})$ . The P-IRLS steps can be summarized as follows:

- The initial weights  $w_i^{[k]}$  as defined above, and the pseudo data,  $z_i^{[k]} = g'(\mu_i^{[k]})(y_i - \mu_i^{[k]}) + \eta_i^{[k]}$  are estimated using the current model estimate  $\mu_i^{[0]} = E(y_i)$ .
- The next estimate  $\boldsymbol{\beta}^{k+1}$  is obtained by minimizing equation 5.17 with respect to  $\boldsymbol{\beta}$  and hence the next estimate of  $\eta_i^{[k+1]} = X\boldsymbol{\beta}^{[k+1]}$  and  $\mu_i^{[k+1]} = g^{-1}(\eta_i^{[k+1]})$ .
- After each iteration, we obtain new values of the coefficients  $\mu$  and  $\boldsymbol{\beta}$  and update the weights  $w_i$  and pseudo data  $z_i$ . Iteration is repeated until convergence .

Selection of the smoothing parameter is also paramount in GAM estimation and can minimize the GVC, the AIC or the Mallows's  $C_p$  commonly known as the unbiased risk estimation (UBRE) (Wood, 2006). The GVC score is given by:

$$V_g = \frac{nD(\hat{\boldsymbol{\beta}})}{[n - \text{tr}(A)]^2} \quad (5.18)$$

where  $A = \mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}^2\mathbf{X} + \mathbf{S})^{-1}$  is the influence matrix of the fitted model and  $D(\hat{\beta}) = 2\phi(l_{sat} - l_{fit})$ ,  $\phi$  is the dispersion parameter,  $l_{sat}$  is the maximum value of the log-likelihood of the model and  $l_{fit}$  is the log-likelihood of the fitted model. The UBRE score is defined as:

$$V_u = \frac{1}{n}D(\hat{\beta}) - \sigma^2 + \frac{2}{n}tr(A)\sigma^2 \quad (5.19)$$

The following AIC can also be minimized to obtain the value of  $\lambda_J$ :

$$V_a = D(\hat{\beta}) + 2tr(A)\phi \quad (5.20)$$

#### 5.2.4 Generalized additive mixed effects models

The generalized additive mixed effects model (Lin and Zhang, 1999) are an extension of the generalized linear mixed effects model (Breslow and Clayton, 1993), that are used to analyse data that extra variability as a result of correlation between and amongst observations. They include an additive function in the linear predictor in addition to the fixed and random effects. The general form of a GAMM is therefore:

$$g(\mu_i^b) = \eta_{ij}|b = \beta_0 + \sum_{j=1}^p f_j(x_{ij}) + \sum_{k=1}^q z_k b_k \quad (5.21)$$

Suppose that  $y_i$  are the independent response variable with conditional mean  $E(y_i|b) = \mu_i^b$  and variance,  $Var(y_i|b) = \phi v(\mu_i^b)$ , where  $v(\mu_i^b)$  is the variance function, and  $\phi$ , the dispersion parameter. Presume also that  $x_i = (1, x_{i1}, \dots, x_{ip})^T$  are the  $p \times 1$  vectors of the covariates associated with the fixed effects. Therefore, equation 5.21 can be described as follows;  $g(\cdot)$  is a monotonic, differentiable and invertible link function, with  $b = (b_1, \dots, b_k)^T$ , a  $q \times 1$  vector of the random effects,  $b \sim N(0, D(\theta))$ . The vector  $z_k = (z_1, \dots, z_k)^T$  are the design covariates associated with the random effects and  $f_j(\cdot)$ , are the centered twice differentiable smooth

functions.

In matrix notation, Lin and Zhang (1999) shows that equation 5.21 simplifies to

$$g(\mu^b) = 1\beta_0 + N_1f_1 + \dots + N_pf_p + \mathbf{Z}b \quad (5.22)$$

where  $1$  is an  $n \times 1$  vector of ones,  $N_j$  is an  $n \times r_j$  incidence matrix with the  $i^{th}$  component  $N_jf_j$  is defined by  $f_j(x_j)$ ,  $\mu^b = (\mu_1^b, \dots, \mu_n^b)^T$ ,  $g(\mu^b) = (\mu_1^b, \dots, \mu_n^b)^T$  and  $\mathbf{Z} = (z_1, \dots, z_n)^T$ .

Regression spline methodology allows the unknown smooth functions to be approximated by a set of basis functions as defined in equation 5.9. Hence, obtaining the fit of a GAMM, requires a choice of the basis function and estimation of the smoothing parameter. One also needs to make inference about the variance component  $\theta$ . The estimator of  $f_j(\cdot)$ , can be obtained using natural cubic smoothing splines discussed in section 5.2.2 above by maximizing the penalized log-likelihood function given by:

$$l(y_j; \beta_0, f_1(\cdot), \dots, f_p(\cdot), \theta) - \frac{1}{2} \sum_{j=1}^p \lambda_j f_j^T k_j f_j \quad (5.23)$$

where the roughness penalty of the penalized sum of squares  $\int_{s_j}^{t_j} [f_j''(x)]^2 dx$  can be estimated by  $f_j^T k_j f_j$ . The range of the  $j^{th}$  covariate is defined in the interval  $(s_j, t_j)$ . Maximizing this function requires numerical integration hence Laplace methods (Breslow and Clayton, 1993) discussed extensively by (Chen, 2000) can be used to obtain the MLE of  $\beta$ ,  $\phi$ ,  $f(\cdot)$  and  $\theta$ .

Lin and Zhang (1999) proposes maximizing the double penalized quasi likelihood function (D-PQL) w.r.t  $(\beta_0, f_1, \dots, f_p)$  to obtain the cubic spline estimators. The D-PQL is defined as:

$$l_{dpql} = -\frac{1}{2} \sum_{i=1}^n d_i(y_i; \mu_1^b) - \frac{1}{2} \sum_{i=1}^n b^T D^{-1} b - \frac{1}{2} \sum_{j=1}^r \lambda_j f_j^T k_j f_j \quad (5.24)$$

The first penalty term  $\sum_{i=1}^n b^T D^{-1} b$  is due to an approximation of the integrated log-quasi likelihood based on the Laplace method whereas the second penalty term  $\sum_{j=1}^r \lambda_j f_j^T k_j f_j$  determines the smoothness of the functions  $f_j(\cdot)$  that depends on the estimate of the smoothing parameter  $\lambda_j$ .

Since the centered parameter vector  $f_j$  can be re-parameterized in terms of the basis function such that  $f_j = X_j \beta_j + \beta_j a_j$ , the D-PQL becomes

$$l_{dpql}^* = -\frac{1}{2\phi} \sum_{i=1}^n d_i(y; \mu_1^b) - \frac{1}{2} b^T D^{-1} b - \frac{1}{2} a^T \Lambda^{-1} a \quad (5.25)$$

The vector  $X_j$  is an  $r_j \times 1$  vector with the  $r_j$  centered distinct values of  $x_{ij}$  while  $B_j = L_j(L_j L_j)^{-1}$ , with the  $r_j \times (r_j - 2)$  full rank matrix  $L_j$  satisfying the conditions  $L_j L_j^T$  and  $L_j^T X_j = 0$ . The identity  $f_j^T k_j f_j$  reduces to  $a_j^T a_j$  in the parametrized D-PQL, where  $a = (a_1^T, \dots, a_p^T)^T$ . The vector  $\Lambda = \text{diag}(\tau_1 I, \dots, \tau_p I)$  with  $\tau_j = \frac{1}{\lambda_j}$

Equation 5.22 can be generalized to:

$$g(\mu^b) = \mathbf{X}\boldsymbol{\beta} + \mathbf{B}a + \mathbf{Z}b \quad (5.26)$$

The vectors  $a^T$  and  $b^T$  represent the random effects and are both multi-normally distributed with mean equal to zero and variance given by  $\Lambda$  and  $D$  respectively. The vector  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$  is a  $((p+1) \times 1)$  vector of the model coefficients, while  $\mathbf{X} = (1, N_1 X_1, \dots, N_p X_p)^T$  and  $\mathbf{B} = (N_1 B_1, \dots, N_p B_p)$ . This model is simply a GLMM, and the D-PQL estimators of  $f_J$  can be obtained by fitting this model.

Maximising the DPQL with respect to  $\boldsymbol{\beta}$ ,  $a$  and  $b$ , we obtain the following normal equa-

tions:

$$\begin{pmatrix} X^T W X & X^T W B & X^T W Z \\ B^T W X & B^T W B + \Lambda^{-1} & B^T W Z \\ Z^T W X & Z^T W B & Z^T W Z + D^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ a \\ b \end{pmatrix} = \begin{pmatrix} X^T W Y \\ B^T W Y \\ Z^T W Y \end{pmatrix} \quad (5.27)$$

that can be solved through iterative methods such as the fisher scoring algorithm, to obtain the estimators  $\hat{f}_J$ , and the estimators of  $\hat{a}$ , and  $\hat{b}$ , the random effects.

The covariance matrix of  $\hat{f}_j$  is obtained from an approximation of the covariance matrix of  $\hat{\beta}$  and  $\hat{a}$ . The values of  $\hat{\beta}$  and  $\hat{a}$  are obtained by solving:

$$\begin{pmatrix} X^T R^{-1} X & X^T R^{-1} B \\ B^T R^{-1} X & B^T R^{-1} B + \Lambda^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ a \end{pmatrix} = \begin{pmatrix} X^T R^{-1} Y \\ B^T R^{-1} Y \end{pmatrix} \quad (5.28)$$

The smoothing parameter  $\lambda$ , and the variance of the unknown vector of fixed regression parameters  $\theta$  also need to be estimated. To ensure that  $f(\cdot)$  performs well, the choice of  $(\hat{\lambda})$  has to be good (Green and Silverman, 1994). The classical nonparametric method of obtaining  $\hat{\lambda}$  through cross validation can be used although it may be expensive and also hard to make inference on the variance components (Zeger and Diggle, 1994). Zhang et al. (1998) estimate  $\lambda$  and  $\theta$  jointly by using the restricted maximum likelihood (REML), and treat  $\tau = (\frac{1}{\lambda_1}, \dots, \frac{1}{\lambda_j})^T$  as another variance component. Conversely, Lin and Zhang (1999) obtain these estimates simultaneously by maximizing the marginal quasi likelihood, and also presume that  $\tau$  is an extra variance component.

### 5.3 Application of GAMM to the dataset

The GAMM is used to fit the dataset to determine the risk factors associated with malaria in children. The previous chapters applied the parametric methods of survey logistic regres-

sion and the GLMM. GAMM is an important model used to fit non-linear and non-normal forms of data. It combines both features of parametric and non-parametric regression hence the term, semi-parametric regression. The continuous predictor variables are fitted non-parametrically through additive regression while the remaining covariates are fitted linearly through parametric methods.

The response variable is malaria status in children, a binary variable, with a Bernoulli distribution. The associated covariates are: age in years; gender; malaria zone; type of place of residence; cluster altitude in metres; toilet facility; source of water; availability of electricity; ownership of bicycle; mobile phone and radio; type of wall material, floor material and roofing material used in household construction; use of mosquito nets; number of mosquito nets; nets per person; number of rooms per person; and anti malarial spraying.

The effects of age, number of rooms per person, number of nets, nets per person and cluster altitude in metres were modeled non-parametrically while the remaining covariates were modeled in a parametric way. Variable selection was achieved through applying both backward and forward selection procedures. We began with a full model containing all the covariates, eliminating the insignificant covariates (with a p-value  $\leq 0.05$ ) one at a time until we remained with only the significant variables. The forward selection procedure was then used to confirm the choice model. For comparison purposes, the final GLMM in the previous chapter was also fitted as a GAMM, but now with the continuous covariates modeled as additive functions (see Table 4.3).

Modelling was done using the *mgcv* package available in R statistical language using the *gam* formula. (Wood and Wood, 2007). The smooth terms are specified by expressions of the form  $s(X, k=" ", bs=" ")$ , where  $X$  is the covariate associated with the smooth function,

$k$  specifies the dimension of the basis and  $bs$  indicates the basis for the smooth term. The default option is "tp", thin plate regression smoother which is slow and uses a lot of memory for large data sets. The option utilized in this study was cubic spline regression, "cr". The family="binomial", in the formula specifies the distribution (binomial) and the link function (logit link) used to fit the model. The random effect variable is specified through the "random=option" in the formula statement.

The GAMM is therefore given as:

$$g(\mu_{ij}) = \beta_0 + \beta_1(\text{Malaria zone}) + \beta_2(\text{Toilet}) + \beta_3(\text{Residence}) + \beta_4(\text{Education}) + s_1(\text{Age}) + s_2(\text{Altitude}) + \epsilon_{ij} \quad (5.29)$$

the logit link function is defined by  $g(\cdot)$ ,  $s(\cdot)$  represents the centered smooth function and  $\beta_{j's}$  are the regression coefficients for the parametric terms. The random effects [ $b_i \sim N(0, D(\theta))$ ] was measured by the HouseID variable from the dataset. The variable identified the household number from which each child under fourteen was tested for malaria. The model failed to converge for random effect measured by the cluster number and even houseID nested within the cluster effect. This may be as a result of the complexity due to non-linear functions in the linear predictor and dimensionality problems. The total number of rooms per person, number of nets per person, and total number of nets in the household had an effective degrees of freedom (EDF) of one signifying a linear relationship with the response variable. They were then modeled non-parametrically but were dropped from the final model since they were insignificant.

The model based on the GLMM results converged upon removal of the interaction terms. This model (labeled model 2) consisted of effects of age and nets per person modeled non-

parametrically, and the effects of toilet facility, malaria zone, mother's highest education, floor material, number of nets and type of place of residence modeled parametrically. Table 5.1 and 5.2, gives the anova results for the two models:

Table 5.1: Anova results for model 1

<b>Parameter</b>	<b>df</b>	<b>F - value</b>	<b>p-value</b>
Age	3.357	25.728	0.0001
Altitude	6.403	6.851	0.0001
Toilet facility	2	14.543	0.0001
Malaria zone	4	127.792	0.0001
Mother's Highest Education level	2	8.335	0.0002
Type of place of residence	1	14.360	0.0001

Table 5.2: Anova results for model 2

<b>Parameter</b>	<b>df</b>	<b>F - value</b>	<b>p-value</b>
Age	3.335	26.850	0.0001
Nets per person	1.000	8.728	0.0031
Toilet facility	2	8.962	0.0001
Malaria zone	4	120.227	0.0001
Mother's Highest Education level	2	6.310	0.0018
Type of place of residence	1	15.903	0.0001

The first model had a better fit with an adjusted R-square statistic of 0.324. The results of anova tests on both models also revealed that terms of the first model had a more significant fit. The variable nets per person sharing in the second model shows that it has a linear effect (EDF=1). Table 5.3 represents the significant parametric coefficients for the model.

Table 5.3: Parameter estimates, odds ratio and 95 % Confidence intervals for the parametric covariate in the final GAMM

Variable	Estimate	Odds ratio	95% C. I	Standard error	P- value
<b>Intercept</b>	-4.5501			0.6563	0.0001
<b>Toilet facility( Ref = Toilet with flush)</b>					
Pit Latrine	0.4782	1.6132	(0.477,5.460)	0.6221	0.4421
No facility	1.2112	3.3575	(0.980,11.508)	0.6285	0.0540
<b>Malaria zone ( Ref = Highland)</b>					
Lake region	3.2028	24.6013	(17.654,34.282)	0.1693	0.0001
Moderate risk	-0.1160	0.8905	(0.414,1.916)	0.3910	0.7666
Seasonal risk	-2.6121	0.0734	(0.029,0.185)	0.4716	0.0001
Low risk	-1.3260	0.2655	(0.093,0.761)	0.5372	0.0136
<b>Type of place of residence ( Ref = Urban)</b>					
Rural	0.8314	2.2965	(1.494,3.530)	0.2194	0.0002
<b>Mother's highest education ( Ref = No Education)</b>					
Primary	-0.5144	0.5979	(0.433,0.826)	0.1648	0.0018
Higher	-0.8203	0.4403	(0.297,0.653)	0.2011	0.0001

The results show that toilet facility, type of place of residence, malaria zone and mother's highest education level are significant in explaining malaria risk in children. The age of the child and cluster altitude in metres were modeled non-parametrically using smoothing spline regression and were also significant in explaining malaria in children. The results for these smooth terms are displayed in Table 5.4.

Compared to households that had toilets with flush system, those with pit latrines and no facility had higher odds for having children test positive for malaria at 1.6132 ( $e^{0.4782}$ ) and 3.3575 ( $e^{1.2112}$ ) respectively. Respondents living in the lake regions had extremely high odds for malaria infection compared to those living in the highland areas (24.6013,  $e^{3.2028}$ ). Households in the moderate zones were 0.8905 ( $e^{-0.1160}$ ) times less likely to have malaria compared to those in the highland regions. The results are similar for seasonal and low risk regions with odds of 0.0734 ( $e^{-2.6121}$ ) and 0.2655 ( $e^{-1.3260}$ ) respectively.

Households within the rural settings had higher odds for malaria infection compared to households within the urban set up (2.2965). The education level of mother's also had an influence on the risk for malaria. Children whose mother's had attained a primary education and/or had a higher education had decreased odd for malaria (0.5979 and 0.4403 respectively) compared to households with mother's with no education at all. Both age and cluster

Table 5.4: Approximate significance of the smooth terms for GAMM

<b>Variable</b>	<b>EDF*</b>	<b>F - Value</b>	<b>Pr(&gt;t)</b>
s(Age)	3.357	25.728	0.0001
s(Altitude)	6.403	6.851	0.0001

\* Effective degrees of freedom

altitude in metres were significant non-linear effects. Figure 5.1 represents the smooth terms for these effects.

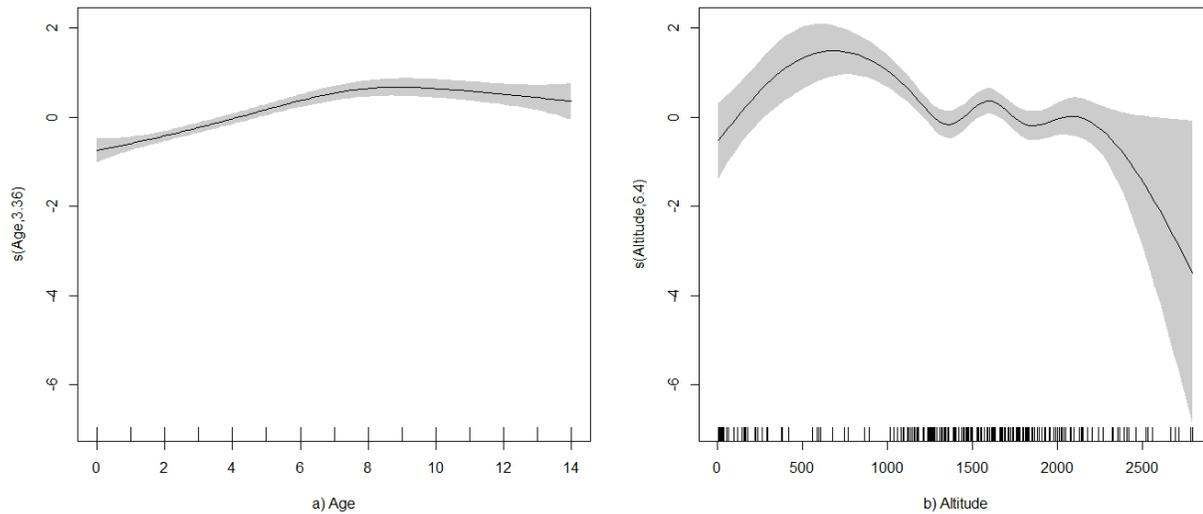


Figure 5.1: Smoothing components for malaria status with age and cluster altitude

Non-linearity in GAMMs is measured using the effective degrees of freedom (EDF) of the smoothing terms. The y-axis in figure 5.1 shows the effect of each smooth term  $s()$  and the value in the parentheses shows the EDF. The shaded region in the figure gives the 95% confidence intervals for the smooth terms. The EDF's for age and cluster altitude in metres are: 3.357 and 6.403 respectively. An  $EDF = 1$  signifies that the relationship between the covariates and the response is linear. Age has a quadratic effect, the risk for malaria increases steadily with increasing age in years in children, and reaches its peak at age seven. It's F-value is given by 25.728 with a p-value of 0.0001 suggesting with a 95% level of significance that its not a linear effect as assumed in the GLMM (see Table 4.3) and even survey logistic regression model (see Table 3.3).

From the figure, we again see that altitude is also non-linearly related with malaria in children. At lower altitudes, the risk of malaria is higher, increasing with an increase in altitude. It reaches its peak between 600m to 700m in altitude and steadily declines as

the altitude increases. Moreover its F-value = 6.851 with a p-value = 0.0001 suggests that cluster altitude in metres is not linearly associated with malaria status in children.

### 5.3.1 Summary and discussion

In this section, data was analyzed using the generalized additive mixed effects model (GAMM). The effects of the continuous covariates, age and cluster altitude in metres were analyzed non-parametrically while toilet facility, malaria zone, type of place of residence and mother's highest education level were modeled using parametric methods. The findings from this analysis support the results from the previous models. We see that age is an important demographic factor that affects malaria incidence. Older children seem to be more susceptible to malaria infections than the younger ones.

Geographical factors also contribute to an increase in malaria parasitamea in children. Living in lower altitudes and regions that provide conducive environments for the parasite to thrive such as lake region leads to higher chances of malaria infections. The socio-economic factors also play an important role in determining whether a child has malaria or not. Poorer households with no toilet facilities are more likely to be affected by malaria than well to do homesteads with toilets. Households in rural areas were also highly affected by malaria than urban households. The education level of mother seems to provide a protective cover to their children against malaria infections.

The results also continue to highlight on the importance of social welfare programs. There is a need for the government as well as the various stakeholders to continue to develop policies and programs that provide education to communities on malaria control measures such as the use of insecticide treated mosquito nets, antimalarial spraying, and early diagnosis and

treatment procedures.

# Chapter 6

## Discussion and conclusion

### 6.0.2 Discussion

The main objective of the thesis was to determine the prevalence and risk factors associated with malaria in children under the age of fourteen years in Kenya. In order to achieve this objective, two broad statistical analyses techniques were applied to the data. These were the parametric methods that included survey logistic regression and generalized linear mixed effects models (GLMM). The semi-parametric regression method used for analysis was the generalized additive mixed effects model (GAMM).

The data used in this study was obtained from a household survey conducted by the Kenyan government in the year 2010. The response variable was malaria status in children indicating whether they tested positive or negative for malaria. The associated demographic, geographical and socio-economic risk factors were: age of child in years, gender, cluster altitude in metres, malaria zone, type of place of residence, type of wall, floor, and roofing material used in household construction, toilet facility, water source, number of rooms per person sharing, wealth quantile, availability of electricity, ownership of mobile phone, bicycle, radio

and television, number of nets in the household, use of mosquito nets while sleeping, number of nets per person sharing and incidence of anti-malarial spraying.

A detailed discussion for each of the methods used was provided in the previous chapters, hence only a summary of the main contributions, limitations and recommendations for future research will be provided. In order to make valid statistical inference, it was important to account for the survey design such as stratification, clustering, non-response, and use of probability weights in the analysis. Hence the use of survey logistic regression (SLR) model, which is a special type of generalized linear model (GLM) for analyzing survey data with a binary outcome. To cater for possible correlations between observations from the same households and clusters, the generalized linear mixed effects model (GLMM) was used. In addition to the fixed effects, it also includes random effects in its linear predictor to account for such variabilities. The SLR model and the GLMM are both parametric methods that assume a linear relationship between the predictors and the outcome variables. This may not be the case for some predictor variables that may exhibit non-linearity. Hence, the semi-parametric generalized additive mixed effects model was also applied to the data.

The three methods used in the analysis of the data may be different with inherent strengths and weaknesses, but provide similar findings. Generally, age of the child, cluster altitude in metres, malaria zone, toilet facility, type of floor material and wall material used in house construction, ownership of bicycle, and mother's highest education level are significant as direct or indirect risk factors of malaria. The risk for malaria infection was higher in older children aged between 5 years and 8 years than the younger children. The findings favor the results from other studies showing malaria infection is not only persistent to children in <5 age groups, but also older age groups (Peterson et al., 2009; Ghebreyesus et al., 2000; Deressa et al., 2007). The geographical factors of altitude and malaria zone ultimately predetermine

the risk of malaria. The highly malaria endemic regions, mostly located in lower altitude zones and characterized by high rainfall with moderate temperatures provide conducive environments for parasite development and mosquito survival. Malaria infection was common in children at lower altitudes regions, peaking at 600 – 700 metres in altitude regions, and steadily declined thereafter. Many studies have also found malaria prevalence to increase with a decrease in altitude (Akhwale et al., 2004).

Evidently, malaria is related to poor socio-economic factors. Households that had poor housing structures, located in rural settings and lacked toilet facilities had higher chances of positive malaria tests. This was consistent with findings from literature ((Ayele et al., 2012, 2013; Gahutu et al., 2011; Peterson et al., 2009; Njau et al., 2014)). Poorly constructed households provide habitable environments for the mosquito vector to feed, breed and also rest. Interestingly, ownership of a bicycle, a symbol of wealth in most communities, did not indicate lower odds for malaria in such households. These factors highlight on the role and contribution of poverty in the malaria epidemic. However, households that incorporated control measures such as number of nets per person and total number of nets in the household had lower risks for malaria. The mother's education level also seemed to provide a protective cover to children. Households with mother's with no education had positive malaria results than those with primary and higher education levels. The mothers with at least a basic education are perhaps more aware of malaria infection, its diagnosis, treatment procedures and are ore likely to take up any of the intervention methods such as the use of mosquito bed nets.

### 6.0.3 Conclusion

Poverty continues to portray its influence on health. From the study, the more affluent families living in urban settings, with better housing structures and with basic amenities such as toilet facilities, had lower risks of malaria. These factors have highlighted on the importance of social policies that advocate for awareness of malaria, its control measures and alleviation of poverty. Several efforts have already been put in place by the Kenyan government and donor organizations, to help control and eventually eliminate malaria. In particular the study has shown that through proper use of insecticide treated mosquito nets, malaria prevalence can be reduced drastically. This is highlighted in literature too (Atieli et al., 2011; Nyarango et al., 2006). Other control measures such as indoor residual spraying, clearing of vegetation and bushes around the homestead and overall hygienic practices can also be implemented. These control measures can particularly be focused on rural communities, and highly malarious regions. The conclusion would therefore be for the Kenyan government to improve their health policies to focus on regional development, and improving the socio-welfare of communities. This can be achieved through routine education on malaria control measures.

The government can also focus on rural development, and encourage social-welfare programs that aim at providing both formal and informal education to the poor communities. This may in turn improve the socio-economic status of most families and also influence behavioral change towards the use of malaria control measures. Policies should also focus on encouraging more women and girls in particular, to gain formal education. The study has shown that maternal education has a huge impact on malaria control. Malaria intervention methods and control programs can also focus on individual household care in addition to the broad population based programs. This may be costly but the study has indicated a

variation in malaria prevalence due to household effect.

One of the shortcomings of this study was that some of the categorical variables had small sample sizes across the different levels, which may have produced insignificant results. Closely associated with this, is the issue of missing data. The assumption that data was missing at random, may have lead to biased parameter estimates. Hence, a future research area would be to consider statistical techniques that handle missing data.

The different statistical methods applied to analyze complex survey data with a binary outcome have highlighted the risk factors for malaria in children under the age of fourteen years. The same methods can also be advanced to all age groups in the population to highlight risk factors for each age group and the vulnerable groups as well. An important future research area would be to spatially model the geographical regions to identify the malaria patterns for each zone. Malaria has been seen to be more prevalent in some zones such as the lake regions, hence malaria mapping would help identify the variation in malaria risk. This would then help the government formulate policies for each zone and fairly allocate resources.

# Appendix A

## SAS and R codes for the models

**SAS code for analyzing the survey logistic regression model:**

```
Proc surveylogistic data=malaria data;  
Strata District;  
Cluster Cluster;  
Weight Weight;  
Class Cluster, X5, X6, X7, X8, X11;  
Model Y(Ref=FIRST) = X1, X2, X5, X6, X7, X8, X11, X6 * X8;  
Run;
```

**SAS code for analyzing the final GLMM with household random effect**

```
Proc glimmix data=malaria data method = quadrature(20points);  
Class Household, X3, X4, X5, X6, X8, X9, X10;  
Model Y(Ref=First) = X1, X3, X4, X5, X6, X8, X9, X10, X6 * X9 / link=logit dist = binary  
oddsratiosolution;
```

```
Random intercept /subject=Household type=chol;
```

```
Covtest zerog;
```

```
Run;
```

### **R code for analyzing final GAMM**

```
Gamm.Model <- gamm (Y ~ S(X1, bs="cr") + S(X2, bs="cr")+X3 + X4 + X5 + X8  
, family = binomial(link=logit), data = malaria data)
```

where:

$Y$  = Malaria status

$X_1$  = Age of child

$X_2$  = Cluster altitude in metres

$X_3$  = Malaria zone

$X_4$  = Type of place of residence

$X_5$  = Toilet facility

$X_6$  = Floor material

$X_7$  = Wall material

$X_8$  = Mother's highest education level

$X_9$  = Number of nets

$X_{10}$  = Nets per person

$X_{11}$  = Ownership of bicycle

# Bibliography

- Agresti, A. (1990). *Categorical Data Analysis*. John Wiley and Sons, Inc., United States of America.
- Agresti, A. (2002). *Categorical Data Analysis*. John Wiley and Sons, Inc., Hoboken, New Jersey, second edition.
- Agresti, A., Booth\*, J. G., Hobert\*, J. P., and Caffo\*, B. (2000). Random-effects modeling of categorical response data. *Sociological Methodology*, 30(1):27–80.
- Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723.
- Akhwale, W. S., Lum, J., Kaneko, A., Eto, H., Obonyo, C., Björkman, A., and Kobayakawa, T. (2004). Anemia and malaria at different altitudes in the western highlands of Kenya. *Acta Tropica*, 91(2):167–175.
- Amato, U., Antoniadis, A., and Feis, I. D. (2002). Fourier series approximation of separable models. *Journal of Computational and Applied Mathematics*, 146(2):459–479.
- Antonio, K. and Beirlant, J. (2007). Actuarial statistics with generalized linear mixed models. *Insurance: Mathematics and Economics*, 40(1):58–76.
- Archer, K. J., Lemeshow, S., and Hosmer, D. W. (2007). Goodness-of-fit tests for logistic

- regression models when data are collected using a complex sampling design. *Computational Statistics and Data Analysis*, 51(9):4450–4464.
- Atieli, H., Menya, D., Githeko, A., and Scott, T. (2009). House design modifications reduce indoor resting malaria vector densities in rice irrigation scheme area in Western Kenya. *Malaria Journal*, 8(1):1–9.
- Atieli, H. E., Zhou, G., Afrane, Y., Lee, M.-C., Mwanzo, I., Githeko, A. K., and Yan, G. (2011). Insecticide-treated net (itn) ownership, usage, and malaria transmission in the highlands of Western Kenya. *Parasit Vectors*, 4(1):113.
- Ayele, D. G., Zewotir, T. T., and Mwambi, H. G. (2012). Prevalence and risk factors of malaria in Ethiopia. *Malar J*, 11(195):10–1186.
- Ayele, D. G., Zewotir, T. T., and Mwambi, H. G. (2013). The risk factor indicators of malaria in Ethiopia. *International Journal*, 5(7):335–347.
- Beck-Johnson, L. M., Nelson, W. A., Paaijmans, K. P., Read, A. F., Thomas, M. B., and Bjørnstad, O. N. (2013). The effect of temperature on *Anopheles* mosquito population dynamics and the potential for malaria transmission. *PLoS ONE*, 8(11):e79276.
- Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review/Revue Internationale de Statistique*, pages 279–292.
- Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Steven, M. H. H., and White, J.-S. S. (2008). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and evolution*, 24(3):pp. 127–135.
- Bray, R. and Garnham, P. (1982). The life-cycle of primate malaria parasites. *British medical bulletin*, 38(2):117–122.

- Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391):pp. 580–598.
- Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):pp. 9–25.
- Breslow, N. E. and Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, 82(1):pp. 81–91.
- Brooker, S., Clarke, S., Njagi, J. K., Polack, S., Mugo, B., Estambale, B., Muchiri, E., Magnussen, P., and Cox, J. (2004). Spatial clustering of malaria and associated risk factors during an epidemic in a highland area of Western Kenya. *Tropical Medicine & International Health*, 9(7):757–766.
- Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, 17(2):pp. 453–510.
- Bursac, Z., Gauss, H. C., Williams, D. K., and Hosmer, D. W. (2008). Purposeful selection of variables in logistic regression. *Source code for Biology and Medicine*, 3(1):17.
- Caldwell, J. and McDonald, P. (1982). Influence of maternal education on infant and child mortality: levels and causes. *Health policy and education*, 2(3-4):251–267.
- Capanu, M., Gönen, M., and Begg, C. B. (2013). An assessment of estimation methods for generalized linear mixed models with binary outcomes. *Statistics in Medicine*, 32(26):4550–4566.
- Carneiro, I., Roca-Feltrer, A., Griffin, J. T., Smith, L., Tanner, M., Schellenberg, J. A., Greenwood, B., and Schellenberg, D. (2010). Age-patterns of malaria vary with severity,

- transmission intensity and seasonality in sub-Saharan Africa: a systematic review and pooled analysis. *PLoS One*, 5(2):e8988.
- Carota, C. and Parmigiani, G. (2002). Semiparametric regression for count data. *Biometrika*, 89(2):265–281.
- Carter, R. and Mendis, K. N. (2002). Evolutionary and historical aspects of the burden of malaria. *Clinical microbiology reviews*, 15(4):564–594.
- Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46(3):pp. 167–174.
- Ceesay, S. J., Casals-Pascual, C., Erskine, J., Anya, S. E., Duah, N. O., Fulford, A. J., Sesay, S. S., Abubakar, I., Dunyo, S., Sey, O., Palmer, A., Fofana, M., Corrah, T., Bojang, K. A., Whittle, H. C., Greenwood, B. M., and Conway, D. J. (2008). Changes in malaria indices between 1999 and 2007 in The Gambia: a retrospective analysis. *The Lancet*, 372(9649):1545–1554.
- Chaves, L. F., Satake, A., Hashizume, M., and Minakawa, N. (2012). Indian ocean dipole and rainfall drive a moran effect in East Africa malaria transmission. *Journal of Infectious Diseases*, 205(12):1885–1891.
- Chen, C. (2000). Generalized additive mixed models. *Communications in Statistics - Theory and Methods*, 29(5-6):1257–1271.
- Chen, M. H. and Dipak, K. D. (2003). Variable selection for multivariate logistic regression models. *Journal of Statistical planning and Inference*, 111(1):37–55.
- Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):pp. 327–335.

- Chirebvu, E., Chimbari, M. J., and Ngwenya, B. N. (2014). Assessment of risk factors associated with malaria transmission in Tubu village, northern Botswana. *Malaria research and treatment*, 2014.
- Clark, T. D., Greenhouse, B., Njama-Meya, D., Nzarubara, B., Maiteki-Sebuguzi, C., Staedke, S. G., Seto, E., Kanya, M. R., Rosenthal, P. J., and Dorsey, G. (2008). Factors determining the heterogeneity of malaria incidence in children in Kampala, Uganda. *Journal of Infectious Diseases*, 198(3):393–400.
- Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.
- Cochran, W. G. (1964). *Sampling techniques*. John Wiley and sons inc, New York.
- Cox, F. E. (2010). History of the discovery of the malaria parasites and their vectors. *Parasit Vectors*, 3(1):5.
- de Castro, M. C. and Fisher, M. G. (2012). Is malaria illness among young children a cause or a consequence of low socioeconomic status? Evidence from the United Republic of Tanzania. *Malaria journal*, 11(1):1.
- Deressa, W., Ali, A., and Berhane, Y. (2007). Household and socioeconomic factors associated with childhood febrile illnesses and treatment seeking behaviour in an area of epidemic malaria in rural Ethiopia. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 101(9):939 – 947.
- Division of Malaria control (2009). *National malaria strategy 2009 - 2017*. Division of public health, Ministry of public health and Sanitation, Nairobi, Kenya.
- Division of Malaria control [Ministry of public health and sanitation], Kenya National Bureau

- of Statistics, and ICF Macro (2011). *2010 Kenya Malaria Indicator survey*. DOMC, KNBS, and ICF Macro, Nairobi, Kenya.
- Dobson, A. J. (1990). *An introduction to Generalized Linear Models*. Chapman and Hall, New South Wales, Australia.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7(1):1–26.
- Efron, B. (1980). The jackknife, bootstrap and other resampling plans. Technical report, Stanford University.
- Engle, R. F., Granger, C. W. J., Rice, J., and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*, 81(394):pp. 310–320.
- Ernst, K. C., Adoka, S. O., Kowuor, D. O., Wilson, M. L., and John, C. C. (2006). Malaria hotspot areas in a highland Kenya site are consistent in epidemic and non-epidemic years and are associated with ecological factors. *Malaria Journal*, 5(1):78.
- Ernst, K. C., Lindblade, K. A., Koech, D., Sumba, P. O., Kuwuor, D. O., John, C. C., and Wilson, M. L. (2009). Environmental, socio-demographic and behavioural determinants of malaria risk in the Western Kenyan highlands: a case–control study. *Tropical Medicine and International Health*, 14(10):1258–1265.
- Faraway, J. J. (2006). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall/CRC, USA.
- Fortin, A., Stevenson, M., and Gros, P. (2002). Susceptibility to malaria as a complex trait: big pressure from a tiny creature. *Human Molecular Genetics*, 11(20):2469–2478.
- Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, 76(376):pp. 817–823.

- Gad, A. M. and El Kholy, R. B. (2012). Generalized linear mixed models for longitudinal data. *International Journal of Probability and Statistics*, 1(3):67–73.
- Gahutu, J.-B., Steininger, C., Shyirambere, C., Zeile, I., Cwinya-Ay, N., Danquah, I., Larsen, C. H., Eggelte, T. A., Uwimana, A., Karema, C., et al. (2011). Prevalence and risk factors of malaria among children in southern highland Rwanda.
- Gallup, J. and Sachs, J. (2001). The economic burden of malaria. *The American Journal of Tropical Medicine and Hygiene*, 64(1 suppl):85–96.
- Gamage-Mendis, A. C., Carter, R., Mendis, C., De Zoysa, A., Herath, P., and Mendis, K. N. (1991). Clustering of malaria infections within an endemic population: risk of malaria associated with the type of housing construction. *The American journal of tropical medicine and hygiene*, 45(1):77–85.
- Gemperli, A. (2013). *Development of Spatial statistical methods for modelling point-referenced spatial Data in malaria epidemiology*. PhD thesis, University of Basel.
- Ghebreyesus, T. A., Haile, M., Witten, K. H., Getachew, A., Yohannes, M., Lindsay, S. W., and Byass, P. (2000). Household risk factors for malaria among children in the Ethiopian highlands. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 94(1):17–21.
- Ghosh, A., Edwards, M., and Jacobs-Lorena, M. (2000). The journey of the malaria parasite in the mosquito: Hopes for the new century. *Parasitology Today*, 16(5):196–201.
- Gilles, H. (1981). The six diseases of who. malaria. *BMJ*, 283(6303):1382–1385.
- Gimnig, J. E., Otieno, P., Were, V., Marwanga, D., Abong’o, D., Wiegand, R., Williamson, J., Wolkon, A., Zhou, Y., Bayoh, M. N., et al. (2016). The effect of indoor residual spraying on the prevalence of malaria parasite infection, clinical malaria and anemia in

- an area of perennial transmission and moderate coverage of insecticide treated nets in Western Kenya. *PloS one*, 11(1).
- Githeko, A. K., Ayisi, J. M., Odada, P. K., Atieli, F. K., Ndenga, B. A., Githure, J. I., and Yan, G. (2006). Topography and malaria transmission heterogeneity in Western Kenya highlands: prospects for focal vector control. *Malaria Journal*, 5(1):1–9.
- Graves, P. M., Richards, F. O., Ngondi, J., Emerson, P. M., Shargie, E. B., Endeshaw, T., Ceccato, P., Ejigsemahu, Y., Mosher, A. W., Hailemariam, A., Zerihun, M., Teferi, T., Ayele, B., Mesele, A., Yohannes, G., Tilahun, A., and Gebre, T. (2009). Individual, household and environmental risk factors for malaria infection in Amhara, Oromia and {SNNP} regions of Ethiopia. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 103(12):1211–1220.
- Green, P., Jennison, C., and Seheult, A. (1985). Analysis of field experiments by least squares smoothing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 47(2):pp. 299–315.
- Green, P. and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall, USA.
- Greenwood, B., Bradley, A., Greenwood, A., Byass, P., Jammeh, K., Marsh, K., Tulloch, S., Oldfield, F., and Hayes, R. (1987). Mortality and morbidity from malaria among children in a rural area of The Gambia, West Africa. *Transactions of The Royal Society of Tropical Medicine and Hygiene*, 81(3):478–486.
- Greenwood, B. M., Bojang, K., Whitty, C. J., and Targett, G. A. (2005). Malaria. *The Lancet*, 365(9469):1487–1498.

- Gross, S. (1980). Median estimation in sample surveys. In *Proceedings of the Section on Survey Research Methods*, volume 1814184. American Statistical Association Ithaca, NY.
- Gurka, M. J., Edwards, L. J., and Muller, K. E. (2011). Avoiding bias in mixed model inference for fixed effects. *Statistics in medicine*, 30(22):2696–2707.
- Haque, U., Sunahara, T., Hashizume, M., Shields, T., Yamamoto, T., Haque, R., and Glass, G. E. (2011). Malaria prevalence, risk factors and spatial distribution in a hilly forest area of Bangladesh. *PLoS One*, 6(4):e18908.
- Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1(3):pp. 297–310.
- Hastie, T. and Tibshirani, R. (1987). Generalized additive models: Some applications. *Journal of the American Statistical Association*, 82(398):pp. 371–386.
- Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall/CRC, USA.
- Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):pp. 97–109.
- Hosmer, D. W., Hosmer, T., Le Cessie, S., and 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. and Lemeshow, S. (1989). *Applied Logistic Regression*. John Wiley and Sons, Inc., United states of America.
- Howell, P. I. and Chadee, D. D. (2007). The influence of house construction on the indoor abundance of mosquitoes. *Journal of Vector Ecology*, 32(1):69–74.

- Härdle, W. (1990). *Applied nonparametric regression*. Cambridge University press, New York, USA.
- Härdle, W., Mammen, E., and Müller, M. (1998). Testing parametric versus semiparametric modeling in generalized linear models. *Journal of the American Statistical Association*, 93(444):1461–1474.
- Hunter, P. (2003). Climate change and waterborne and vector-borne disease. *Journal of Applied Microbiology*, 94:37–46.
- Izenman, A. J. (1991). Recent developments in nonparametric density estimation. *Journal of the American Statistical Association*, 86(413):pp. 205–224.
- Jorgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 49(2):pp. 127–162.
- Kateera, F., Mens, P. F., Hakizimana, E., Ingabire, C. M., Muragijemariya, L., Kalinda, P., Grobusch, M. P., Mutesa, L., and van Vugt, M. (2015). Malaria parasite carriage and risk determinants in a rural population: a malariometric survey in Rwanda. *Malaria journal*, 14(1):16.
- Keating, J., Macintyre, K., Mbogo, C. M., Githure, J. I., and Beier, J. C. (2005). Self-reported malaria and mosquito avoidance in relation to household risk factors in a Kenyan coastal city. *Journal of Biosocial Science*, 37:761–771.
- Kish, L. (1965). *Survey Sampling*. John Wiley and sons, New York.
- Kish, L. and Frankel, M. R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(1):pp. 1–37.
- KNBS (2010). *The 2009 Kenya population and housing census*. Kenya National Bureau of Statistics, Nairobi, Kenya.

- Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2005). *Applied Linear statistical models*. McGraw - Hill Irwin.
- Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):pp. 963–974.
- Lavergne, C., Martinez, M.-J., and Trottier, C. (2008). Empirical model selection in generalized linear mixed effects models. *Computational Statistics*, 23(1):99–109.
- Lin, D. Y. and Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, 96(453):pp. 103–113.
- Lin, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika*, 84(2):pp. 309–326.
- Lin, X. (2007). Estimation using penalized quaslikelihood and quasi-pseudo-likelihood in poisson mixed models. *Lifetime Data Analysis*, 13(4):533–544.
- Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):381–400.
- Lindsey, J. (1997). *Applying Generalized Linear Models*. Springer texts in Statistics, Newyork, USA.
- Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46(3):pp. 673–687.
- Linton, O. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, 82(1):pp. 93–100.

- Liu, Q. and Pierce, D. A. (1994). A note on Gauss-Hermite quadrature. *Biometrika*, 81(3):pp. 624–629.
- Lu, M. and Yang, W. (2012). Multivariate logistic regression analysis of complex survey data with application to brfss data. *Journal of Data science*, 10:157–173.
- Lu, W. W. (2004). *Confidentiality and variance estimation in complex surveys*. PhD thesis, Simon Fraser University.
- McCullagh, P. and Nelder, J. (1983). *Generalized Linear Models*. Chapman and Hall Ltd, New York, USA.
- McCulloch, C., Searle, S., and N., N. (2008). *Generalized, Linear, and Mixed Models*. John Wiley and Sons, Inc., New Jersey, USA.
- Medrano, P., Rodríguez, C., and Villa, E. (2008). Does mother’s education matter in child’s health? Evidence from South Africa. *South African Journal of Economics*, 76(4):612–627.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.
- Ministry of Health (2015). *Kenya annual malaria report 2013/2014*. National malaria control program, ministry of Health, Nairobi, Kenya.
- Ministry of Health, K. Kenya malaria fact sheet. <http://www.nmcp.or.ke/index.php/kenya-malaria-fact-sheet>. Accessed:2015-11-13.
- Ministry of public health and sanitation (2010). *National guidelines for the diagnosis, treatment and prevention of malaria in Kenya*. Division of malaria control, Ministry of public health and sanitation, Nairobi, Kenya.

- Molineaux, L., Gramiccia, G., et al. (1980). The Garki project: research on the epidemiology and control of malaria in the Sudan savanna of West Africa.
- Mwenesi, H., Harpham, T., and Snow, R. W. (1995). Child malaria treatment practices among mothers in Kenya. *Social Science & Medicine*, 40(9):1271–1277.
- Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):pp. 370–384.
- Nevill, C., Some, E., Mung'Ala, V., Muterni, W., New, L., Marsh, K., Lengeler, C., and Snow, R. (1996). Insecticide-treated bednets reduce mortality and severe morbidity from malaria among children on the Kenyan coast. *Tropical Medicine & International Health*, 1(2):139–146.
- Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, 16(1):pp. 1–32.
- Njau, J. D., Stephenson, R., Menon, M. P., Kachur, S. P., and McFarland, D. A. (2014). Investigating the important correlates of maternal education and childhood malaria infections. *The American journal of tropical medicine and hygiene*, 91(3):509–519.
- Noor, A. M., Omumbo, J. A., Amin, A. A., Zurovac, D., and Snow, R. W. (2006). Wealth, mother's education and physical access as determinants of retail sector net use in rural Kenya. *Malar J*, 5(5):5.
- Nyarango, P. M., Gebremeskel, T., Mebrahtu, G., Mufunda, J., Abdulummini, U., Ogbamariam, A., Kosia, A., Gebremichael, A., Gunawardena, D., Ghebrat, Y., et al. (2006). A steep decline of malaria morbidity and mortality trends in Eritrea between 2000 and 2004: the effect of combination of control methods. *Malaria journal*, 5(1):33.

- Okrah, J., Traoré, C., Palé, A., Sommerfeld, J., and Müller, O. (2002). Community factors associated with malaria prevention by mosquito nets: an exploratory study in rural Burkina Faso. *Tropical Medicine & International Health*, 7(3):240–248.
- Olliaro, P. (2008). Mortality associated with severe plasmodium falciparum malaria increases with age. *Clinical Infectious Diseases*, 47(2):158–160.
- Olsson, U. (2002). *Generalized linear models: An applied approach*. Studentlitteratur, Sweden.
- O’Meara, W. P., Bejon, P., Mwangi, T. W., Okiro, E. A., Peshu, N., Snow, R. W., Newton, C. R., and Marsh, K. (2008). Effect of a fall in malaria transmission on morbidity and mortality in Kilifi, Kenya. *The Lancet*, 372(9649):1555–1562.
- Pan, Z. and Lin, D. Y. (2005). Goodness-of-fit methods for generalized linear mixed models. *Biometrics*, 61(4):1000–1009.
- Park, H.-A. (2013). An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain. *Journal of Korean Academy of Nursing*, 43(2):154–164.
- Pendergast, J. F., Gange, S. J., Newton, M. A., Lindstrom, M. J., Palta, M., and Fisher, M. R. (1996). A survey of methods for analyzing clustered binary response data. *International Statistical Review / Revue Internationale de Statistique*, 64(1):pp. 89–118.
- Peng, C.-Y. J., Lee, K. L., and Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, 96(1):3–14.
- Peterson, I., Borell, L. N., Wafaa, E.-S., and teklehaimanot., A. (2009). Individual and household level factors associated with malaria incidence in a highland region of Ethiopia:

- A multilevel analysis. *The American Journal of tropical Medicine and Hygiene*, 80(1):103–111.
- Quenouille, M. H. (1949). Problems in plane sampling. *The Annals of Mathematical Statistics*, 20(3):pp. 355–375.
- Rao, J. (1997). Developments in sample survey theory: An appraisal. *Canadian Journal of Statistics*, 25(1):1–21.
- Rao, J. N. K. and Wu, C. F. J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83(401):pp. 231–241.
- Rodriguez, G. and Goldman, N. (2001). Improved estimation procedures for multilevel models with binary response: A case-study. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 164(2):pp. 339–355.
- Rust, K. (1986). Efficient replicated variance estimation. In *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pages 81–87.
- Sardy, S. and 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.
- SAS Institute Inc. (2015). *SAS/STAT<sup>®</sup> 14.1 User's guide*. SAS Institute Inc., Cary, NC, USA.
- Schellenberg, D., Menendez, C., Aponte, J., Guinovart, C., Mshinda, H., Tanner, M., and Alonso, P. (2004). The changing epidemiology of malaria in Ifakara Town, southern Tanzania. *Tropical Medicine & International Health*, 9(1):68–76.
- Schofield, C. and White, G. (1984). Engineering against insect-borne diseases in the domestic environment. *Trans. R. Soc. Trop. Med. Hyg*, 78:285–292.

- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):pp. 461–464.
- Shiff, C. (2002). Integrated approach to malaria control. *Clinical microbiology reviews*, 15(2):278–293.
- 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):pp. 1–52.
- Sintasath, D. M., Ghebremeskel, T., Lynch, M., Kleinau, E., Bretas, G., Shililu, J., Brantly, E., Graves, P. m., and Beier, J. C. (2005). Malaria prevalence and associated risk factors in Eritrea. *The American Journal of Tropical Medicine and Hygiene*, 72(6):682–687.
- Siri, J. G. (2014). Independent associations of maternal education and household wealth with malaria risk in children. *Ecol Soc*, 19(1):33.
- Sitter, R. R. (1992a). Comparing three bootstrap methods for survey data. *Canadian Journal of Statistics*, 20(2):135–154.
- Sitter, R. R. (1992b). A resampling procedure for complex survey data. *Journal of the American Statistical Association*, 87(419):755–765.
- Skinner, C., Holt, D., and Smith, T. (1989). *Analysis of Complex surveys*. John Wiley and sons, New York.
- Smyth, G. K. (2003). Pearson’s goodness of fit statistic as a score test statistic. *Lecture Notes-Monograph Series*, 40:pp. 115–126.
- Snow, R. W. (2015). Global malaria eradication and the importance of Plasmodium falciparum epidemiology in Africa. *BMC medicine*, 13(1):1.

- Snyman, K., Mwangwa, F., Bigira, V., Kapisi, J., Clark, T. D., Osterbauer, B., Greenhouse, B., Sturrock, H., Gosling, R., Liu, J., et al. (2015). Poor housing construction associated with increased malaria incidence in a cohort of young Ugandan children. *The American journal of tropical medicine and hygiene*, 92(6):1207–1213.
- Spall, J. (2003). Estimation via Markov chain Monte Carlowi. *Control Systems, IEEE*, 23(2):34–45.
- Staniswalis, J. G. and Lee, J. J. (1998). Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, 93(444):1403–1418.
- Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2):pp. 689–705.
- Stram, D. O. and Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50(4):pp. 1171–1177.
- Sutradhar, B. C. and Rao, R. (2001). On marginal quasi-likelihood inference in generalized linear mixed models. *Journal of Multivariate Analysis*, 76(1):1–34.
- Ter Kuile, F. O., Terlouw, D. J., Phillips-Howard, P. A., Hawley, W. A., Friedman, J. F., Kolczak, M. S., Kariuki, S. K., Shi, Y. P., Kwena, A. M., Vulule, J. M., and Nahlen, B. L. (2003). Impact of permethrin-treated bed nets on malaria and all-cause morbidity in young children in an area of intense perennial malaria transmission in Western Kenya: A cross-sectional survey. *The American Journal of Tropical Medicine and Hygiene*, 68(Suppl 4):68–77.
- Tuerlinckx, F., Rijmen, F., Verbeke, G., and De Boeck, P. (2006). Statistical inference in generalized linear mixed models: A review. *British Journal of Mathematical and Statistical Psychology*, 59(2):225–255.

- Tukey, J. W. (1958). Bias and confidence in not-quite large samples. In *Annals of Mathematical statistics*, volume 29, pages 614–614. Inst Mathematical Statistics IMS business office suite 7, 3401 Investment Blvd, Hayward, CA 94545.
- Unicef (2007). Malaria in children-progress in intervention coverage. *New York City, USA*.
- Verbeke, G. and Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics*, 59(2):pp. 254–262.
- Verbeke, G. and Molenberghs, G. (2009). *Linear Mixed models for Longitudinal data*. Springer series in Statistics, USA.
- Vonesh, E. F., Chinchilli, V. M., and Pu, K. (1996). Goodness-of-fit in generalized nonlinear mixed-effects models. *Biometrics*, 52(2):pp. 572–587.
- Wahba, G. (1975). Smoothing noisy data with spline functions. *Numerische Mathematik*, 24(5):383–393.
- WHO (2014). *World malaria report 2013*. World Health Organization.
- WHO (2015). *World malaria report 2015*. World Health Organization.
- Wilber, S. T. and Fu, R. (2010). Risk ratios and odds ratios for common events in cross-sectional and cohort studies. *Academic Emergency Medicine*, 17(6):649–651.
- Winskill, P., Rowland, M., Mtove, G., Malima, R. C., and Kirby, M. J. (2011). Malaria risk factors in north-east Tanzania. *Malaria journal*, 10(1):1.
- Wolter, K. M. (1985). *Introduction to variance estimation*. Springer-Verlag, New York.
- Wood, S. (2006). *Generalized additive models: an introduction with R*. CRC press.
- Wood, S. and Wood, M. S. (2007). The mgcv package. *www.r-project.org*.

- Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36.
- Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66(334):411–414.
- Worrall, E., Basu, S., and Hanson, K. (2003). The relationship between socio-economic status and malaria: a review of the literature.
- Worrall, E., Basu, S., and Hanson, K. (2005). Is malaria a disease of poverty? a review of the literature. *Tropical Medicine & International Health*, 10(10):1047–1059.
- Woyessa, A., Deressa, W., Ali, A., and Lindtjørn, B. (2013). Malaria risk factors in Butajira area, south-central Ethiopia: a multilevel analysis. *Malar J*, 12:273.
- Yadav, K., Dhiman, S., Rabha, B., Saikia, P., and Veer, V. (2014). Socio-economic determinants for malaria transmission risk in an endemic primary health centre in Assam, India. *Infectious diseases of poverty*, 3(19).
- Yung, W. and Rao, J. N. K. (2000). Jackknife variance estimation under imputation for estimators using poststratification information. *Journal of the American Statistical Association*, 95(451):903–915.
- Zeger, S. L. and Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics*, 50(3):pp. 689–699.
- Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects; A Gibbs sampling approach. *Journal of the American Statistical Association*, 86(413):pp. 79–86.

- Zhang, D. and Lin, X. (2008). Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics. In *Random Effect and Latent Variable Model Selection*, volume 192 of *Lecture Notes in Statistics*, pages 19–36. Springer New York.
- Zhang, D., Lin, X., Raz, J., and Sowers, M. (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association*, 93(442):710–719.
- Zhang, H., Lu, N., Feng, C., Thurston, S. W., Xia, Y., Zhu, L., and Tu, X. M. (2011). On fitting generalized linear mixed effect models for binary response using different statistical packages. *Statistics in Medicine*, 30(20):pp. 2562–2572.