Prevalence and Risk Factors of Malaria in Children Under the Age of Five Years Old in Uganda

Danielle Roberts

December, 2014

Prevalence and Risk Factors of Malaria in Children Under the Age of Five Years Old in Uganda

by

Danielle Roberts

A thesis submitted to the

University of KwaZulu-Natal

in fulfilment of the requirements for the degree

of

MASTER OF SCIENCE IN STATISTICS

SCHOOL OF MATHEMATICS, STATISTICS AND COMPUTER SCIENCE

UNIVERSITY OF KWAZULU-NATAL

WESTVILLE CAMPUS, DURBAN, SOUTH AFRICA

Declaration - Plagiarism

I, Danielle Roberts, declare that

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

_______________________________   _______________________________

Danielle Roberts (Student)   Date

_______________________________   _______________________________

Prof. G.B. Matthews (Supervisor)   Date

Disclaimer

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

Abstract

Malaria is considered to be one of the main global health problems, with it causing close to a million deaths each year. Ninety percent of these deaths occur in Sub-Saharan Africa and 70% are of children under the age of 5 years. Uganda, ranked 6th worldwide in the number of malaria cases and 3rd in the number of malaria deaths in 2008, experiences weather conditions that often allow malaria transmission to occur all year round with only a few areas that experience low or unstable transmission. Malaria is the leading cause of morbidity in Uganda with 95% of the population at risk and it killing between 70,000 and 100,000 children every year. Children under the age of five years are among the most vulnerable to malaria infection as they have not yet developed any immunity to the disease.

In order to apply successful implementations to eradicate malaria, there is a continuous need to understand the epidemiology and risk factors associated with the disease. Although a large number of studies done worldwide have identified a wide variety of risk factors; socioeconomic, environmental, demographic, and others, associated with malaria infection, there is still a great need to identify the influence of these factors in a local context to allow a successful formulation of a national malaria-control strategy. There have, however, been very few studies done in Uganda on malaria indicators and risk factors. These studies have also been specific to one community at a time. Most recent studies on malaria in Uganda have been hospital-based, investigating clinical malaria among young children and pregnant women. One of the aims of this thesis was to identify significant socio-economic, demographic and environmental risk factors associated with malaria infection, based on the result of a microscopy test conducted on 3,972 children under the age of five during a nationally represented Malaria Indicator Survey (MIS) done in Uganda in 2009.

The MIS sample was stratified according to 10 regions of Uganda and was not spread geographically in proportion to the population, but rather equally across the regions.

The survey consisted of a two-stage sample design where the first stage involved selecting clusters, with probability proportional to size, from a list of enumeration areas. The second stage involved systematic sampling of households from a list of households in each cluster. Surveys carried out using these sampling techniques are referred to as having complex survey designs.

The response variable of interest is binary, indicating whether a child tested positive or negative for malaria. Logistic regression is commonly used to explore the relationship between a binary response variable and a set of explanatory variables. However, this method of analysis is not valid if the data come from complex survey designs. Failure to account for the complex design of a study may result in an overestimation of standard errors, therefore leading to incorrect results. There are many methods of dealing with this design of the study. Two such commonly used approaches are design-based and model-based statistical methods. A designed-based method, which involves the extension of logistic regression to complex survey designs, is survey logistic regression. For design-based methods, parameter estimates and inferences are based on the sampling weights, and only inferences concerning the effects of certain covariates on the response variable are of interest. However, model-based methods are used when interest is also on estimating the proportion of variation in the response variable that is attributable to each of the multiple levels of sampling. In this case, inference on the variance components of the model may also be of interest. Such methods include generalized linear mixed models and generalized estimating equations. This thesis discusses these three methods of analyzing complex survey designs and compares the results of each applied to the MIS data.

Acknowledgements

I would first like to thank my supervisor, Prof. Glenda Matthews, for all the encouragement and guidance. This work could not have been accomplished without her continued support. The solid foundation and keen interest that Prof. Matthews gave me in statistical models, during my undergraduate and Honours degree, also played a huge part in this thesis, for which I am very grateful. Thank you to SACEMA (South African Centre for Epidemiological Modelling and Analysis) for all the financial and academic support. Their research days were inspiring to me and gave me the opportunity to meet many wonderful people. Many thanks to the DHS (Demographic and Health Surveys) Program for allowing me access to their data.

I am truly thankful to all the staff members of the Department of Statistics at the University of KwaZulu-Natal. Everyone has always been so friendly and supportive. A special thanks goes to Prof. Delia North, who encouraged me to do this Masters degree in the first place.

Lastly, I wish to thank all my friends and family for their continued support and encouragement, especially my mom and fiancé who always believed in me, it has truly touched me and I will be forever grateful.

Contents

Page

List of Figures

List of Tables

Abbreviations

aOR	Adjusted Odds Ratio
BRR	Balanced Repeated Replication
CDC	Centers for Disease Control and Prevention
CI	Confidence Interval
DHS	Demographic and Health Surveys
GDP	Gross Domestic Product
GEEs	Generalized Estimating Equations
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Model
IRWLS	Iterative Reweighted Least Squares
ITNs	Insecticide-Treated Nets
IWLS	Iterative Weighted Least Squares
JRR	Jackknife Repeated Replication
LLIN	Long-lasting Insecticidal Net
LM	General Linear Model
LMM	Linear Mixed Model
ML	Maximum Likelihood
MCP	Malaria Control Programme
MCU	Malaria Control Unit
MIS	Malaria Indicator Survey
OLS	Ordinary Least Square
OR	Odds Ratio
PML	Pseudo-Maximum Likelihood
PSU	Primary Sampling Unit
QL	Quasi-Likelihood
RBM	Roll Back Malaria
REML	Restricted Maximum Likelihood
SLR	Survey Logistic Regression
UBOS	Uganda Bureau of Statistics
WHO	World Health Organization

Background

Malaria, which is caused by the *Plasmodium* parasite that is transmitted via the bite of infected Anopheles mosquitoes, is considered to be one of the main global health problems, with it causing close to a million deaths each year (Malaria Foundation International, 2013). Ninety percent of these deaths occur in Sub-Saharan Africa and 70% are of children under the age of 5 years. This is equivalent to one child in Africa dying of malaria every 30 seconds (Against Malaria Foundation, 2013). According to the World Health Organization (WHO), 7% of all deaths in children under the age of 5 years in 2010 was caused by malaria.

Malaria imposes substantial costs to both individuals and governments (Centers for Disease Control and Prevention (CDC, 2012)). It has a considerable effect on poverty by affecting young people who would otherwise enter the workforce and contribute to the local economy (Hochman & Kim, 2009). These infected individuals require diagnosis, treatment and sometimes hospital care therefore creating a burden on health services. The individual is prevented from going to work or children from going to school, thus creating a knock-on effect on the economy (Malaria.com, 2013). High levels of absenteeism from school can hinder efforts to improve literacy rates and stall the progress of education systems. Malaria is also considered a contributing factor in decreasing the gross domestic product (GDP) in countries with high infection rates. Over the long run these economic losses add up, resulting in substantial differences in the GDP between countries with and without malaria, particularly in Africa. This presents an enormous challenge to efforts of lifting people out of poverty (Gallup & Sachs, 2001).

In order to apply successful implementations to eradicate malaria, there is a continuous need to understand the epidemiology and risk factors associated with the disease (Pullan et al., 2010). Within the last decade, increasing numbers of partners and resources have rapidly increased malaria control efforts (CDC, 2012). (WHO, 2013).

This increase in interventions has led to decreased morbidity and mortality in a number of countries, with mortality rates decreasing by more than 25% worldwide between 2000 and 2010, and a decrease of 33% in the region of Africa where the burden of the disease is greatest (WHO, 2013). Extensive research has been able to define new tools and strategies for malaria control, such as long-lasting insecticide-treated nets (LLIN), indoor residual spraying, intermittent preventative treatment in pregnancy and infants, and prompt and effective treatment of infected individuals (Johansson et al., 2007). Although a large number of studies have identified a wide variety of risk factors; socioeconomic, environmental, demographic, and others, associated with malaria infection (Nahum et al., 2010), there is still a great need to identify the influence of these factors in a local context to allow a successful formulation of a national malaria-control strategy. Despite the large efforts being made to alleviate the burden of malaria, there are many barriers still to overcome. These include drug and insecticide resistance, the availability of health care professionals to administer tests and treatments, and according to Mayah (2011) climate change has intensified the threat of malaria.

Malaria transmission occurs primarily in tropical and subtropical regions but usually not higher than 1,500 - 2,000m above sea level. Climate affects both the parasite and the mosquito. Mosquitoes are unable to survive in low humidity and their breeding grounds are expanded by rainfall. *Plasmodium* parasites are affected by temperature where their development slows as the temperature drops and stops at high temperatures, which explains why parasites can be found in temperate areas (National Institute of Allergy and Infectious Diseases, 2007). Uganda, ranked 6th worldwide in the number of malaria cases and 3rd in the number of malaria deaths in 2008 (WHO, 2008), experiences weather conditions that often allow transmission to occur all year round with only a few areas that experience low or unstable transmission (Malaria Control Programme, 2005 - 2010). Malaria is the leading cause of morbidity in Uganda with 95% of the population at risk and it killing between 70,000 and 100,000 children every year (Malaria Control Programme, 2005 - 2010). Children under the age of five years are among the most vulnerable to malaria infection as they have not yet developed any immunity to the disease (CDC, 2012).

In Uganda, malaria control received little attention from the Ministry of Health before 1995, after which the Malaria Control Programme (MCP) was established in order to direct and guide the day to day implementation of the National Malaria Control Strategy (Malaria Control Programme, 2005 - 2010). Uganda was one of the first countries to introduce a waiver of taxes and tariffs for insecticide-treated nets (ITNs).

Today, the fight against malaria is part of the overall effort of the Government of Uganda to improve health. This effort is multi-sectoral and involves a broad partnership which forms the Roll Back Malaria Country Partnership (Ministry of Health Online, 2013). The Roll Back Malaria (RBM) Partnership, which was launched in 1998 by numerous partners[1], is a forum of all stakeholders in malaria control with a goal to alleviate the burden of the disease (Ministry of Health Online, 2013). In 2002, Roll Back Malaria established the Monitoring and Evaluation Reference Group (MERG) which was responsible for developing the Malaria Indicator Survey (MIS) in an effort to monitor and evaluate efforts of malaria control (MEASURE DHS, MEASURE Evaluation, Presidents Malaria Initiative, Roll Back Malaria and United Nations Childrens Fund, 2013). MIS is a stand-alone household survey which collects national and regional or provincial data from a representative sample of respondents, and depending on the needs of the country, it may also include a measurement of malaria parasites among household members most at risk. The MIS done in Uganda in 2009, which includes results of two test procedures for malaria in children under the age of five years, will provide the data set for this thesis.

There have been very few studies done in Uganda on malaria indicators and risk factors. These studies have also been specific to one community at a time (Pullan et al., 2010; Namanya, 2013; Clark et al., 2008). Most recent studies on malaria in Uganda have been hospital-based, investigating clinical malaria among young children and pregnant women (Idro et al., 2006, 2005; Ndyomugyenyi & Magnussen, 2001, 2004; Kiggundu et al., 2013). The Malaria Indicator Survey in 2009 was the first nationally representative survey of malaria to be done in Uganda and the Ministry of Health plans to carry out this survey every two to three years (Uganda Bureau of Statistics (UBOS) and ICF Macro, 2010).

[1] World Health Organization (WHO), United Nations Children's Fund (UNICEF) and United Nations Development Programme (UNDP)

Chapter 1

Introduction

1.1 Country Profile

The Republic of Uganda is a small, landlocked country located on the equator in East Africa. It shares borders with Sudan in the north, Kenya in the east, the Democratic Republic of Congo (DRC) in the west, and Tanzania and Rwanda in the south (Uganda Bureau of Statistics (UBOS) and ICF Macro, 2010). The country has an area of 241,039 square kilometres with approximately 45,000 square kilometres being taken up by open water and swamps, as can be seen in Figure 1.1 on page 5. This abundance of water bodies in the country provides great breeding grounds for the Anopheles mosquito. Uganda is administratively divided up into 111 districts and one capital city, Kampala. However, at the time of designing and collecting data for the MIS in Uganda in 2009, there were only 80 districts (Uganda Bureau of Statistics (UBOS) and ICF Macro, 2010).

Uganda experiences a favourable tropical climate due to its relatively high altitude with most of the Southwest lying between altitudes of 1,300 and 1,500m above sea level. High mountain ranges above 1,800m are found in the border region in the Southwest with Rwanda and the DRC, the Rwenzori Mountains in the West and Mount Elgon in the East (Malaria Control Programme, 2005 - 2010). These areas are sometimes prone to epidemics, experiencing low or unstable malaria transmission. Uganda has mean annual temperatures of 16°C in the Southwest, 25°C in the Centre, East and Northwest and close to 30°C in the Northeast. Average relative humidity varies between 54% and 88%.

Uganda experiences two rainy seasons per year, with heavy rains from March to May and light rains between September and December. The peak incidence of clinical malaria follows the peak of the rains with a delay of about 4 to 6 weeks, therefore most cases are seen between December and February, and May and July. However, rainfall decreases in the North region of the country, turning it into just one rainy season per year, thus the malaria season is more between May and November (Malaria Control Programme, 2005 - 2010).

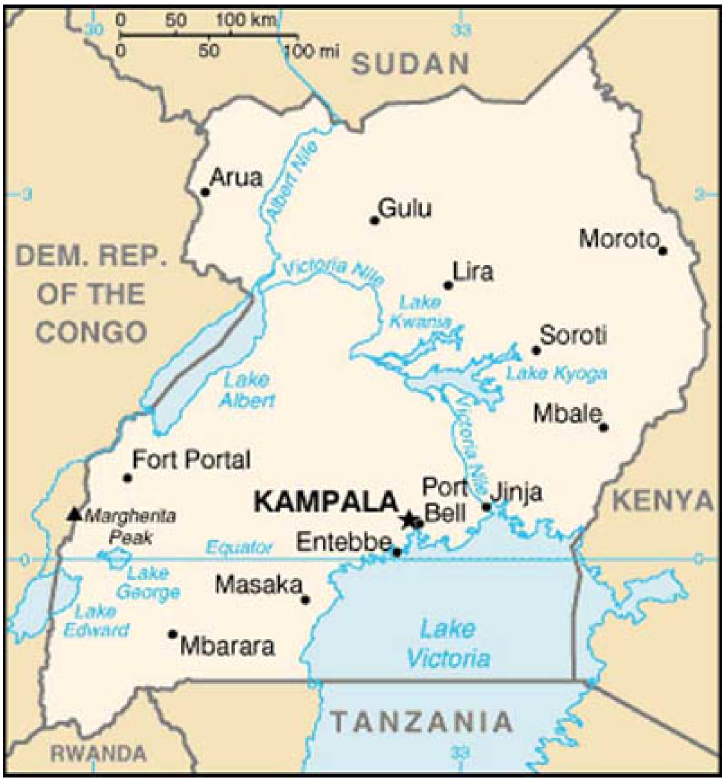

Figure 1.1: Map of Uganda

Uganda is currently densely populated with a population of 35 million, and with a population of 24,5 million in 2002 (2002 Uganda Population and Housing Census Report), the annual population growth is currently 3.5%. Uganda has a very young population, with a median age of 15 years. It also has the second highest fertility rate in the world, at 6.2 births per woman (Population Secretariat (POPSEC) of Uganda, 2012). Some research has suggested this is an effect of high childhood mortality from malaria, where parents plan on extra births in case a child is lost (Weil, 2010). The economy of Uganda is predominately agricultural, with the majority of the population dependent on subsistence farming (Uganda Bureau of Statistics (UBOS) and ICF International Inc., 2012).

Due to regular rainfall, the Southwest and Centre regions are rich in vegetation and fertile soil, resulting in high population densities. Thus, 87.9% of the population is exposed to moderate to very high malaria transmission (Malaria Control Programme, 2005 - 2010). Uganda is considered one of the poorest countries in the world with 24.5% of the population in 2009 living below the national poverty line[1] (The World Bank, 2013). The majority of the population lives in rural areas, where only 13.3% in 2010 were living in urban areas (Uganda Bureau of Statistics (UBOS) and ICF International Inc., 2012).

1.2 The Data Set

The MIS was designed to provide national, regional, urban and rural estimates of key malaria indicators (Uganda Bureau of Statistics (UBOS) and ICF Macro, 2010). The survey was carried out during November and December 2009 to correspond with peak malaria transmission. The survey was nationally represented with the following objectives:

- Measure the extent of ownership and use of mosquito nets.
- Assess coverage of the intermittent preventative treatment (IPT) programme for pregnant women.
- Identify practices used to treat malaria among children under the age of five and the use of specific antimalarial medications.
- Measure prevalence of malaria and anaemia among children 0-59 months.
- Determine the species of *Plasmodium* parasite most prevalent in Uganda.
- Assess knowledge, attitudes, and practices of malaria in the general population.

Two questionnaires were used in the MIS, a Household Questionnaire and a Woman's Questionnaire for all women aged 15 to 49 in the sample. Both instruments were based on the standard MIS Questionnaire developed by the RBM and DHS programmes, with a few modifications to reflect issues relevant to malaria in Uganda. The National Malaria Control Programme (NMCP) was responsible for developing the survey protocol.

[1] Defined as the percentage of people living on $1.20 per day.

1.2.1 Sampling Procedure

The sample was stratified into 9 survey regions of the country, plus the capital city, Kampala, which, due to it being entirely an urban district, comprised a separate region. Each of the 9 other regions consisted of 8 to 10 administrative districts of Uganda that shared similar languages and cultural characteristics. The following figure represents the 10 regions:

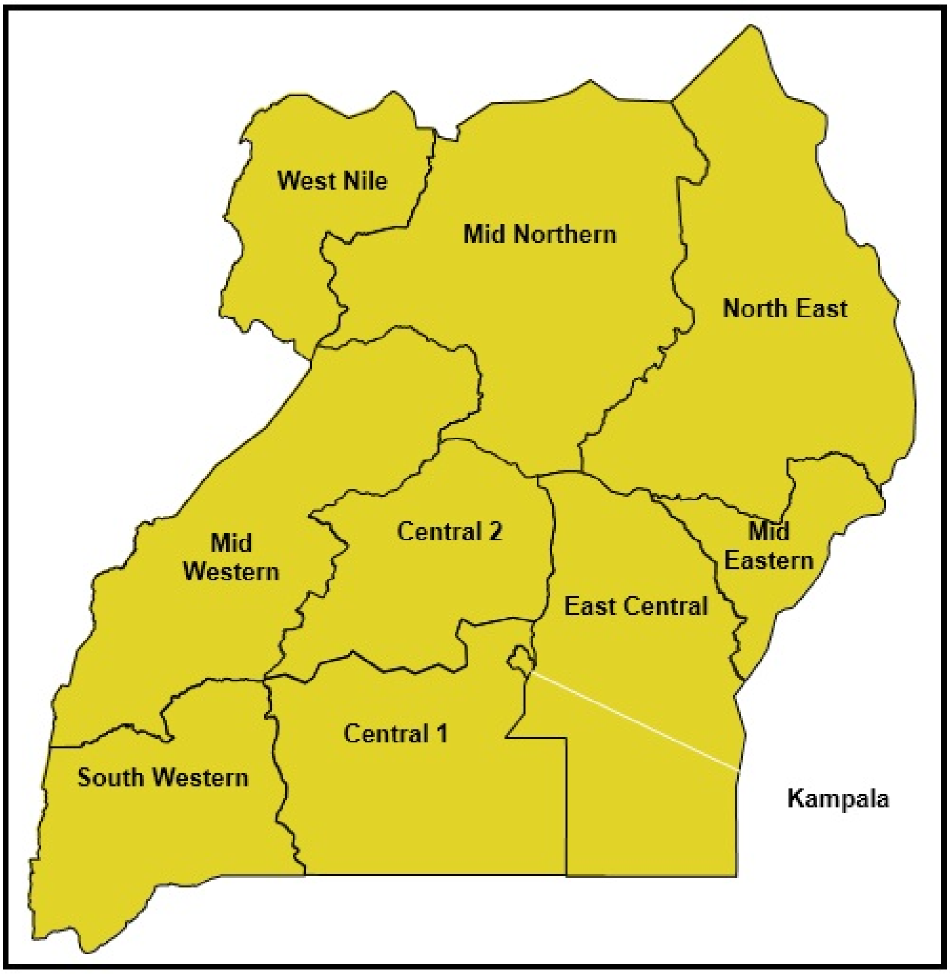

Figure 1.2: MIS 2009 Sample Regions in Uganda

The sample was not spread geographically in proportion to the population, but rather equally across the regions, with 17 clusters per region. The survey consisted of a two-stage sample design. The first stage involved selecting clusters from a list of enumerations areas (EA) covered in the 2002 Population Census, these areas made up the primary sampling units (PSUs). A total of 170 clusters (17 clusters for each of the 10 regions) with probability proportional to size were collected. These clusters consisted of 26 urban areas and 144 rural areas.

Table 1.1: Allocation of clusters by region and type of residence.

Region	Urban	Rural
Central 1	0	17
Central 2	1	16
East Central	1	16
Kampala	17	0
Mid Eastern	1	16
Mid Northern	0	17
Mid Western	1	16
North East	1	16
South Western	1	16
West Nile	3	14
Total	26	144

Table 5.7 above shows the number of urban and rural clusters selected in each region. Several months before the survey took place, a list of all households in the 170 clusters was drawn up and provided the sampling frame from which the households were then selected for the survey. The second stage of the selection process involved systematic sampling of households from the list of households in each cluster. Twenty-eight households were selected in each cluster. Thus, the final sample consisted of 4,760 households. Appendix A on page 102 shows the calculation of the sample size as well as the sampling probabilities and weights associated with the households.

1.2.2 Data Collection

The selected households were visited and interviewed by trained staff. The Household Questionnaire collected basic information on the characteristics of each member and recent visitors of the household, including age, sex, and relationship to the head of the household. The Household Questionnaire also collected information on characteristics of the household's dwelling unit, such as source of water; type of toilet facilities; materials used for the floor, roof and walls of the house; ownership of various durable goods; and ownership and use of mosquito nets. The Woman's Questionnaire was used to collect a range of information from all eligible women in the sample. With the consent of a parent or guardian in the household, all children between the ages of 0 and 59 months were tested for malaria and anaemia.

Two types of testing procedures were used to determine the prevalence of malaria in the children; a rapid diagnostic test (RDT) and microscopy.

The RDT consisted of testing a drop of blood using the Paracheck Pf^{TM} rapid diagnostic test, which tests for the parasite *Plasmodium falciparum*, the most dangerous *Plasmodium* parasite. The result of the test was available in 15 minutes. This type of test has become more widely used as a diagnostic test where a reliable microscopy test is not available (Uganda Bureau of Statistics (UBOS) and ICF Macro, 2010). Children who tested positive for malaria using the RDT were offered a full course of treatment according to standard procedures for testing malaria in Uganda, provided they were not currently on treatment and had not completed a full course of Artemisinin-based Combination Therapy (ACT) during the last 2 weeks.

The second test procedure involved taking two blood smears; one thick and one thin. These blood smears were then sent to the Uganda Malaria Surveillance Project (UMSP) Molecular Research Lab at Mulago Hospital in Kampala. The thick smears were first examined by microscopy to determine malaria infection, then the thin smears of all positive thick smears were examined to determine the species of *Plasmodium* parasite.

1.2.3 Variables of Interest

Unlike the microscopy tests, RDTs are more readily available and do not require technicians with advanced skills and laboratories. However, the RDT detects the *Plasmodium falciparum*-specific protein (not the parasite itself), which can remain in the blood for several weeks after treatment. Therefore, this test is less sensitive and often results in slightly higher rates of malaria. Microscopy is considered the gold standard and is a highly sensitive test (Reyburn et al., 2007). Therefore, for the purpose of this thesis, the prevalence of malaria in children under the age of 5 years will be according to the microscopy test results. Thus, the response variable is binary, indicating whether a child tested positive or negative for malaria.

Many studies have shown malaria transmission is not evenly distributed and some regions, households or individuals may be more at risk than others (Greenwood, 1989). Much of these variations have been found to be related to a number of factors. These include a variety of environmental factors such as housing conditions (main material used for the floors, roof and walls) and proximity to vector-breeding sites (Guthmann et al., 2001; Schofield & White, 1984; Sintasash et al., 2005; Peterson et al., 2009), as well as socio-economic factors where low socio-economic conditions have been associated with a higher risk of malaria infection (Koram et al., 1995; Ghebremeskel et al., 2000; Gahutu et al., 2011; Ayele et al., 2012). Mendez et al. (2000) and Koram et al. (1995) have shown an individual's knowledge of malaria has a significant influence on their malaria status. The use of mosquito bed nets has been a significate factor in very many studies done across the world.

Thus, the independent variables that are considered in the modelling of malaria status in this thesis comprise of a number of socio-economic, demographic and environmental factors. Such variables include:

- gender and age of the child
- number of members in the household
- caregiver's age, education level and knowledge of malaria
- type of place of residence: rural or urban
- cluster altitude and region of Uganda
- main source of drinking water
- type of toilet facilities
- whether or not the household had electricity
- whether or not the household had a refrigerator, bicycle, television and a radio
- main material of the floors, walls and roof of the household
- incidence of anti-malarial spraying in the last 12 months
- use of mosquito nets and total number of mosquito nets used in the household

1.3 Thesis Objectives

Surveys carried out using sampling techniques such as multistage sampling, stratified random sampling, cluster sampling or sampling with unequal weights are often referred to as having complex survey designs (Nadimpalli & Hubbell, 2012). Modeling of data obtained from these surveys must take into consideration the design of the study for the following reasons:

- Observations within the same cluster or household may be correlated and thus the assumption of independence in the data cannot be met.
- A limited number of clusters are sampled thus leaving a significant portion of the population unsampled. This may result in certain characteristics not being represented in the study.
- Sample units may be selected with unequal weights or probabilities.
- Often surveys are subjected to non-response. This may result in unmeasured characteristics which could lead to biased results.

Logistic regression, which is a class of generalized linear models, is commonly used to explore the relationship between a binary response variable and a set of explanatory variables. However, this method of analysis is not valid if the data come from complex survey designs (An, 2002). There are many methods of dealing with this design of the study. Two such commonly used approaches are design-based and model-based statistical methods (Ghosh & Pahwa, 2006). A designed-based method, which involves the extension of logistic regression to complex survey designs, is survey logistic regression, first introduced by Binder (1983). For design-based methods, parameter estimates and inferences are based on the sampling weights, and only inferences concerning the effects of certain covariates on the response variable are of interest. However, model-based methods are used when interest is also on estimating the proportion of variation in the response variable that is attributable to each of the multiple levels of sampling (Heeringa et al., 2010). In this case, inference on the variance components of the model may also be of interest. Such methods include generalized linear mixed models and generalized estimating equations. These methods are both extensions of generalized linear models. One of the objectives of this thesis is to examine and compare the different methods used for the analysis of data from complex survey designs. However, the primary objectives of the thesis are:

- to investigate the distribution of malaria infection in children under the age of 5 years old in the different regions of Uganda.
- to investigate the relationship between malaria status of children under the age of 5 years old in Uganda and selected socio-economic, demographic and environmental factors.
- to determine which factors significantly increase the risk of malaria infection in children under the age of 5 years old in Uganda.

1.4 Thesis Overview

In the last section of this chapter, some exploratory data analyses is carried out on the MIS data set. In this section, the observed malaria prevalence is determined for the data set, as well as the prevalence by gender, region, type of residence, and many other variables of interest. Chapter 2 gives a brief overview of linear models where the general linear model and linear mixed model are discussed. Chapter 3 discusses generalized linear models and the use of logistic regression in modeling a binary outcome. In this chapter, the survey logistic regression model is introduced, along with the different methods of variance estimation of the model's parameter estimates.

Chapter 4 gives an overview of generalized estimating equations and generalized linear mixed models and their use of modeling data where a correlation structure may exist. Chapter 5 involves applying the survey logistic regression model to the MIS data set where adjusted odds ratios are determined for significant variables. Generalized estimating equations and a generalized linear mixed model are also fitted to the data in this chapter. The last chapter discusses the results of the different methods used in the analyses of the data. This chapter also discusses the conclusions of the study, as well as possible areas of further study.

1.5 Exploratory Data Analysis

Before any statistical modeling of the data is done, it is ideal to first carry out some exploratory analyses. This enables one to get a general understanding of the data. Out of the 4,760 households interviewed, a total of 4,146 children under the age of 5 were eligible for testing. Of those 4,146 eligible children, 3,972 were tested for malaria using the microscopy test, thus resulting in a response rate of 95.8%. These 3,972 children from a total of 2,491 households make up the sample that will be used in the analyses.

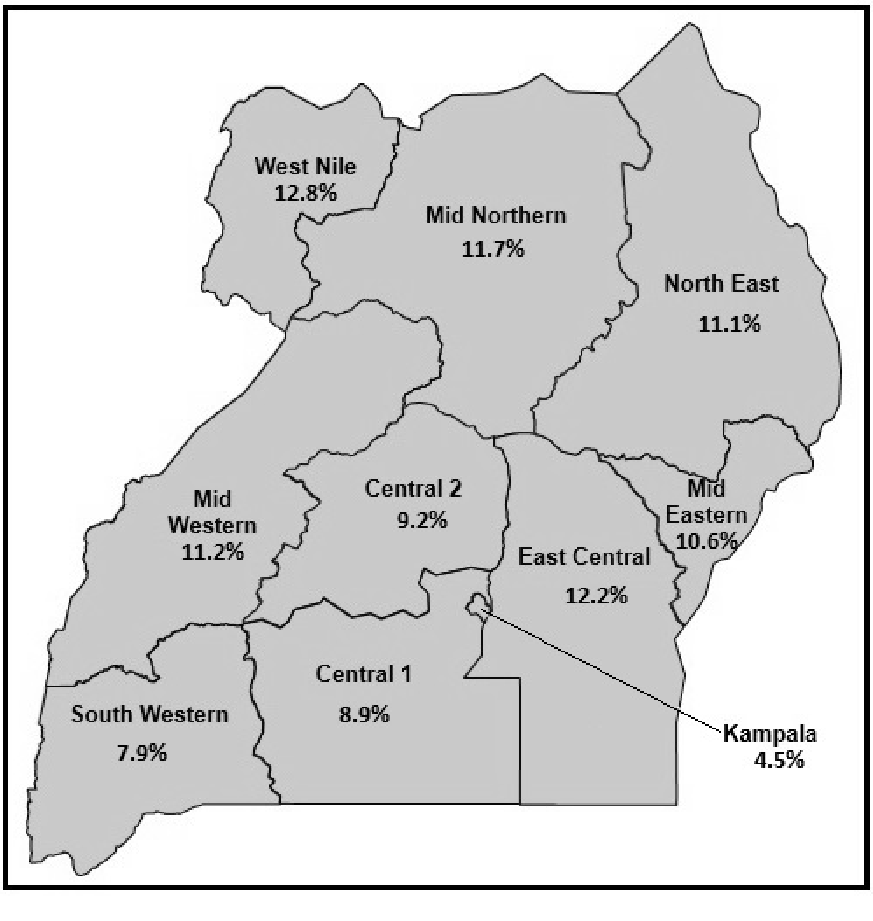

Figure 1.3: Distribution of children under the age of 5 tested for malaria across the regions Uganda.

Figure 1.3 on the previous page shows the percentage of children under the age of 5 tested for malaria in each region. Kampala, which is entirely urban and the smallest region, had the lowest number of children at 4.5%, with the rest of the regions ranging from 7.9% to 12.8%.

A total of 1,725 children tested positive for malaria, which was 43.4% of the sample. More than half of the children in the sample in regions Central 2, East Central and Mid Northern tested positive for malaria, with the East Central region of Uganda having the highest observed prevalence of 67.7% of the 12.2% of children tested in the region. This region, which experiences high malaria transmission, borders Lake Victoria. Some studies have suggested Lake Victoria is a fertile breeding ground for malaria vectors (Minakawa et al., 2012).

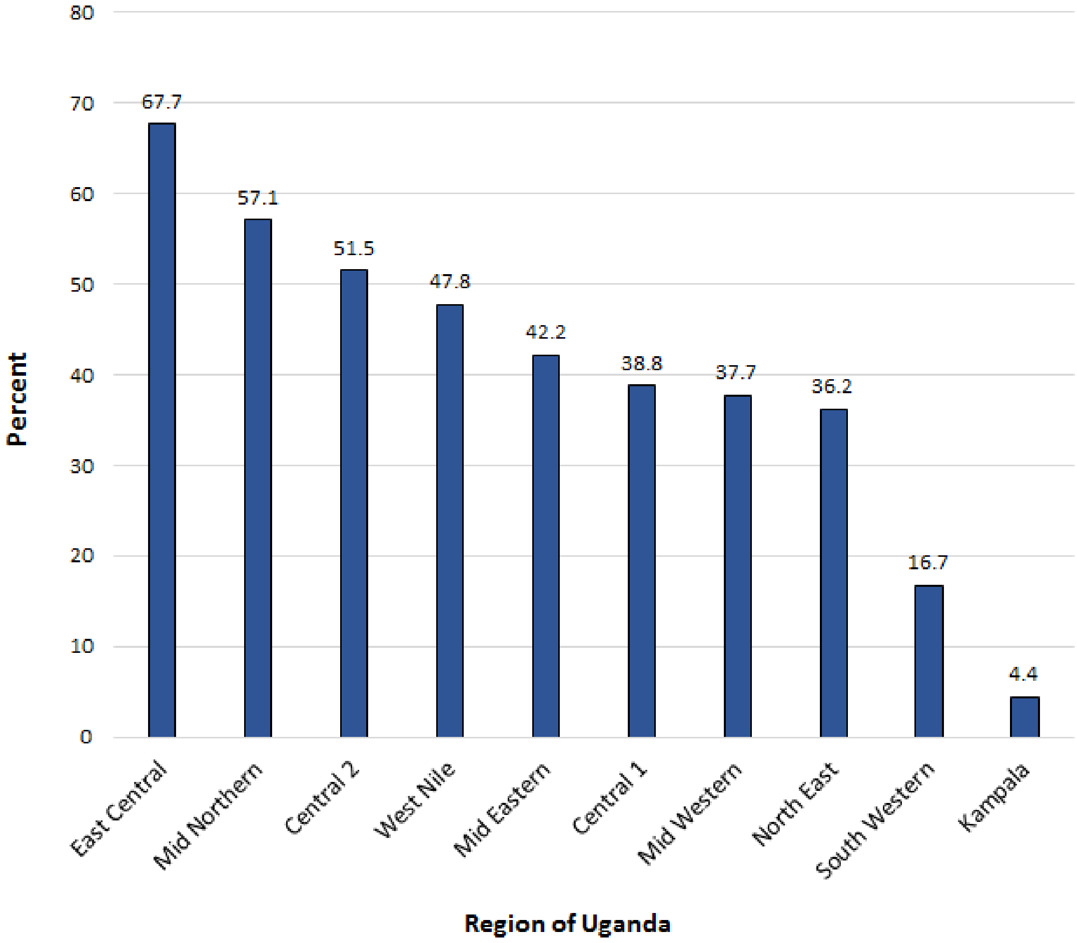

Figure 1.4: Observed prevalence of malaria according to region of Uganda.

Kampala had the lowest prevalence with only 4.4% of the children in the region testing positive. This could be a result of the low number of children under the age of 5 tested in the region or due to the fact that the region consists of only urban areas.

The table below shows that urban areas only made up 10.8% of the households in the sample, which is a very low portion and therefore needs to be kept in mind during the analyses. The majority of the sampled households were in clusters with altitudes ranging between 1000m and 1500m. A very small portion of the households (7.4%) were in clusters with altitudes higher than 1500m, where malaria transmission is lower.

Table 1.2: Percentage of households in the sample according to type of place of residence and cluster altitude.

Cluster Altitude	Type of place of Residence		**Total**
	Urban	Rural	
< 1000m	2.2	7.8	10.1
1000 - 1500m	8.6	73.9	82.5
1500 - 2000m	0	6.2	6.2
> 2000m	0	1.2	1.2
Total	10.8	89.2	100

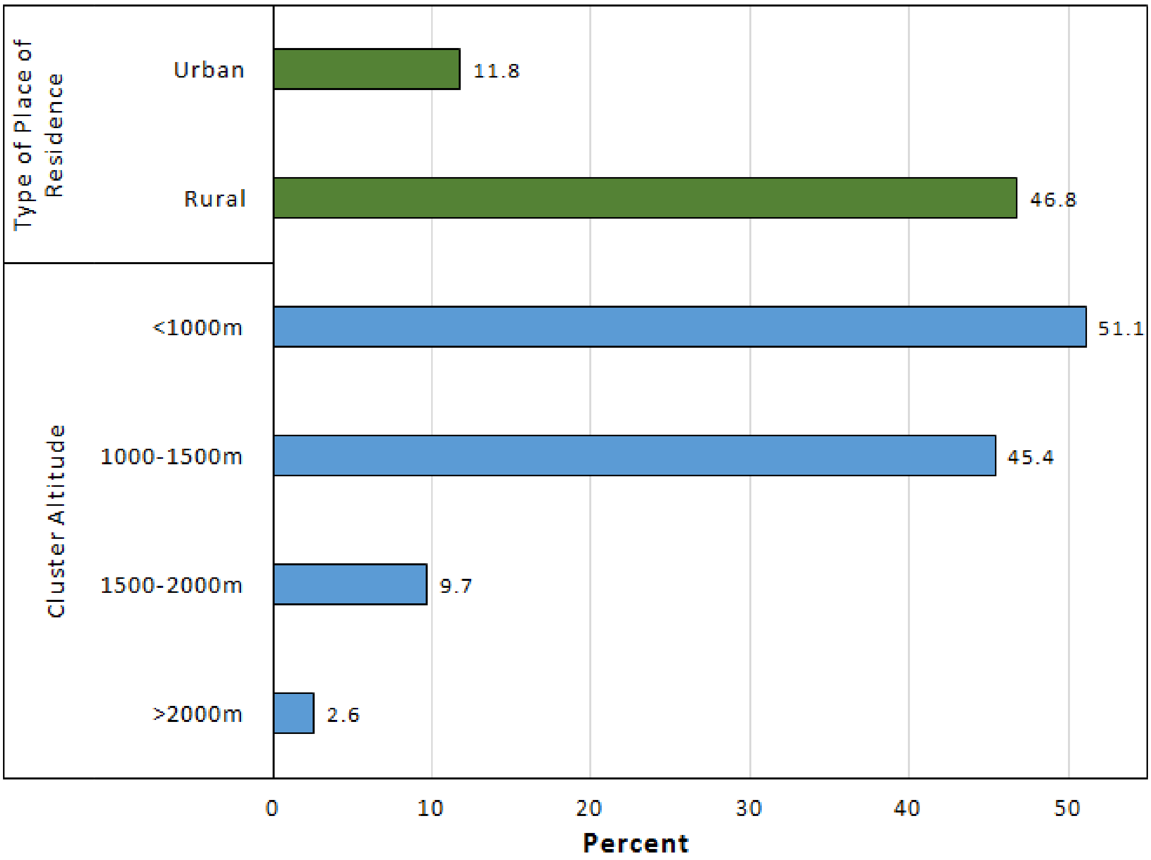

Figure 1.5: Observed prevalence of malaria according to type of place of residence and cluster altitude.

As seen in Figure 1.5 on the previous page, the observed prevalence was much higher within rural areas compared to urban areas. Furthermore, the figure shows there was a decrease in malaria prevalence as altitude increased, which is to be expected as malaria transmission decreases as altitude increases. However, even at the higher altitudes of 1500m and above, children tested positive for malaria. The sampled households at these altitudes consisted of only rural areas, as seen in Table 1.2, thus suggesting a rural place of residence may be associated with a higher risk of malaria.

Figure 1.6 below reveals that households having access to electricity, a television or a refrigerator was a rare event, however having access to a radio or bicycle was more common. With only 7.7% of the households in the sample having had access to electricity, shows the extent of how rural Uganda was at the time of this survey.

Out of the 3.2% of the households that had refrigerators, 1% did not have electricity. Therefore, these refrigerators could possibly have been powered by gas.

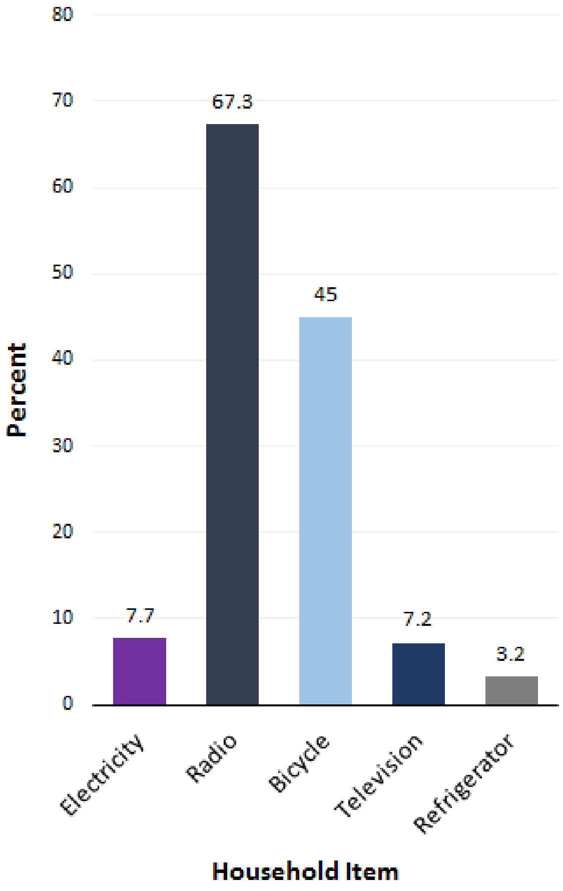

Figure 1.6: Percentage of households with access to certain household items.

Table 1.3 on the next page also shows that a significantly higher percentage of households within urban areas had access to electricity, a television or a refrigerator compared to households within rural areas, thus revealing access to these items are associated with a higher socio-economic status. Whereas there was a higher percentage of households within rural areas that had access to a bicycle, which may have been used as a mode of transport rather than as a luxury. Ownership of a radio was common in both rural and urban areas.

Table 1.3: Percentage of households within each type of place of residence with access to certain household items.

Household	Type of place of Residence	
Item	Urban	Rural
Electricity	43.9	3.3
Radio	79.3	65.9
Bicycle	29.3	46.9
Television	37.4	3.5
Refrigerator	15.3	1.7

Figure 1.7 on the next page reveals that the prevalence of malaria within the group of children who resided in households with electricity was much lower compared to those in households without electricity. Similarly, the prevalence amongst those in households with a refrigerator or television was much lower than those in households without these items. Thus, suggesting a higher socio-economic status may be associated with a lower risk of malaria. This can also be observed by the higher prevalence of malaria amongst those in households with a bicycle compared to those in households without a bicycle.

Table 1.4 on page 18 shows how the households in the sample are distributed according to the socio-economic variables: source of drinking water, toilet facility and main floor material, main wall material and main roof material of the house. The majority of the households in the sample (59.1%) obtained their drinking water from a protected source, which included protected wells (private and public), boreholes and protected springs.

Unprotected water sources included open wells (private and public), unprotected springs, rainwater and surface water (rivers/streams, ponds/lakes and dams). Furthermore, only 13.4% of the households obtained their drinking water from a tap, which included public taps or standpipes, water piped into the dwelling and water piped to the yard. The category 'other' represents unspecified water sources and only makes up 0.3% of the sample of households.

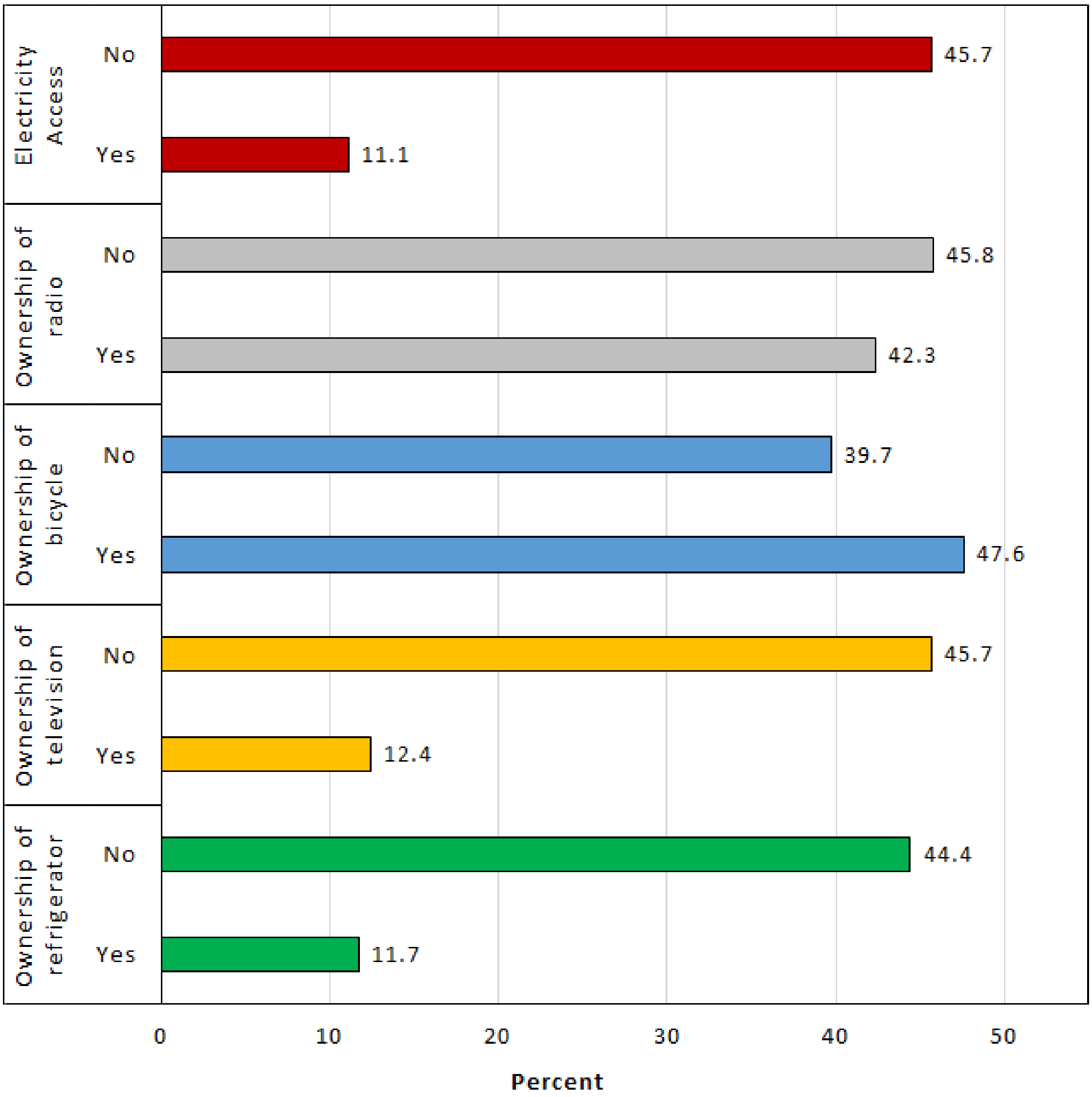

Figure 1.7: Observed prevalence of malaria according to electricity access and ownership of certain household items.

Table 1.4: Percentage of households in the sample according to source of drinking water, toilet facility and main material for construction of house.

Source of Drinking Water	Unprotected Water	27.2
	Protected Water	59.1
	Tap Water	13.4
	Other	0.3
Toilet Facility	No Facility	11
	Uncovered Pit Latrine	22.4
	Covered Pit Latrine	61.5
	VIP Latrine	4
	Flush Toilet	0.9
	Other	0.3
Main Floor Material	Earth/Sand	36.1
	Earth and Dung	40.6
	Cement	22
	Other	1.3
Main Wall Material	Thatch/Straw	1
	Mud and Poles	35.8
	Unburnt Bricks	27.7
	Burnt Bricks	33.3
	Cement Blocks	1.2
	Other	1
Main Roof Material	Thatch	43.6
	Iron Sheets	55
	Tiles	0.6
	Other	0.8

The most common type of toilet facility was a pit latrine with 61.5% and 22.4% of the households using a covered and uncovered pit latrine respectively, both of which included those with and without cement slabs. Only 4% of the households used a VIP (ventilated improved) pit latrine which is an improvement to the simple pit latrines in order to overcome their disadvantages. Just less than 1% of the households had a flush toilet whereas 11% had no toilet facility.

The most commonly used material for a household's floor was a mixture of earth (sand) and dung at 40.6% of households in the sample followed by only earth at 36.1%. There were three main materials used for the walls of the households, with a combination of mud and poles being the most commonly used material at 35.8% of the households, followed by burnt bricks at 33.3% and unburnt bricks at 27.7%. Very few households used cement blocks for their walls (1.2%). The two main materials used for the roofs was iron sheets (55%) and thatch (43.6%) with only 0.6% of households using tiles.

Figure 1.8 on page 20 represents the observed malaria prevalence for each of the different types of socio-economic variables discussed in Table 1.4 above. The prevalence amongst those children using unprotected water sources (46.6%) and protected water sources (47.8%) for drinking water was not much different, however the prevalence was a lot lower amongst those using tap water (14.8%). Those using uncovered pit latrines as their toilet facility had the highest prevalence at 52.6%, followed by those with no toilet facilities (48.5%) and those using covered pit latrines (41.1%). There was no prevalence of malaria amongst those children in households with flush toilets, however these households only made up 0.9% of the sample. Those with unspecified toilet facilities ('other') had a prevalence of 60%, however this category too made up a very small percentage of the households in the sample (0.3%).

More than half (50.5%) of the children living in households with just earth (sand) as the main floor material tested positive for malaria. The prevalence amongst those living in households with earth and dung as the main floor material was 47.3%, not much lower than those living in households with just earth. Only 23.8% of those living in households with cement as the main floor material tested positive for malaria, which is a much lower prevalence than the other materials. The prevalence amongst those living in households with thatch/straw as the main wall material was 59.6%, however these households only made up 1% of the sample. The three most commonly used main wall materials; mud and poles, unburnt bricks and burnt bricks, had prevalences ranging from 38.9% to 51.4%, with unburnt bricks having the highest prevalence.

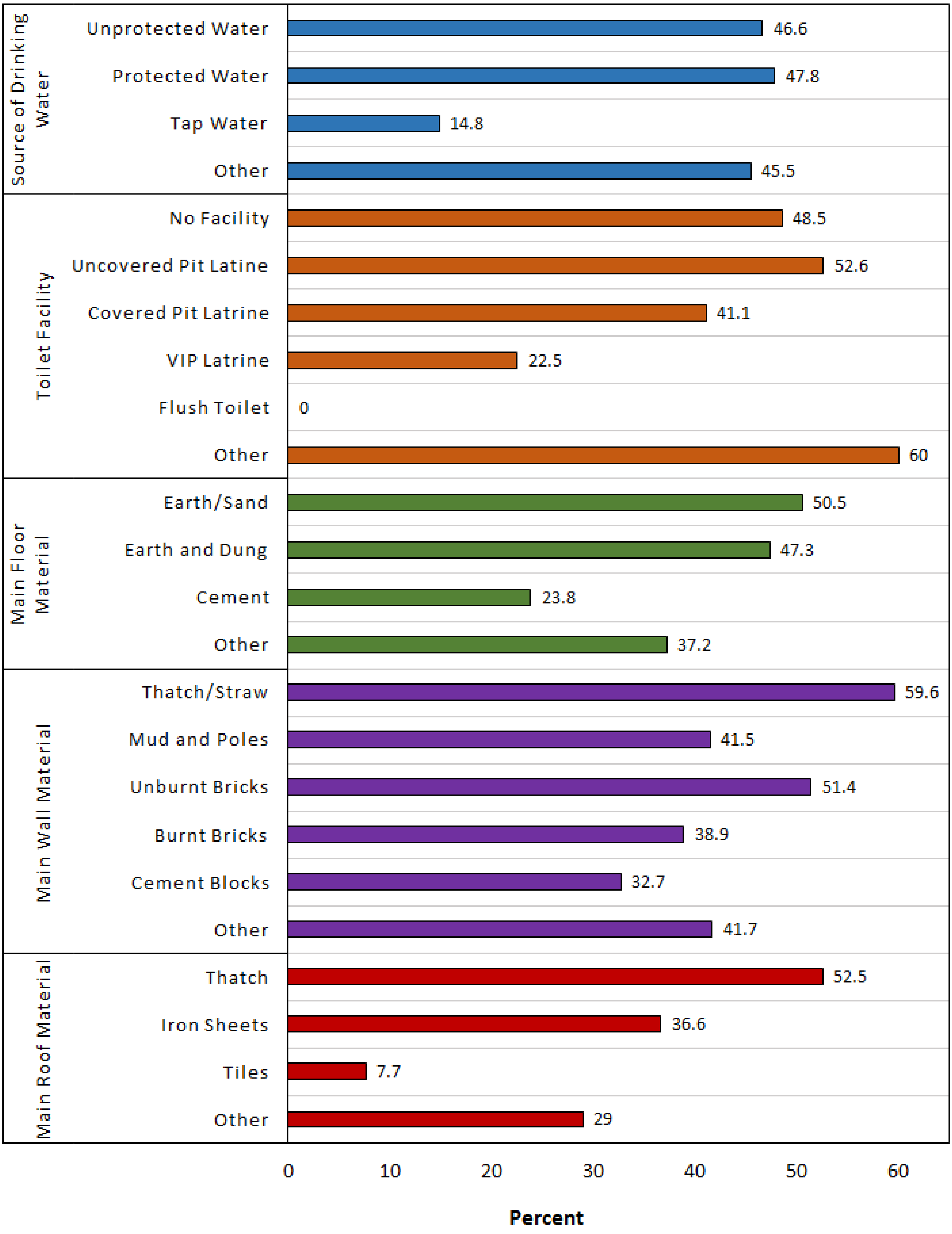

Figure 1.8: Observed malaria prevalence according to source of water, toilet facility and main floor, wall and roof material of the house.

Out of the children living in households with thatch as the main roof material, 52.5% tested positive for malaria. Thus, suggesting this material may be a significant risk factor. The prevalence amongst those in households with iron sheets as the main roof material was 36.6% and with tiles was 7.7%, which only made up 0.6% of the sample of households. The categories 'other', which represent unspecified wall, floor and roof materials, had prevalences ranging from 29% to 41.7%, however these categories only made up 1.3%, 1% and 0.8% of the main floor, wall and roof materials of the houses, respectively.

The total number of mosquito nets available in the household was recorded. The maximum number of nets in a household was 7 with a median of 1 and a mean of 1.3 nets. Figure 1.9 below represents the percentage of households corresponding to the number of nets recorded in the household. Over a third of the households in the sample had no nets and very few households had four or more. Whether or not the child in the household that was tested for malaria slept under a mosquito net was recorded. 64.7% of the children in the sample were recorded to sleep under a net.

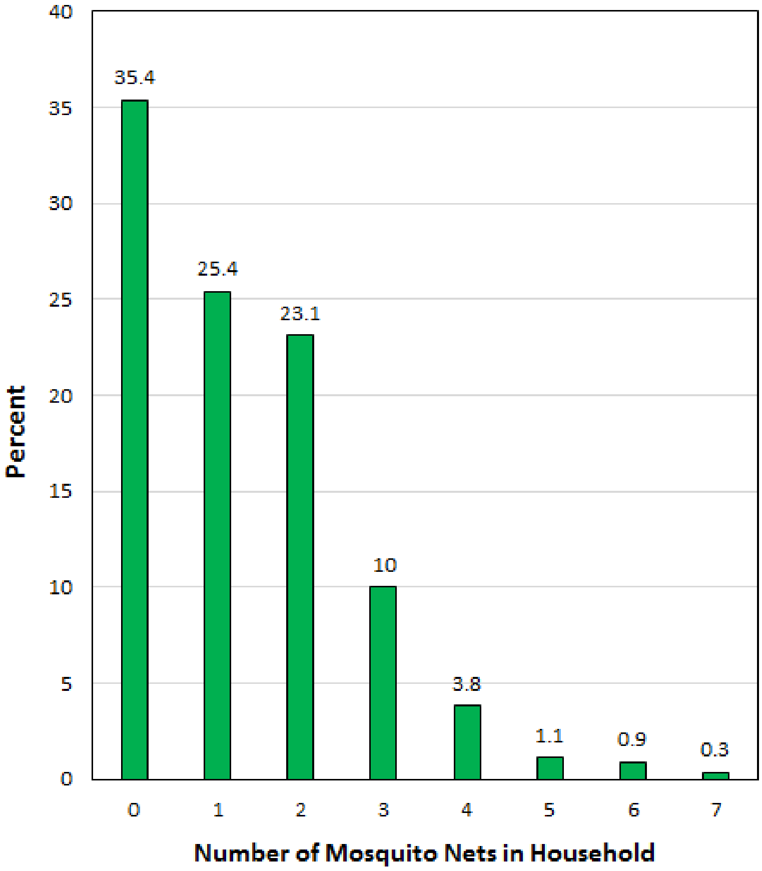

Figure 1.9: Percentage of households according to the total number of mosquito nets in the household.

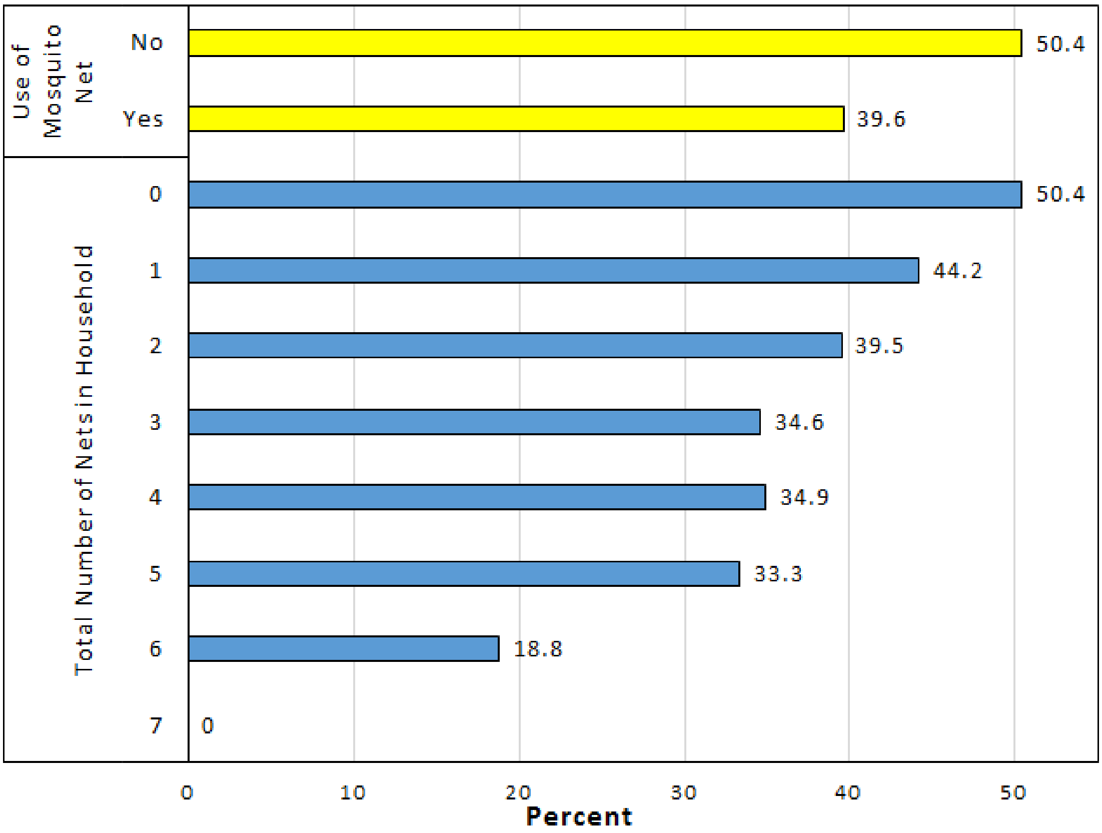

Figure 1.10: Observed prevalence of malaria according to use of mosquito nets and total number of mosquito nets in the household.

As expected, Figure 1.10 reveals that the prevalence of malaria amongst the children who did not sleep under a mosquito net was much greater than those who did sleep under a mosquito net. There was a decreasing trend in the prevalence as the number of mosquito nets in the household increased to a total of three. However, there was a slight increase of 0.3% in prevalence as the total number of mosquito nets increased from three to four, after which the prevalence decreased again. There was no malaria prevalence amongst the children in households with seven mosquito nets, although these households only made 0.3% of the sample.

Information about indoor residual spraying of the interior walls within the last 12 months prior to the survey was collected for each household. 94.6% of the households had no incidences of indoor spraying and 5.1% of the households had been sprayed at least once within the 12 months. 0.3% of the households were recorded to not have any knowledge of the incidence of spraying. Figure 1.11 shows there was a higher prevalence amongst the children in households that had been sprayed within the last 12 months compared to those in households with no incidences of spraying.

However, Table 1.5 reveals that over 62% of the households that had incidences of spraying were from the three regions of Uganda (East Central, Mid Northern and Central) with the highest prevalence. Whereas only 31.4% of the households that had no incidences of spraying were located in these three regions.

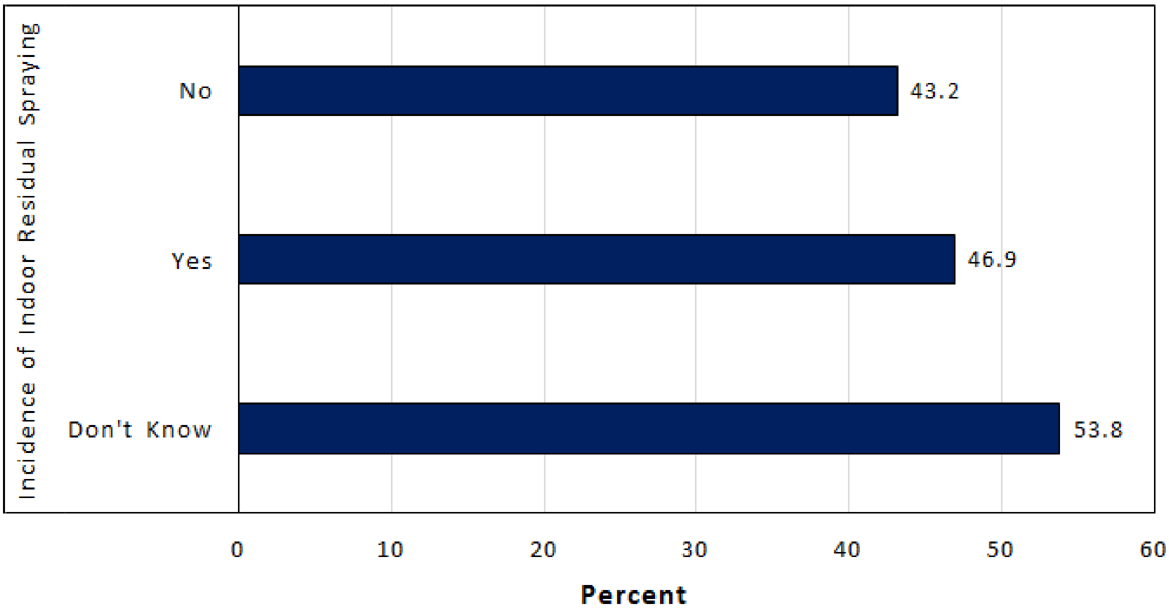

Figure 1.11: Observed prevalence of malaria according to incidence of indoor residual spraying within the last 12 months.

Table 1.5: Distribution of incidence of indoor residual spraying in households across the regions of Uganda.

Region of Uganda	Incidence of Indoor Residual Spraying		
	No	Yes	Don't Know
East Central	12.8%	1.4%	0%
Mid Northern	9.6%	48.8%	23.1%
Central 2	9%	12.6%	15.4%
West Nile	13.5%	0%	15.4%
Mid Eastern	11.2%	1%	0%
Central 1	9.3%	0%	30.8%
Mid Western	11.7%	0%	7.7%
North East	10%	32.4%	0%
South Western	8.2%	1.9%	0%
Kampala	4.7%	1.9%	7.7%

Out of the children tested for malaria, 50.3% were female and 49.6% were male. Figure 1.12 displays the prevalence of malaria for males and females. 43% of the males tested positive and 43.9% of the females tested positive. Thus, the prevalence was not much different between males and females.

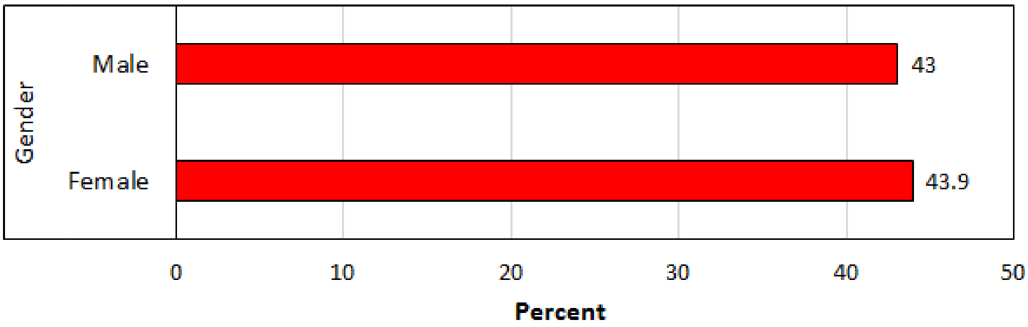

Figure 1.12: Observed prevalence of malaria according to gender.

The percentage of children in the sample within the different age groups, given in Table 1.6, ranged from 8.7% between 0 and 5 months to 20.9% between 36 and 47 months. Only 2.8% of the children in the sample had caregivers of the age 45 to 49 years and the majority had caregivers of the age 20 to 29 years. Figure 1.13 on the following page reveals there was an increase in the prevalence of malaria as the age in months of a child increased. More than half the children in the age groups 36 to 47 months and 48 to 54 months tested positive. The prevalence according to the caregiver's age ranged from 38.6% to 51.4%, with the prevalence within the age groups 20 to 24 years, 25 to 29 years and 30 to 34 years not differing by much. Although the prevalence was highest amongst the group of children who had caregivers aged 40 to 44 years and 45 to 49 years, these groups only made up 8.3% of the sample.

Table 1.6: Distribution of children in the sample according to age and caregiver's age.

Age in Months of Child	Percent	Age in Years of Caregiver	Percent
0 to 5	8.7	15 to 19	14.7
6 to 11	10	20 to 24	25.9
12 to 17	10.6	25 to 29	24.3
18 to 23	9.4	30 to 34	16.4
24 to 35	20	35 to 39	10.4
36 to 47	20.9	40 to 44	5.5
48 to 54	20.3	45 to 49	2.8

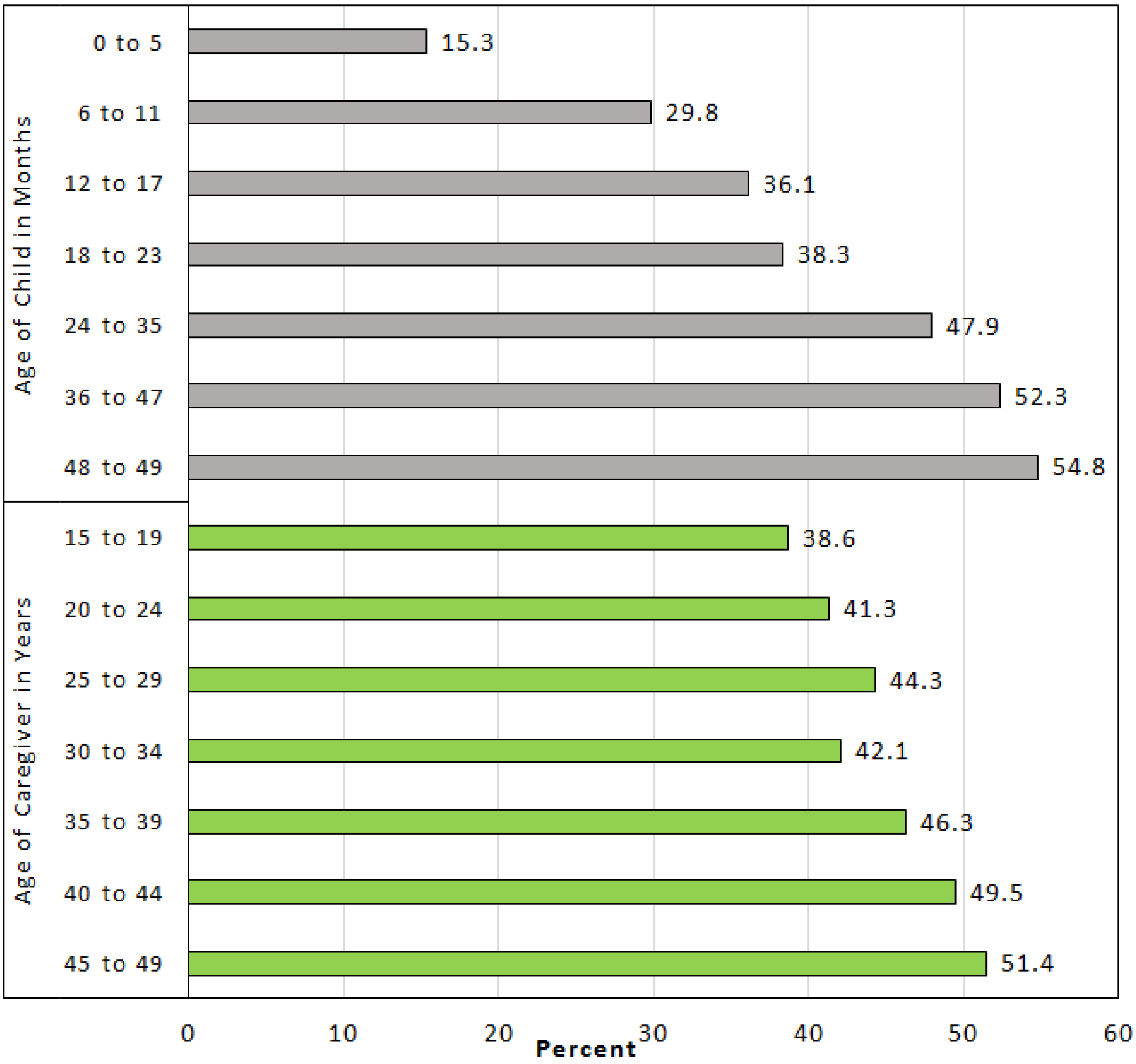

Figure 1.13: Observed prevalence of malaria according to age of child and caregiver.

A total of 84.2% of the children in the sample had caregivers who knew mosquito bites can cause malaria, and 85.8% had caregivers who knew there are ways of preventing malaria. Figure 1.14 below shows the prevalence of malaria according to the caregiver's knowledge of malaria. The prevalence was highest amongst the children whose caregivers did not believe malaria can be caused by mosquito bites. The prevalence was also highest amongst the children whose caregivers did not believe there are ways of preventing malaria. With more than half the children in these two groups testing positive for malaria, this may suggest the caregiver's inadequate knowledge of malaria may be a significant risk factor.

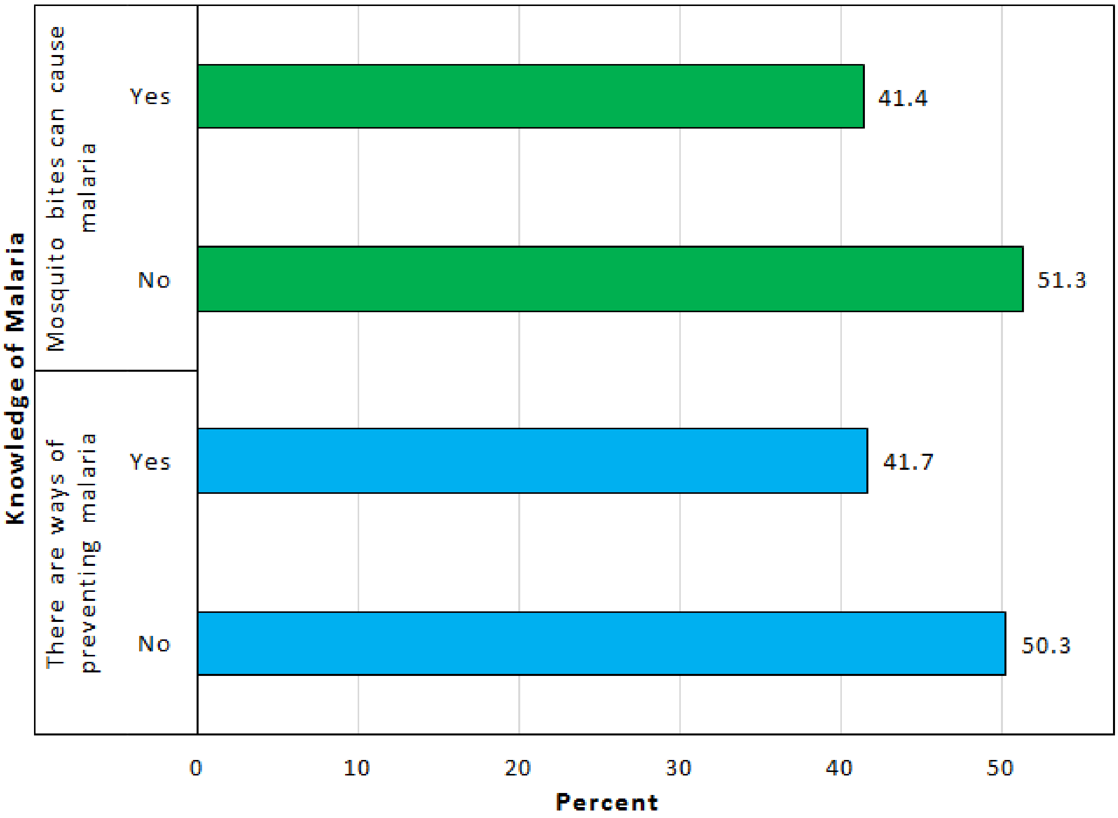

Figure 1.14: Observed prevalence of malaria according to the caregiver's knowledge of malaria.

Table 1.7 on the next page shows the distribution of the children in the sample according to the highest education level of their caregiver. This table also shows the distribution of education level for each type of place of residence (urban and rural) as well as for malaria result. The majority of the children had caregivers with only primary school level and only 1.4% had caregiver's with an education higher than secondary school. None of the children in the sample from urban areas whose caregivers had higher education tested positive for malaria, however only 0.5% of the children in the sample were from urban areas and had caregivers with higher education. Almost half the children in the sample from rural areas with caregivers who only had primary education tested positive for malaria (27.1% from a total of 57.4%). Similarly, more than half of those in rural areas who had caregivers with no education tested positive (10.9% from a total of 20.9%). Thus, suggesting these children were most at risk for malaria.

Figure 1.15 on the next page reveals that the prevalence decreased as the education level of the child's caregiver increased, which is what one would expect. Just over half (50.6%) of the children whose caregivers had no education and just under half (45.2%) of those with caregivers who only had primary education tested positive.

Table 1.7: Percentage of children in the sample according to caregiver's highest education level, type of place of residence and malaria result.

Education Level	Type of Place of Residence						Total
	Urban			Rural			
	Negative	Positive	Total	Negative	Positive	Total	
No Education	0.8	0.2	1	10	10.9	20.9	21.9
Primary	3.4	0.7	4.1	30.3	27.1	57.4	61.5
Secondary	4.1	0.3	4.4	7.2	3.6	10.8	15.2
Higher	0.5	0	0.5	0.7	0.2	0.9	1.4

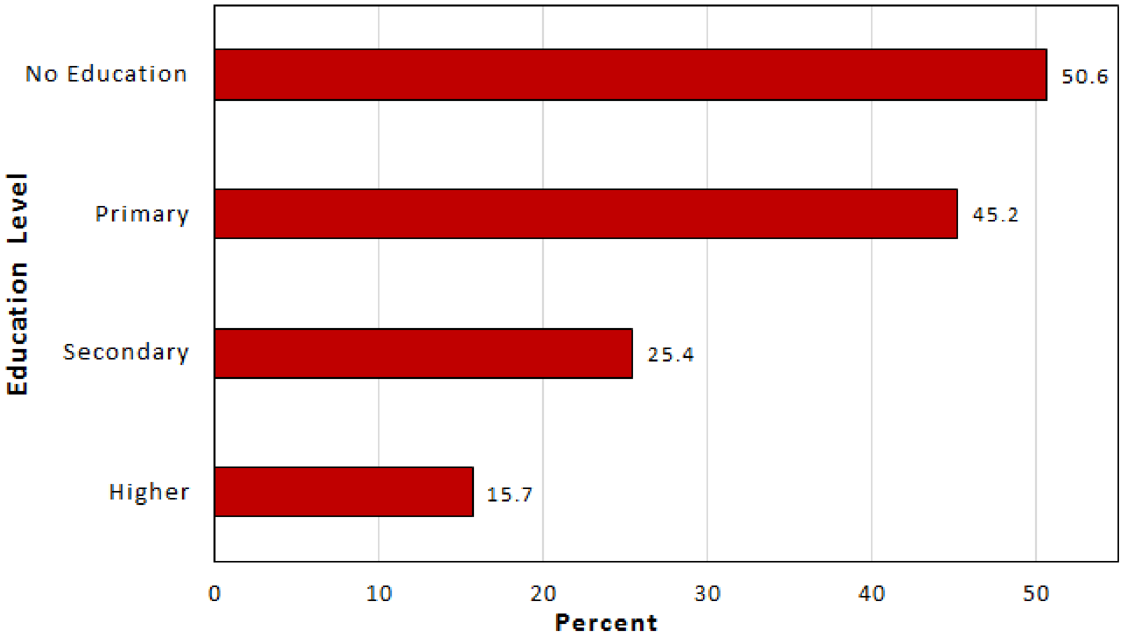

Figure 1.15: Observed prevalence of malaria according to the caregiver's highest education level.

The distribution of the children in the sample according to the number of members within their households is shown in Table 1.8 below. The majority of the sample (55.2%) had between 6 and 10 members within their households. Figure 1.16 on the next page reveals that the prevalence amongst the different groups of children according to the number of members within their households did not differ significantly, with the prevalence ranging from 38.1% to 44%. However, as seen in Table 1.8, only 6.6% of the children had more than 10 members within their households.

Table 1.8: Distribution of the sample according to the number of members in each household.

Number of Members	Percent
1 - 5	38.2
6 - 10	55.2
11 - 15	6.1
16 - 20	0.5

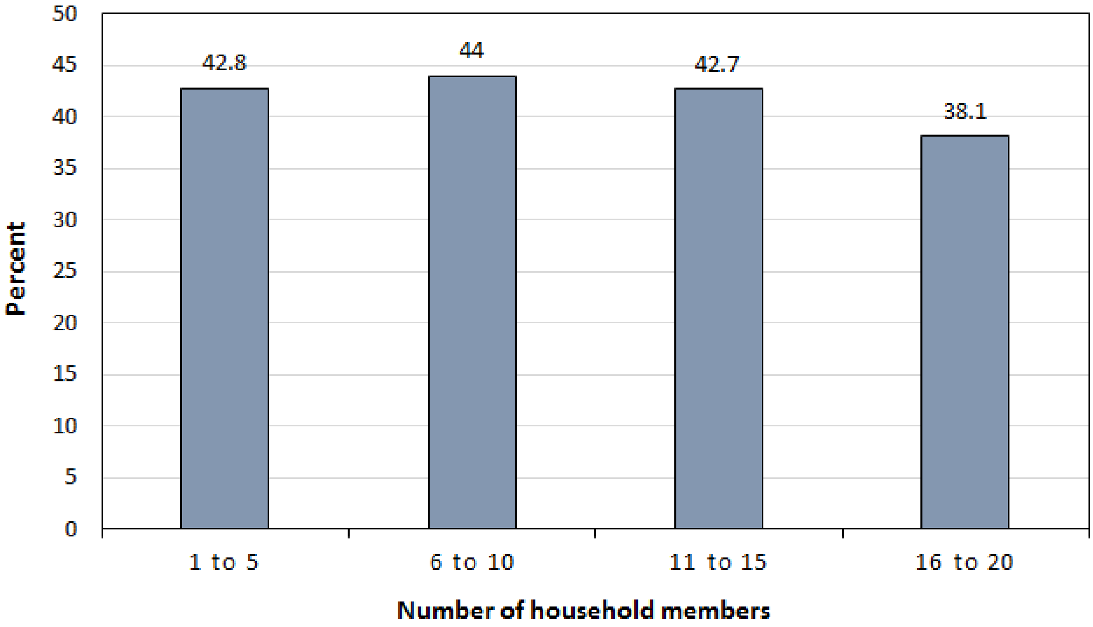

Figure 1.16: Observed prevalence of malaria according to the number of household members.

The variables time taken to collect water, proximity to vector-breeding sites and total number of rooms per household have been shown to be significant risk factors for malaria (Ayele et al., 2013; Peterson et al., 2009), however, these variables were not recorded in this MIS survey. Whether or not an insecticide-treated net (ITN) was used has also been shown to be a significant risk factor, where its use is known to be highly effective in reducing malaria morbidity and mortality (Atieli et al., 2011; Nevill et al., 1996; Binka et al., 1996). However, this variable was missing from 58.9% of the MIS data for this thesis.

In the next chapter, a review of linear models is given, where some of the theory of general linear models and linear mixed models is discussed.

Chapter 2

Linear Models

Linear models are some of the most widely used statistical techniques to analyze data sets, where such models can be used to test almost any hypothesis concerning a response (dependent) variable and, or, the independent/explanatory variables (Miller & Haden, 2006). The selection of an appropriate model is dependent on the scale of measurements of the variables in the data set, particularly the response variable.

2.1 General Linear Models

In this chapter we start by considering the general linear model (LM). The LM models a continuous response variable Y_i, for $i = 1, ..., n$, in terms of a linear combination of its corresponding explanatory variables $x_{ij},\ j = 1, ..., p$. The LM assumes the Y_i are independent and follow a normal distribution with a constant variance. The explanatory variables may be continuous or categorical.
The linear model for such variables is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i \tag{2.1}$$

where $\beta_1, \ldots, \beta_p$ are the regression coefficients and ϵ_i is the error for the i^{th} observation.

Equation (2.1) can also be written as follows:

$$Y_i = \boldsymbol{x}_i' \boldsymbol{\beta} + \epsilon_i$$

where $\boldsymbol{x}_i' = (1, x_{i1}, \ldots, x_{ip})$ and $\boldsymbol{\beta'} = (\beta_0, \beta_1, \ldots, \beta_p)$

In matrix form the model for all the observations is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.2}$$

where $\mathbf{y} = [Y_1, Y_2, \ldots, Y_n]'$ is an $n \times 1$ vector of response variables, $\mathbf{X}$ is the $n \times (p+1)$ design matrix, $\boldsymbol{\beta}$ is a $(p+1) \times 1$ vector of parameters and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of errors. The errors are assumed to be independently and identically normally distributed with mean 0 and variance σ^2. In other words, $\boldsymbol{\epsilon} \sim \boldsymbol{N}(\mathbf{0}, \sigma^2\boldsymbol{I})$ where $\boldsymbol{I}$ is an $n \times n$ identity matrix.
Therefore,

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$$

and

$$\begin{aligned} Var(\mathbf{y}) &= Var(\boldsymbol{\epsilon}) \\ &= \sigma^2\boldsymbol{I} \end{aligned}$$

Thus, the response variables are independently distributed and $\mathbf{y} \sim \boldsymbol{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$. The fitted model for the data is given by

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$$

where $\widehat{\boldsymbol{\beta}}$ is the estimate of parameter $\boldsymbol{\beta}$.

The value of $\widehat{\boldsymbol{\beta}}$ can be found using two commonly used methods: the method of ordinary least squares and the maximum likelihood method. The method of least squares is based on minimizing the residual error $\widehat{\epsilon_i} = y_i - \widehat{y_i}$, the difference between the observed and fitted values. This is done by minimizing the error sums of squares

$$\begin{aligned} \sum_{i=1}^{n} \widehat{\epsilon}_i^2 &= \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 \\ &= \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})^2 \end{aligned}$$

or equivalently

$$\widehat{\boldsymbol{\epsilon}}'\,\widehat{\boldsymbol{\epsilon}} = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \tag{2.3}$$

The method of maximum likelihood (ML) chooses the values of the parameters that are most consistent with the sample data under the assumption $\boldsymbol{\epsilon} \sim \boldsymbol{N}(\mathbf{0}, \sigma^2\boldsymbol{I})$. This is done by maximizing the likelihood function of $\mathbf{y}$ given below:

$$L(\mathbf{y}) = (2\pi\sigma^2)^{-\frac{n}{2}}\, exp\{-\frac{1}{2\sigma^2}\,(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\}$$

This is equivalent to maximizing the log of the likelihood function:

$$ln\,L(\mathbf{y}) = -\frac{n}{2}\,ln(2\pi) - \frac{n}{2}\,ln(\sigma^2) - \frac{1}{2\sigma^2}\,(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \tag{2.4}$$

The value of $\widehat{\boldsymbol{\beta}}$ that minimizes Equation 2.3 and maximizes Equation 2.4 is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Thus, both methods obtain the same estimate for $\boldsymbol{\beta}$. The mean and variance of this estimate of $\boldsymbol{\beta}$ can be found as follows

$$\begin{aligned} E(\widehat{\boldsymbol{\beta}}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

and

$$\begin{aligned} Var(\widehat{\boldsymbol{\beta}}) &= Var[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\sigma^2 \end{aligned}$$

Thus, $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator for $\boldsymbol{\beta}$. Since $\widehat{\boldsymbol{\beta}}$ is a linear combination of a normally distributed random variable, it follows $\widehat{\boldsymbol{\beta}} \sim \boldsymbol{N}(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2)$. Therefore, knowing the distributional properties of this parameter estimator will allow one to perform hypothesis tests to determine which independent variables are significant.

The method of maximum likelihood can also be used to estimate σ^2, the variance of the error term. This can be done by minimizing Equation 2.4 with respect to σ^2 where it can easily be shown that the ML estimate of σ^2 is given by

$$\widehat{\sigma}^2 = \frac{(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})}{n} \tag{2.5}$$

This value of $\widehat{\sigma}^2$ can therefore be used to determine the variance, and thus the standard errors, of the parameter estimate $\widehat{\boldsymbol{\beta}}$.

2.2 Linear Mixed Models

The LM assumes the different levels of the independent variables or factors are fixed. This model is only applicable when interest is on the effect of the specific factor levels included in the model (Kutner et al., 2005). However, when the factor levels represent a sample from a larger population of potential factor levels, and inferences are to be made on the whole population of the levels, then these factors included in the model are no longer considered fixed, but rather random, and therefore the LM needs to be extended to represent this. When fixed effects and random effects are included in the model, the resulting model is referred to as a linear mixed model (LMM), sometimes also referred to as a linear mixed effect model. Random effects are used to model the random variation in the dependent variable at different levels of the factors (West et al., 2007) and can be used to represent unobserved effects or characteristics influencing the pattern of responses of an individual when measures on that individual are being repeated (Der & Everitt, 2006). LMMs are often utilized in the modeling of hierarchical or multilevel data, where observations can be placed in levels of hierarchy in the data (Ker, 2014). Such data include clustered, repeated-measures and longitudinal data where observations within the same cluster or from the same individual tend to be more homogeneous with one another than those from another cluster or individual. Thus, these observations can no longer be treated as independent. The inclusion of a random effect in the model allows the correlation structure of the observations to be modeled.

LMMs may be expressed in different but equivalent forms (Fox, 2002). When modeling hierarchical data, the common form of the LMM is

$$Y_{ij} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \ldots + \beta_p x_{ijp} + \gamma_{i1} z_{ij1} + \ldots + \gamma_{iq} z_{ijq} + \epsilon_{ij} \qquad (2.6)$$

where

- Y_{ij} is the value of the response variable for the j^{th} of n_i observations in the i^{th} of m clusters or individuals.
- $\beta_1, \ldots, \beta_p$ are the fixed effect coefficients, which are common for all clusters or individuals.
- $x_{ij1} \ldots x_{ijp}$ are the p fixed effect regressors for observation j in cluster/individual i.
- $\gamma_{i1} \ldots \gamma_{iq}$ are the random effect coefficients for cluster/individual i.
- $z_{ij1} \ldots z_{ijq}$ are the q random effect regressors for cluster/individual i.
- ϵ_{ij} is the error for observation j in cluster/individual i.

Alternatively, Equation 2.6 can be expressed in the form

$$\boldsymbol{y}_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i$$

Or more compactly, in matrix form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \tag{2.7}$$

where

- $\mathbf{y}$ is an $n \times 1$ vector of response variables, where $\sum_{i=1}^{m} n_i = n$ is the total number of observations.
- $\mathbf{X}$ is the $n \times (p+1)$ design matrix for the fixed-effects.
- $\boldsymbol{\beta}$ is a $(p+1) \times 1$ vector of fixed effect coefficients.
- $\mathbf{Z}$ is the $n \times q$ design matrix for the random effects.
- $\boldsymbol{\gamma}$ is a $q \times 1$ vector of random effect coefficients.
- $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of errors.

The random effects are random variables. Thus, there are two random effects in the LMM with the following assumptions:

$$\boldsymbol{\epsilon} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{R}_{n\times n})$$

$$\boldsymbol{\gamma} \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{G}_{q\times q})$$

and

$$Cov(\boldsymbol{\epsilon}, \boldsymbol{\gamma}) = \mathbf{0}_{n\times q}$$

Therefore

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$$

and

$$\begin{aligned} Var(\mathbf{y}) &= Var(\mathbf{Z}\boldsymbol{\gamma}) + Var(\boldsymbol{\epsilon}) \\ &= \mathbf{Z}\boldsymbol{G}\mathbf{Z}' + \boldsymbol{R} \\ &= \boldsymbol{V} \text{ (say)} \end{aligned}$$

Which results in $\mathbf{y} \sim \boldsymbol{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{V})$.

For LMMs, one could be interested in the following:

1. Making inferences about the fixed effects, thus estimating and performing hypothesis tests on $\boldsymbol{\beta}$.
2. Making inferences about the random effect's variance, thus estimating the variance components of $\boldsymbol{V}$.
3. Predicting the realization of the random effects.

Estimating $\boldsymbol{\beta}$:

There are many methods used to estimate $\boldsymbol{\beta}$, however only the maximum likelihood (ML) method will be considered.

Similar to solving for $\boldsymbol{\beta}$ in the LM, the ML method solves for the value of $\boldsymbol{\beta}$ which maximizes the log-likelihood function, and thus the likelihood function, of $\mathbf{y}$. For the LMM with $\mathbf{y} \sim \boldsymbol{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{V})$, the log-likelihood function is

$$ln\, L(\mathbf{y}) = -\frac{n}{2}\, ln(2\pi) - \frac{1}{2}\, ln(\boldsymbol{V}) - \frac{1}{2}\,(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\,\boldsymbol{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{2.8}$$

Maximizing Equation 2.8 with respect to $\boldsymbol{\beta}$ will result in the following ML estimate

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\,\boldsymbol{V}^{-1}\,\mathbf{X})^{-1}\mathbf{X}'\,\boldsymbol{V}^{-1}\,\mathbf{y} \tag{2.9}$$

Thus, using the same procedure as that from the previous section for the LM to find the mean, variance and distribution of $\widehat{\boldsymbol{\beta}}$, it follows $\widehat{\boldsymbol{\beta}} \sim \boldsymbol{N}(\boldsymbol{\beta}, (\mathbf{X}'\boldsymbol{V}^{-1}\mathbf{X})^{-1})$.

The value of $\widehat{\boldsymbol{\beta}}$, and the variance of $\widehat{\boldsymbol{\beta}}$, requires the value of $\boldsymbol{V}$ to be known, thus the variances of $\boldsymbol{\epsilon}$ and $\boldsymbol{\gamma}$, $\boldsymbol{R}$ and $\boldsymbol{G}$ respectively, need be known beforehand. In practice, these variances are usually unknown and therefore need to be estimated. Thus, when $\boldsymbol{V}$ is unknown, we can replace it by its estimate $\widehat{\boldsymbol{V}}$. The ML estimate of $\boldsymbol{\beta}$ will then become

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\,\widehat{\boldsymbol{V}}^{-1}\,\mathbf{X})^{-1}\mathbf{X}'\,\widehat{\boldsymbol{V}}^{-1}\mathbf{y} \tag{2.10}$$

with

$$Var(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}'\widehat{\boldsymbol{V}}^{-1}\mathbf{X})^{-1} \tag{2.11}$$

Estimating $\boldsymbol{V}$:

Let $\boldsymbol{\theta}$ be a vector of unknown variance components in $\boldsymbol{V}$ to be estimated. The ML method can be used to estimate these variance components by finding the value of $\boldsymbol{\theta}$ in $\boldsymbol{V}$ that maximizes the log-likelihood function in Equation 2.8. If $\boldsymbol{\beta}$ is unknown, it can be replaced by its estimate $\widehat{\boldsymbol{\beta}}$ from Equation 2.9. This will result in the following log-likelihood function:

$$\ell_p = -\frac{n}{2}\,ln(2\pi) - \frac{1}{2}\,ln(\boldsymbol{V}) - \frac{1}{2}\,\mathbf{y}'\mathbf{P}\mathbf{y} \tag{2.12}$$

where $\mathbf{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\mathbf{X}\,(\mathbf{X}'\,\boldsymbol{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\,\boldsymbol{V}^{-1}$

Equation 2.12 is known as a profile log-likelihood function. Maximization of this equation with respect to $\boldsymbol{\theta}$ will result in a non-linear optimization (West et al., 2007). Thus, iterative procedures such as Newton Raphson and Fisher Score (also commonly referred to as Fisher Scoring) are required to solve for the estimate of $\boldsymbol{\theta}$ that maximizes this profile log-likelihood function.

2.2.1 Newton Raphson

The iterative equation for Newton Raphson is given by

$$\widehat{\boldsymbol{\theta}}^{(t+1)} = \widehat{\boldsymbol{\theta}}^{(t)} - (\boldsymbol{H}^{(t)})^{-1}\,\boldsymbol{U}^{(t)} \tag{2.13}$$

where $\widehat{\boldsymbol{\theta}}^{(t)}$ is the approximation of $\boldsymbol{\theta}$ at the t^{th} iteration. $\boldsymbol{U}^{(t)} = \dfrac{\partial \ell_p}{\partial \boldsymbol{\theta}}$ evaluated at $\widehat{\boldsymbol{\theta}}^{(t)}$, where $\boldsymbol{U}$ is called the score. $\boldsymbol{H}^{(t)}$ is the Hessian matrix, $\boldsymbol{H}$, with the following elements evaluated at $\widehat{\boldsymbol{\theta}}^{(t)}$:

$$\boldsymbol{H}_{jk} = \frac{\partial^2 \ell_p}{\partial\theta_j\,\partial\theta_k} \tag{2.14}$$

2.2.2 Fisher Score

The Fisher Score iterative equation is given by

$$\widehat{\boldsymbol{\theta}}^{(t+1)} = \widehat{\boldsymbol{\theta}}^{(t)} + (\boldsymbol{\mathcal{I}}^{(t)})^{-1}\,\boldsymbol{U}^{(t)} \tag{2.15}$$

where $\boldsymbol{\mathcal{I}} = -E(\boldsymbol{H})$ is referred to as the information matrix.

Both these iterative methods require an appropriate starting value $\widehat{\boldsymbol{\theta}}^{(0)}$, whereafter the process will continue until the algorithm converges. That is, until the difference between successive approximations is negligible.

This maximum likelihood method of estimating $\boldsymbol{\theta}$ does not take into consideration the loss of degrees of freedom from estimating the fixed effect parameters in $\boldsymbol{\beta}$, and thus the ML estimate of $\boldsymbol{\theta}$ is biased (West et al., 2007). The restricted maximum likelihood (REML) method, first introduced by Patterson & Thompson (1971), is a modification of the ML method and can be used as an alternative for finding the estimate of $\boldsymbol{\theta}$. This method takes into consideration the loss of degrees of freedom from the estimation of $\boldsymbol{\beta}$, and thus produces an unbiased estimate of $\boldsymbol{\theta}$.
The REML profile log-likelihood function is

$$\ell_{reml} = \ell_p - \frac{1}{2}\, ln|\mathbf{X}'\,\boldsymbol{V}^{-1}\mathbf{X}| + \frac{rank(\mathbf{X})}{2}\, ln(2\pi) \tag{2.16}$$

where ℓ_p is the profile log-likelihood function given in Equation 2.12 and $|\mathbf{X}'\,\boldsymbol{V}^{-1}\mathbf{X}|$ is the determinant of matrix $\mathbf{X}'\,\boldsymbol{V}^{-1}\mathbf{X}$.

Once again, this results in a non-linear optimization with respect to $\boldsymbol{\theta}$, thus iterative procedures are required to solve for its estimate. The Newton Raphson Iterative Equation 2.13 and Fisher Score Iterative Equation 2.15 can be used where the score $\boldsymbol{U}$ and Hessian matrix $\boldsymbol{H}$ can be found by replacing ℓ_p, the profile log-likelihood function, by ℓ_{reml}, the REML profile log-likelihood function. Once we have obtained the REML estimate $\widehat{\boldsymbol{\theta}}$ in order to obtain $\widehat{\boldsymbol{V}}$, we can calculate $\widehat{\boldsymbol{\beta}}$ by substituting the estimate $\widehat{\boldsymbol{V}}$ into Equation 2.10.

The next chapter discusses some of the theory of generalized linear models and introduces the survey logistic regression model to be applied to the MIS data set.

Chapter 3

Generalized Linear Models

As discussed in Chapter 1, the LM assumes the response variable is continuous and follows a normal distribution. Suppose instead, the response variable represents a count or a binary outcome and thus is discrete. It can no longer be modeled using the LM as the predictions using this model can fall outside the range of the response variable. A possible method of modeling a discrete/categorical outcome is the use of generalized linear models (GLMs) first introduced by Nelder & Wedderburn (1972). GLMs model response variables with non-normal distributions through a transformation called the link function (Nelder & Wedderburn, 1972).

3.1 The GLM Model

The GLM assumes the response variable $Y_i, i = 1, \ldots, n$, follows a distribution that belongs to the exponential family with the following general form

$$f(y_i; \theta_i, \phi) = exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\} \tag{3.1}$$

where θ_i is referred to as a natural or canonical parameter and $a_i(\phi)$, $b(\theta_i)$ and $c(y_i, \phi)$ are known functions. $a_i(\phi)$ has the form $a_i(\phi) = \phi/w_i$, where w_i is a known weight depending on whether the data is grouped and ϕ is referred to as the dispersion or scale parameter. It can be shown that if a response Y_i has a distribution belonging to the exponential family, then its mean and variance are

$$E(Y_i) = \mu_i = b'(\theta_i) \tag{3.2}$$

$$Var(Y_i) = a_i(\phi)\, b''(\theta_i) \tag{3.3}$$

where $b'(\theta_i)$ and $b''(\theta_i)$ are the first and second derivatives of $b(\theta_i)$ with respect to θ_i. $b''(\theta_i)$ is a function of the mean, thus it is referred to as the variance function denoted by $v(\mu_i)$.

Therefore, Equation 3.3 can be expressed in the form

$$\begin{aligned} Var(Y_i) &= a_i(\phi)\, v(\mu_i) && (3.4) \\ &= \frac{\phi}{w_i}\, v(\mu_i) \qquad \text{since } \; a_i(\phi) = \phi / w_i && (3.5) \end{aligned}$$

Thus, another property of the GLM is that of a non-constant variance where the variance may vary across the responses. When $a_i(\phi) > 1$ the model is said to be overdispersed since $Var(Y_i) > v(\mu_i)$. Similarly, the model will be underdispersed when $a_i(\phi) < 1$. Therefore, standard errors calculated on the assumption $a_i(\phi) = 1$ would be incorrect when $a_i(\phi) \neq 1$.

The GLM is specified by the following three components:

- *The Random Component:*
  This consists of the response variable Y_i belonging to the exponential family with probability distribution in the form given in Equation 3.1. The observations of Y are assumed to be independent.

- *The Systematic Component:*
  This component of the GLM relates a linear predictor $\boldsymbol{\eta}$ to the explanatory variables through a linear model as follows

  $$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

- *The Link Function:*
  The expected value or mean of the random component, $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)'$, and the systematic component of the GLM are connected through a link function

  $$\boldsymbol{\eta} = g(\boldsymbol{\mu})$$

  where g is a monotone, differentiable function. This link function $g(\boldsymbol{\mu})$ is invertible where its inverse is often referred to as the mean function

  $$g^{-1}(\boldsymbol{\eta}) = \boldsymbol{\mu}$$

  The canonical link function is that function that makes the linear predictor $\boldsymbol{\eta}$ the same as the canonical parameter $\boldsymbol{\theta}$. Therefore, function g such that $g(\boldsymbol{\mu}) = \boldsymbol{\theta}$ is called the canonical link function.

Many distributions belong to the exponential family, such distributions include the Binomial, Poisson, Gamma and Chi-Square distribution. The Normal distribution also belongs to the exponential family, thus the LM is a class of GLMs where $\boldsymbol{\eta} = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$. Here the canonical link is referred to as the identity link since $E(\mathbf{y}) = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$.

3.1.1 Parameter Estimation

Once again, the ML method can be used for the parameter estimation in GLMs. Due to advances in statistical theory and computer software, this method of estimation has become the most popular technique in applied statistics (Wu, 2005). The log-likelihood function for a single observation is given by

$$\ell_i = ln\, f(y_i; \theta_i, \phi) = \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \tag{3.6}$$

Since $Y_i, i = 1, \ldots, n$, are independent, the joint log-likelihood function is

$$\ell(\boldsymbol{\beta}, \mathbf{y}) = \sum_{i=1}^{n} \ell_i \tag{3.7}$$

The ML estimate of $\beta_j, j = 0, \ldots, p$, is the solution to the score equation

$$\frac{\partial \ell_i}{\partial \beta_j} = 0$$

To obtain this solution, we use the chain rule

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

Using Equation 3.6, we get

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a_i(\phi)}$$

Since $\mu_i = b'(\theta_i),\ Var(Y_i) = a_i(\phi)\, v(\mu_i)$ and $\eta_i = \sum_j \beta_j\, x_{ij}$,

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{y_i - \mu_i}{a_i(\phi)}$$

$$\frac{\partial \mu_i}{\partial \eta_i} = b''(\theta_i) = v(\mu_i)$$

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

Thus,

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\beta}, \mathbf{y})}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{y_i - \mu_i}{a_i(\phi)} \, \frac{1}{v(\mu_i)} \, \frac{\partial \mu_i}{\partial \eta_i} \, x_{ij} \\ &= \sum_{i=1}^{n} (y_i - \mu_i) \, W_i \, \frac{\partial \eta_i}{\partial \mu_i} \, x_{ij}\end{aligned}$$

where W_i is referred to as the iterative weights given by

$$\begin{aligned}W_i &= \frac{1}{a_i(\phi)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 v_i^{-1} &(3.8)\\ &= \frac{1}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 &(3.9)\end{aligned}$$

and $v_i = v(\mu_i)$ is the variance function. Since $\eta_i = g(\mu_i)$, $\dfrac{\partial \mu_i}{\partial \eta_i}$ depends on the link function for the model.

Therefore, solving for the score equation below will give the ML estimate of $\boldsymbol{\beta}$:

$$\sum_{i=1}^{n} (y_i - \mu_i) \, W_i \, \frac{\partial \eta_i}{\partial \mu_i} \, x_{ij} = 0 \tag{3.10}$$

This score equation is a nonlinear function of $\boldsymbol{\beta}$, and therefore requires iterative procedures to be solved. Again, the Newton Raphson and Fisher Score iterative Equations 2.13 and 2.15 from Chapter 2 can be used, where the score $\boldsymbol{U}$ is given by the left hand side of Equation 3.10. Thus, the Newton Raphson iterative equation will be

$$\widehat{\boldsymbol{\beta}}^{(t+1)} = \widehat{\boldsymbol{\beta}}^{(t)} - (\boldsymbol{H}^{(t)})^{-1} \, \boldsymbol{U}^{(t)} \tag{3.11}$$

and the Fisher Score iterative equation

$$\widehat{\boldsymbol{\beta}}^{(t+1)} = \widehat{\boldsymbol{\beta}}^{(t)} + (\boldsymbol{\mathcal{I}}^{(t)})^{-1} \, \boldsymbol{U}^{(t)} \tag{3.12}$$

with information matrix

$$\begin{aligned}\boldsymbol{\mathcal{I}} &= -E(\boldsymbol{H}) &(3.13)\\ &= -E\left(\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \, \partial \boldsymbol{\beta}'} \right) &(3.14)\\ &= \mathbf{X}' \, \mathbf{W} \, \mathbf{X} &(3.15)\end{aligned}$$

where $\mathbf{W}$ is known as the weight matrix with diagonal elements given in Equation 3.8. Equation 3.12 can also be represented as

$$\boldsymbol{\mathcal{I}}^{(t)}\,\widehat{\boldsymbol{\beta}}^{(t+1)} = \boldsymbol{\mathcal{I}}^{(t)}\,\widehat{\boldsymbol{\beta}}^{(t)} + \boldsymbol{U}^{(t)} \tag{3.16}$$

It can be shown that the right hand side of Equation 3.16 can be written as

$$\mathbf{X}'\,\mathbf{W}^{(t)}\,\mathbf{z}^{(t)}$$

where $\mathbf{W}^{(t)}$ is weight matrix evaluated at $\widehat{\boldsymbol{\beta}}^{(t)}$, and $\mathbf{z}^{(t)}$ has the following elements evaluated at $\widehat{\boldsymbol{\beta}}^{(t)}$

$$z_i = \eta_i + (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right) \tag{3.17}$$

This variable z_i is often called the adjusted dependent variable or the working variable. Therefore, we can obtain

$$\widehat{\boldsymbol{\beta}}^{(t+1)} = (\mathbf{X}'\,\mathbf{W}^{(t)}\,\mathbf{X})^{-1}\,\mathbf{X}'\,\mathbf{W}^{(t)}\,\mathbf{z}^{(t)} \tag{3.18}$$

Thus, each iteration step is the result of a weighted least squares regression of the adjusted variable z_i on the predictors x_i with working weight W_i. Fisher scoring can therefore be regarded as iteratively reweighted least squares (IRWLS) carried out on a transformed version of the dependent variable (Bates, 2010).

It follows that the asymptotic variance (also known as the asymptotic covariance) of this estimate of $\boldsymbol{\beta}$ is the inverse of the information matrix given in Equation 3.15 and can be estimated by

$$\widehat{Var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}'\,\widehat{\mathbf{W}}\,\mathbf{X})^{-1} \tag{3.19}$$

where $\widehat{\mathbf{W}}$ is $\mathbf{W}$ evaluated at $\widehat{\boldsymbol{\beta}}$ and depends on the link function of the model. The dispersion parameter ϕ, in function $a_i(\phi)$ that is used in the calculation of W_i, gets cancelled out of the IRWLS procedure, thus the value of $\widehat{\boldsymbol{\beta}}$ is the same under any value of ϕ. However, the value of ϕ is required for the calculation of the variance of $\widehat{\boldsymbol{\beta}}$, therefore when ϕ is unknown, it can be estimated using a moment estimator (McCulloch & Searle, 2001), given by

$$\widehat{\phi} = \frac{1}{n-p-1}\sum_{i=1}^{n}\frac{w_i\,(y_i - \widehat{\mu}_i)^2}{v(\widehat{\mu}_i)} \tag{3.20}$$

where w_i is the weight defined in Equation 3.1.

3.1.2 Measure of Fit

An important step in statistical analyses is to assess the goodness-of-fit of the model of interest. One way in which this could be done is by using the *deviance*, a measure of discrepancy between the predicted values from the fitted model and the actual values from the data set. If, for the fitted model with $p+1$ parameters, $\ell(\widehat{\boldsymbol{\mu}}, \phi, \mathbf{y})$ is the log-likelihood function maximized over $\widehat{\boldsymbol{\beta}}$ for a fixed value of the dispersion parameter ϕ, and $\ell(\mathbf{y}, \phi, \mathbf{y})$ is the maximum log-likelihood achievable under the saturated model where the number of parameters equals the number of observations, the scaled deviance is

$$D^s = \frac{-2[\ell(\widehat{\boldsymbol{\mu}}, \phi, \mathbf{y}) - \ell(\mathbf{y}, \phi, \mathbf{y})]}{\phi} \tag{3.21}$$

If $\phi = 1$, the the deviance is defined as

$$D = -2[\ell(\widehat{\boldsymbol{\mu}}, \phi, \mathbf{y}) - \ell(\mathbf{y}, \phi, \mathbf{y})] \tag{3.22}$$

The (scaled) deviance converges asymptotically to a χ^2 distribution with $n-p-1$ degrees of freedom. Thus, when testing at a level of significance of α, the fitted model is rejected if the calculated deviance is greater than or equal to $\chi^2_{n-p-1;\,\alpha}$

Another commonly used measure of goodness-of-fit is the *generalized Pearson's chi-square statistic* given by

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{v(\widehat{\mu}_i)} \tag{3.23}$$

where $v(\widehat{\mu}_i)$ is the estimated variance function for the distribution in question. This statistic also asymptotically follows a χ^2 distribution with $n-p-1$ degrees of freedom. Similar to the deviance, the smaller the value of the χ^2 statistic, the better the fit of the model. The scaled Pearson's χ^2 statistic is $\dfrac{\chi^2}{\phi}$ (Wu, 2005). For linear models, the value of the Pearson's χ^2 statistic is the residual sum of squares since $v(\widehat{\mu}_i)$ is generally taken as one, and both the deviance and Pearson's χ^2 statistic have exact χ^2 distributions. For other distributions, these measures of goodness-of-fit have asymptotic χ^2 distributions and neither is superior to one another when samples are small. However, the deviance has an advantage over Pearson's χ^2 statistic as it is additive for nested models (Nelder & Wedderburn, 1972).

3.1.3 Likelihood Ratio Test

Suppose one is interested in testing whether certain parameters are equal to zero. In other words, whether the corresponding variables have no effect on the response variable given the other variables in the model. This can be done by comparing the deviances of the full model and the reduced model. Thus, the test statistic is calculated using the following

$$D_{\text{reduced}} - D_{\text{full}} \tag{3.24}$$

Since both the deviances above involve the log-likelihood for the saturated model, this gets cancelled out resulting in the following test statistic

$$\chi^2 = -2[\text{log-likelihood(reduced model)} - \text{log-likelihood(full model)}] \tag{3.25}$$

This test statistic has an asymptotic χ^2 distribution with degrees of freedom equal to the difference in the number of parameters fitted in the full model and the reduced model. This test is referred to as a *Likelihood Ratio Test*.

If $\phi \neq 1$, it was seen in Section 3.1.2 that a scaled deviance can be used. Thus, using this definition of the scaled deviance, the test statistic in Equation 3.25 would become

$$T = \frac{-2[\text{log-likelihood(reduced model)} - \text{log-likelihood(full model)}]}{\phi} \tag{3.26}$$

When $\phi \neq 1$ and unknown, the value of ϕ can be estimated using equation 3.20.

3.1.4 Wald Test

When a hypothesis test on a single parameter, β_j, is to be carried out, a commonly used method is the Wald test. The test statistic for this test is

$$z_0 = \frac{\widehat{\beta}_j}{se(\widehat{\beta}_j)} \tag{3.27}$$

The standard error of $\widehat{\beta}_j$ is the square root of the diagonal elements in the inverse of the information matrix given in Equation 3.15. This test statistic follows an approximate standard normal distribution. Some software packages square this value of the Wald test statistic and thus compare it to a chi-square distribution with 1 degree of freedom (Heeringa et al., 2010). Thus, for large values of the test statistic, one would reject the null hypothesis $H_0 : \beta_j = 0$ and conclude its corresponding variable is significant to the model.

3.2 Quasi-Likelihood Function

The method of maximum likelihood requires the probability distribution of Y to be known in advance. Sometimes there is not enough information about the data for a probability distribution to be specified (McCullagh & Nelder, 1989). In this case, the quasi-likelihood (QL) function is a commonly used method of estimating the parameters. Wedderburn (1974) showed that only the relationship between the mean and variance of the observations needs to be specified in order to define the quasi-likelihood function for the data. Thus, it allows relaxation of the usual assumptions, such as overdispersion which may be caused by correlated data (Agresti, 2007).

In determining the QL function for the data, only the first and second moments of Y_i are required (McCullagh, 1983). It is also assumed that for each observation, μ_i can be represented in terms of some known function of the explanatory variables $\boldsymbol{x}_i'$ and regression parameters $\boldsymbol{\beta}$. Wedderburn (1974) used the following relation to determine the quasi-likelihood (specifically the quasi-log likelihood) function $Q(y_i;\mu_i)$ for each observation

$$\frac{\partial Q(y_i;\mu_i)}{\partial \mu_i} = \frac{w_i\,(y_i-\mu_i)}{\phi\, v(\mu_i)} \tag{3.28}$$

where w_i is the known weight associated with observation Y_i.

Therefore, from Equation 3.28 we can obtain

$$Q(y_i;\mu_i) = \int_{y_i}^{\mu_i} \frac{w_i\,(y_i-t)}{\phi\, v(t)}\, dt + \text{some function of } y_i \tag{3.29}$$

Thus, the maximum quasi-likelihood estimates of $\boldsymbol{\beta}$ can be obtained from Equation 3.29 using Fisher Scoring. The estimate of ϕ can be obtained using Equation 3.20.

The above QL method is for the case where the observations are independent,however, this can be extended for the case where the observations are correlated. It can be shown that the properties of the quasi-log likelihood function are similar to those of the ordinary log-likelihood function, thus the asymptotic theory still applies which makes it possible to carry out measures of fit and hypothesis tests using methods discussed in previous sections. Furthermore, when the probability distribution of Y_i belongs to the exponential family, the quasi-log likelihood function of Y_i is identical to the log-likelihood function of Y_i (Wedderburn, 1974).

3.3 Logistic Regression

In the case where the response variable is binary, we can code the outcome as follows:

$$Y_i = \begin{cases} 1 & \text{if an event is observed, e.g. testing positive for malaria} \\ 0 & \text{if an event is not observed, e.g. testing negative for malaria} \end{cases}$$

Thus, Y_i follows a Bernoulli distribution with $P(Y_i = 1) = \pi_i$ and $P(Y_i = 0) = 1 - \pi_i$ Therefore,

$$E(Y_i) = \pi_i \quad \text{and} \tag{3.30}$$

$$Var(Y_i) = \pi_i(1 - \pi_i) \tag{3.31}$$

Suppose we used the LM to model Y_i:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i \qquad i = 1, \ldots, n \tag{3.32}$$

Thus, by Equation 3.30

$$E(Y_i) = \pi_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \tag{3.33}$$

$$= \boldsymbol{x}_i' \boldsymbol{\beta} \tag{3.34}$$

Since π_i is a probability, it is limited by $0 \leq \pi_i \leq 1$. However, using Equation 3.32 to model a binary outcome would result in the value of $E(Y_i)$ outside of its range. Therefore, a model for $E(Y_i)$ bounded between 0 and 1 would be more suitable (Rencher & Schaalje, 2008).

This model can be found by applying a logit transformation to Equation 3.33 as follows:

$$logit(\pi_i) = ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \boldsymbol{x}_i' \boldsymbol{\beta} \tag{3.35}$$

The left hand side of the above equation is referred to as the logit link, denoted by η_i in the GLM. Thus, Equation 3.33 will become

$$\pi_i = \frac{exp(\boldsymbol{x}_i' \boldsymbol{\beta})}{1 + exp(\boldsymbol{x}_i' \boldsymbol{\beta})} \tag{3.36}$$

This is known as the logistic regression model which is a class of the GLM with a logit link. The value of the link η_i is allowed to range freely while restricting that of $E(Y_i) = \pi_i = \mu_i$ between 0 and 1. The maximum likelihood estimates of $\boldsymbol{\beta}$ can be found using the iterative equations discussed in Section 3.1.1, where the score $\boldsymbol{U}$ can be found as follows.

A binary variable has the following probability distribution

$$f(y_i) = {\pi_i}^{y_i}\,(1-\pi_i)^{1-y_i} \tag{3.37}$$

This can be expressed in the form

$$f(y_i) = exp\left[y_i\, ln\left(\frac{\pi_i}{1-\pi_i}\right) + ln(1-\pi_i)\right] \tag{3.38}$$

The equation above is in the same form of Equation 3.1 where $a_i(\phi) = 1$, thus the dispersion parameter $\phi = 1,\ c(y_i,\phi) = 0$ and the canonical parameter $\theta_i = ln\left(\frac{\pi_i}{1-\pi_i}\right)$, which results in $\pi_i = \frac{e^{\theta_i}}{1+e^{\theta_i}}$. Thus, it follows $b(\theta_i) = ln(1+e^{\theta_i})$.

Since, for the logistic regression model, $E(Y_i) = \pi_i = \mu_i$, $Var(Y_i) = \mu_i(1-\mu_i) = v_i$ and the link function

$$\begin{aligned}\eta_i &= ln\left(\frac{\mu_i}{1-\mu_i}\right)\\ &= ln(\mu_i) - ln(1-\mu_i)\end{aligned}$$

It follows

$$\begin{aligned}\frac{\partial\eta_i}{\partial\mu_i} &= \frac{\partial}{\partial\mu_i}[ln(\mu_i) - ln(1-\mu_i)]\\ &= \frac{1}{\mu_i} + \frac{1}{1-\mu_i}\\ &= \frac{1}{\mu_i\,(1-\mu_i)}\end{aligned}$$

and

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

Therefore,

$$W_i = \frac{1}{a_i(\phi)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 v_i^{-1} \tag{3.39}$$

$$= [\mu_i\,(1-\mu_i)]^2\frac{1}{\mu_i(1-\mu_i)} \tag{3.40}$$

$$= \mu_i\,(1-\mu_i) \tag{3.41}$$

Thus, the score $\boldsymbol{U}$ given in Equation 3.10 reduces to

$$\boldsymbol{U} = \sum_{i=1}^{n} (y_i - \mu_i)\, x_{ij}$$

where μ_i, which is equal to π_i, is given by Equation 3.36. Using this value of $\boldsymbol{U}$ in iterative Equations 3.11 and 3.12, the ML estimate of $\boldsymbol{\beta}$ can be obtained. Equation 3.19 can be used to determine the variance of $\widehat{\boldsymbol{\beta}}$ where the diagonal elements in weight matrix $\mathbf{W}$ are given by Equation 3.41.

A useful property of logistic regression is that the link function represents the log of the odds of an event of interest occurring, where $\dfrac{\pi_i}{1-\pi_i}$ is the odds of the event occurring. Therefore, taking e^{β_j} gives the odds ratio corresponding to a one unit increase in the corresponding explanatory variable x_{ij}, while all other explanatory variables remain the same. In general, for a k unit change in the explanatory variable, the odds ratio is $e^{k\,\beta_j}$. This is a helpful way to determine how much more likely an event of interest is to occur when one explanatory variable changes (Kutner et al., 2005).

3.4 Survey Logistic Regression

As discussed previously, logistic regression is a popular method to analyze the relationship between a binary outcome and a set of explanatory variables, however, this method does not take into account the design of a study (An, 2002). Failure to account for the complex design of a study, where stratification or clustering is used, the analysis may result in an overestimation of standard errors, therefore leading to incorrect results (Nadimpalli & Hubbell, 2012). Thus, some adjustments to the ordinary logistic regression model that account for the survey design are necessary in order to make valid inferences (Roberts et al., 1987). Logistic regression that is used in the analysis of complex survey designs is referred to as survey logistic regression (SLR). The theory for ordinary logistic regression and survey logistic regression is the same, however, survey logistic regression uses special methods of estimating the model's parameters and the corresponding variances (Nadimpalli & Hubbell, 2012).

The most commonly used methods of variance estimation for the survey logistic regression model, which will be discussed in the sections to come, are Taylor series approximation which is based on a linearization technique, Jackknife repeated replication (JRR) and balanced repeated replication (BRR) which are based on a resampling technique (Heeringa et al., 2010).

A number of studies have compared the results of these different methods of variance estimation for complex survey designs (Kish & Frankel, 1974; Rao & Wu, 1985; Kovar et al., 1988). Many have shown that none of the methods obtain a better or worse estimate, and the choice of method may depend on the design of the study as well as the availability of resources, such as statistical programs and computing power (Nadimpalli & Hubbell, 2012). Krewski & Rao (1981) and Rao & Shao (1992) have shown that the linearization and resampling techniques are asymptotically equivalent and both of the techniques lead to consistent variance estimators.

3.4.1 The Model

Let's consider the survey logistic regression model for a binary response where Y_{hij}, $j = 1, \dots, n_{hi};\ i = 1, \dots, n_h;\ h = 1, \dots, H$ is an observation for the j^{th} individual in the i^{th} PSU (cluster) within the h^{th} stratum. Therefore, $\pi_{hij} = P(Y_{hij} = 1)$ represents the probability of an event of interest occurring for the j^{th} individual in the i^{th} PSU within the h^{th} stratum, e.g. the individual testing positive for malaria. Thus, the survey logistic regression model is

$$logit(\pi_{hij}) = \boldsymbol{x}'_{hij}\boldsymbol{\beta} \tag{3.42}$$

with

$$\pi_{hij} = \frac{exp(\boldsymbol{x}'_{hij}\boldsymbol{\beta})}{1 + exp(\boldsymbol{x}'_{hij}\boldsymbol{\beta})} \tag{3.43}$$

where $\boldsymbol{x}_{hij}$ is the row of the design matrix corresponding to the response of the j^{th} individual in the i^{th} PSU within the h^{th} stratum, and $\boldsymbol{\beta}$ is the vector of unknown parameters to be estimated.

This survey logistic regression model is in the same form as the ordinary logistic regression model from Section 3.3. Thus, it follows that the probability distribution of the response variable is be given by

$$f(y_{hij}) = \pi_{hij}^{y_{hij}} (1 - \pi_{hij})^{1-y_{hij}} \tag{3.44}$$

with

$$\begin{aligned} E(Y_{hij}) &= \pi_{hij} \\ &= \frac{e^{\boldsymbol{x}'_{hij}\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}'_{hij}\boldsymbol{\beta}}} \end{aligned}$$

and

$$\begin{aligned} Var(Y_{hij}) &= \pi_{hij}(1-\pi_{hij}) \\ &= \frac{e^{\boldsymbol{x}'_{hij}\boldsymbol{\beta}}}{\left(1+e^{\boldsymbol{x}'_{hij}\boldsymbol{\beta}}\right)^2} \end{aligned}$$

Therefore, the log-likelihood function is

$$\ell = ln\, L(\mathbf{y}) = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} ln\, f(y_{hij}) \tag{3.45}$$

The above log-likelihood function does not take the sampling weights into consideration, thus the ML estimates of the model's parameters obtained using this function are only valid for simple random samples where observations are unweighted (Heeringa et al., 2010). Under more complex designs involving sampling weights and clustering, the ML estimates of the parameters and their standard errors are not consistent (Chandra, 2014), and thus the traditional ML method has to be modified to account for weighted observations. The traditional likelihood function is based on standard distributional assumptions about the response variable, however, for complex survey designs, no convenient likelihood functions are available (Chandra, 2014).

Therefore, a likelihood function that incorporates the sampling weights should be used. Such a likelihood function is referred to as a pseudo-likelihood function. The method of estimation that uses this pseudo-likelihood function is known as pseudo-maximum likelihood (PML) estimation.

3.4.2 Pseudo-Likelihood Function

Like the ML method, the PML method requires knowledge of the distribution of the response variable, however it accounts for the sampling weights as follows

$$P\ell = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} w_{hij}\, ln\, f(y_{hij}) \tag{3.46}$$

where w_{hij} is the weight associated with observation Y_{hij} and $P\ell$ represents the pseudo-log likelihood function.

Thus, for the survey logistic regression model, the pseudo-log likelihood function is

$$P\ell = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} w_{hij}\ [y_{hij}\, ln(\pi_{hij}) + (1 - y_{hij})\, ln(1 - \pi_{hij})]$$

In order to obtain the parameter estimates, the above equation is maximized with respect to $\boldsymbol{\beta}$. It can be shown this results in the following estimating equations

$$\boldsymbol{S}(\boldsymbol{\beta}) = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} w_{hij}(y_{hij} - \pi_{hij})\, \boldsymbol{x}'_{hij} = \mathbf{0} \tag{3.47}$$

Thus, the weighted parameter estimates can be obtained using the Newton Raphson and Fisher Score iterative procedures where the score $\boldsymbol{U}$ is given in Equation 3.47 (Agresti, 2002). It has been shown that the parameter estimates based on the PML method of estimation is consistent (Heeringa et al., 2010).

3.4.3 Taylor Series Approximation

Due to weighting and clustering, the estimated variances of the PML parameter estimates are no longer equal to the inverse of the information matrix as discussed in Section 3.1.1 for the GLM. In order to obtain these variance estimates, Binder (1983) proposed making use of the Taylor series approximation method.

Since the parameter estimates, $\widehat{\boldsymbol{\beta}}$, are defined by equations

$$\boldsymbol{S}(\widehat{\boldsymbol{\beta}}) = \mathbf{0} \tag{3.48}$$

the first order Taylor expansion of $\boldsymbol{S}(\widehat{\boldsymbol{\beta}})$ at $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, the population parameter value, is

$$\mathbf{0} = \boldsymbol{S}(\widehat{\boldsymbol{\beta}}) \simeq \boldsymbol{S}(\boldsymbol{\beta}) + \frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \tag{3.49}$$

Therefore

$$\boldsymbol{S}(\boldsymbol{\beta}) \simeq -\frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \tag{3.50}$$

After applying the Delta method, the following result is obtained in the limit

$$Var\left[\boldsymbol{S}(\widehat{\boldsymbol{\beta}})\right] = \left[\frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right] Var(\widehat{\boldsymbol{\beta}}) \left[\frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]' \tag{3.51}$$

or equivalently

$$Var(\widehat{\boldsymbol{\beta}}) = \left[\frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]^{-1} Var\left[\boldsymbol{S}(\widehat{\boldsymbol{\beta}})\right] \left[\frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]^{-1} \tag{3.52}$$

This leads to a sandwich-type variance estimator

$$\widehat{Var}(\widehat{\boldsymbol{\beta}}) = \left[\boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\beta}})\right]^{-1} Var\left[\boldsymbol{S}(\widehat{\boldsymbol{\beta}})\right] \left[\boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\beta}})\right]^{-1} \tag{3.53}$$

where $\boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\beta}}) = \dfrac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \dfrac{\partial^2 P\ell}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}'}$ is the information matrix evaluated at $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ and $Var\left[\boldsymbol{S}(\widehat{\boldsymbol{\beta}})\right]$ is the variance-covariance matrix for the $p+1$ estimating equations. Since each of the estimating equations is a sample total of the individual scores for the n survey respondents, obtained using stratified and cluster sampling, standard formulae to estimate the variances and covariances of the estimating equations can be used (Heeringa et al., 2010). Therefore, it follows

$$Var\left[\boldsymbol{S}(\widehat{\boldsymbol{\beta}})\right] = \frac{n}{n-p-1}\sum_{h=1}^{H}(1-f_h)\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}(\boldsymbol{s}_{hi.} - \bar{\boldsymbol{s}}_{h..})'(\boldsymbol{s}_{hi.} - \bar{\boldsymbol{s}}_{h..}) \tag{3.54}$$

where

$$\boldsymbol{s}_{hi.} = \sum_{j=1}^{n_{hi}} \boldsymbol{s}_{hij} = \sum_{j=1}^{n_{hi}} w_{hij}(y_{hij} - \widehat{\pi}_{hij})\, \boldsymbol{x}'_{hij} \tag{3.55}$$

and

$$\bar{\boldsymbol{s}}_{h..} = \frac{1}{n_h}\sum_{i=1}^{n_h} \boldsymbol{s}_{hi.} \tag{3.56}$$

and the quantity $(1 - f_h)$ is the finite population correction factor, where $f_h = \dfrac{n_h}{N_h}$ is the sampling rate for stratum H with N_h as the the total number of PSUs in stratum h and n_h is the number of sampled PSUs. If N_h is unknown, it is common to assume that it is large enough such that f_h is very small, which results in the correction factor equalling one (Hosmer et al., 2013). The value of $\widehat{\pi}_{hij}$ is calculated by substituting the parameter estimate $\widehat{\boldsymbol{\beta}}$ into Equation 3.43. For large n, Equation 3.54 reduces to

$$Var\left[\boldsymbol{S}(\widehat{\boldsymbol{\beta}})\right] = \sum_{h=1}^{H}(1-f_h)\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}(\boldsymbol{s}_{hi.} - \bar{\boldsymbol{s}}_{h..})'(\boldsymbol{s}_{hi.} - \bar{\boldsymbol{s}}_{h..}) \tag{3.57}$$

The variance estimator in Equation 3.53 is a consistent estimator for the asymptotic variance of $\widehat{\boldsymbol{\beta}}$ (Lipsitz et al., 1994).

3.4.4 Jackknife Repeated Replication

This method is applicable to a wide range of complex survey designs where the form of the jackknife variance estimator depends on the design of the study (Heeringa et al., 2010). There have been two distinct areas of research on the jackknife method. Quenouille (1949) developed the method for bias reduction, and Tukey (1958) used the basis of this research done by Quenouille (1949) for variance estimation. Lee (1973) and Jones (1974), among others, extended Tuckey's idea for stratified multi-stage sampling. A property of the jackknife method is that it is not based on any assumptions of the model and therefore is less susceptible to violation of any assumptions (Shao, 1992).

This JRR method involves estimating parameters of several sub-samples, which are obtained by deleting one observation at a time from the full sample, then determining the variance of the parameter estimate for the full sample using the variability between the sub-sample estimates (Ahmad, 2014). Extensions to deleting more than one observation or a group of observations are available (Ahmad, 2014). The general form of of the jackknife variance estimator is

$$\widehat{Var}(\widehat{\theta}) = \frac{G}{1-G}\sum_{k=1}^{G}(\widehat{\theta}_{(k)} - \widehat{\theta})^2 \tag{3.58}$$

where $\widehat{\theta}_{(k)}$ is the parameter estimate for the sub-sample with the k^{th} observation deleted, $\widehat{\theta}$ is the parameter estimate for the full sample and G is the number of sub-samples, also referred to as the number of replicates.

In the case of stratified cluster sampling without replacement, each of the sub-samples are obtained by deleting one or more of the PSUs (in this case the clusters) from a single stratum. Operationally the observations are not deleted, but rather are assigned a weight of zero (Heeringa et al., 2010). The remaining PSUs in the stratum are assigned new weights, referred to as jackknife weights, while all other sample weights in the other strata remain unchanged. Thus, the jackknife weight for the i^{th} PSU, $i = 1, \dots, n_h$, in the h^{th} stratum, $h = 1, \dots, H$, when the j^{th} PSU from the g^{th} stratum is deleted is given by

$$w_{hi(gj)} = \begin{cases} 0 & (gj) = (hi) \\ \dfrac{n_g}{n_g - 1} w_{gi}, & h = g,\, i \neq j \\ w_{hi} & h \neq g \end{cases}$$

Therefore, these jackknife weights above can replace the sampling weights in Equation 3.47 to determine the estimating equations for each of the sub-samples. Newton Raphson and Fisher Score iterative procedures once again can be used to determine the solution to these new estimating equations in order to obtain the parameter estimates for each of the sub-samples. A common method used to estimate these parameters is the one-step jackknife method where the parameter estimate for the full sample is used as the starting value in the Newton Raphson iterative procedure (Lipsitz et al., 1994).

Once these estimates have been obtained, the jackknife variance estimator for this sampling design can be determined using

$$\widehat{Var}(\widehat{\boldsymbol{\beta}}) = \sum_{g=1}^{H} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\widehat{\boldsymbol{\beta}}_{(gj)} - \widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}_{(gj)} - \widehat{\boldsymbol{\beta}})' \tag{3.59}$$

where $\widehat{\boldsymbol{\beta}}_{(gj)}$ is the parameter estimate for the sub-sample without the j^{th} PSU from the g^{th} stratum and $\widehat{\boldsymbol{\beta}}$ is the parameter estimate for the full sample.

The jackknife method has been shown to produce consistent results in large samples (Miller, 1974). A disadvantage of this jackknife variance estimator is it can be time consuming when calculating the replicate estimates for very large samples (Yung & Rao, 2000). It was therefore proposed by Yung & Rao (1996) to linearize this variance estimator, which is referred to as the the jackknife linearization variance estimator. It was shown that the jackknife linearization variance estimator performed as well as the jackknife variance estimator (Yung & Rao, 2000).

3.4.5 Balanced Repeated Replication

The BRR method, first proposed by McCarthy (1969), was developed specifically for variance estimation under the design of two PSUs (clusters) per stratum. This method is based on forming half-sample replicates by deleting one PSU in each stratum. Thus, for a design with H strata, the full sample can be split into 2^H half-samples that overlap with H sample clusters in each. The parameter can be estimated for each of the half-samples and used to estimate the variance of the parameter estimate for the full sample. However, this may be time consuming and very difficult for large H. Thus, a balanced set of only K half-samples may be constructed, where K is the smallest multiple of 4 that is greater than H. The half-samples are selected by using the first H columns of the $K \times K$ orthogonal Hadamard matrix (SAS Institute Inc., 2009).

The k^{th} half-sample will be selected from the full sample according to the following:

- If the element corresponding to the k^{th} row and h^{th} column in the Hadamard matrix is 1, then the first PSU from the h^{th} stratum is included in the k^{th} half-sample and the second PSU from the h^{th} stratum is excluded.
- If the element corresponding to the k^{th} row and h^{th} column in the Hadamard matrix is -1, then the second PSU from the h^{th} stratum is included in the k^{th} half-sample and the first PSU from the h^{th} stratum is excluded.

The sampling weights of the PSUs included in the half-samples are adjusted by multiplying their original sampling weights by a factor of 2. These new weights are referred to as the replicate weights (Heeringa et al., 2010). Using this BRR method, the variance estimator of the full sample parameter estimate is in the following form

$$\widehat{Var}(\widehat{\theta}) = \frac{1}{K}\sum_{k=1}^{K}(\widehat{\theta}_i - \widehat{\theta})^2 \tag{3.60}$$

where $\widehat{\theta}_i$ is the parameter estimate for the i^{th} half-sample using the new replicate weights and $\widehat{\theta}$ is the parameter estimate for the full sample.

The BRR method is not presently applicable for arbitrary sample sizes, however, it is said to have an advantage over JRR as it leads to asymptotic inferences for both smooth and non-smooth estimates (Rao, 1997).

3.4.6 Assessing the Model

Goodness-of-Fit

The goodness-of-fit tests discussed in Section 3.1.2 for the GLM are based on selected observations that are independent and identically distributed. However, in the case of a complex survey design, it is very common that observations from the same cluster are often more homogeneous than observations from different clusters. Thus, goodness-of-fit tests that take into consideration the design of the study are more appropriate in assessing the fit of the survey logistic regression model. For complex survey designs, Archer & Lemeshow (2006) and Archer et al. (2007) extended the Hosmer-Lemeshow goodness-of-fit test, which was proposed by Hosmer & Lemeshow (1980) specifically for ordinary logistic regression to avoid possible problems associated with the asymptotic distribution of the chi-square tests.

The Hosmer-Lemeshow goodness-of-fit test is based on grouping the observations in "deciles of risk", where the observations are partitioned into 10 equal-sized groups based on their ordered estimated probabilities, $\widehat{\pi}_i$. The Hosmer-Lemeshow test statistic is given by

$$\widehat{C} = \sum_{k=1}^{10} \frac{(O_k - E_k)^2}{E_k\left(1 - \dfrac{E_k}{n_k}\right)} \tag{3.61}$$

where

- n_k is the number of observations in the k^{th} decile.
- $O_k = \sum\limits_i y_i =$ observed number of cases in the k^{th} decile.
- $E_k = \sum\limits_i \widehat{\pi}_i =$ expected number of cases in the k^{th} decile.

This test statistic is has a chi-square distribution with 8 degrees of freedom (Hosmer & Lemeshow, 1980). The extension of this Hosmer-Lemeshow goodness-of-fit test proposed by Archer & Lemeshow (2006) is called the F-adjusted mean residual test, sometimes also referred to as the Archer and Lemeshow goodness-of-fit test, which is estimated as follows.

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

$$\widehat{r}_{ij} = y_{ij} - \widehat{\pi}(x_{ij}) \tag{3.62}$$

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

$$\widehat{M}_g = \frac{\sum\limits_i \sum\limits_j w_{ij}\widehat{r}_{ij}}{\sum\limits_i \sum\limits_j w_{ij}} \tag{3.63}$$

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

The Wald test statistic for testing g categories is given by

$$\widehat{W} = \widehat{\boldsymbol{M}}\,' \left[\widehat{Var}(\widehat{\boldsymbol{M}})\right]^{-1} \widehat{\boldsymbol{M}} \tag{3.64}$$

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

$$F = \frac{(f - g + 2)}{fg} W \tag{3.65}$$

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

$$\widehat{Q}_m = \frac{(f - 8)}{10f} \widehat{\boldsymbol{M}}\,' \left[\widehat{Var}(\widehat{\boldsymbol{M}})\right]^{-1} \widehat{\boldsymbol{M}} \tag{3.66}$$

as $g = 10$ deciles of risk.

Testing Model Parameters

Since the estimates of the model parameters for the survey logistic regression model are determined using the pseudo-likelihood function, which is an approximate to the true likelihood, inferences about the parameters cannot be based on likelihood ratio tests (Hosmer et al., 2013). Thus, it is more appropriate to use Wald tests instead. The general form of the null hypothesis for this test is $H_0 : \boldsymbol{C\beta} = \boldsymbol{0}$ where $\boldsymbol{C}$ is a matrix of constants that defines the hypothesis to be tested. The Wald test statistic is calculated as

$$W = (\boldsymbol{C}\widehat{\boldsymbol{\beta}})' \left[\boldsymbol{C}\,\widehat{Var}(\widehat{\boldsymbol{\beta}})\,\boldsymbol{C}'\right]^{-1} (\boldsymbol{C}\widehat{\boldsymbol{\beta}}) \tag{3.67}$$

where $\widehat{Var}(\widehat{\boldsymbol{\beta}})$ is the estimated variance-covariance matrix for $\widehat{\boldsymbol{\beta}}$ using methods discussed in the previous sections. Under the null hypothesis, this test statistic follows a chi-square distribution with q degrees of freedom, where q is the rank or the number of independent rows of the matrix $\boldsymbol{C}$. It is again common to approximate this Wald test statistic to an F-distribution using Equation 3.65, where $g = q$.

The next chapter discusses some of the theory for the last two methods that will be applied to the MIS data; generalized estimating equations and generalized linear mixed models.

Chapter 4

Modeling Cluster-Correlated Data

The GLM assumes observations are independent. However, in the case of complex survey designs where stratified cluster sampling is carried out, very often the resulting observations within the same cluster tend to be more similar to one another than those from other clusters. Therefore, in this chapter two methods used for analyzing correlated data are discussed, both of which are an extension of GLMs. The choice of the two methods depends on what one is interested in determining from the data. The first method discussed in this chapter, the generalized estimating equations technique (also referred to as marginal modeling), is a population averaged approach where the population average fixed effects are the effects of interest. However, the second method discussed, the use of the generalized linear mixed model, is a subject-specific model which can be used for modeling the effect on an individual unit (Heeringa et al., 2010).

4.1 Generalized Estimating Equations

An extension of GLMs that can take into account intracluster correlation is the method of generalized estimating equations (GEEs). This method was first proposed by Liang & Zeger (1986) and is based on the quasi-likelihood approach (Agresti, 2002). Thus, GEEs do not require full specification of the distribution of the data, but rather only requires the specification of the mean as well as the mean-variance relationship. An advantage of GEEs is that estimates will be consistent even if the correlation structure of the observations has not been correctly specified (Jiang, 2007). This is due to the fact that the covariance structure is treated as a nuisance. In the case of missing data, GEEs require that the data are missing completely at random (MCAR), where the missing observations are completely independent of their values (Agresti, 2002). If the data are missing at random (MAR) instead, GEEs are no longer appropriate and weighted generalized estimating equations are more suitable. Refer to (Robins et al., 1995).

Even though the GEE method does not require any assumptions about the joint distribution of the observations, it does however make an assumption about the marginal distribution of each Y_{ij}, the j^{th} observation, $j = 1, \ldots, n_i$, from the i^{th} cluster, $i = 1, \ldots, m$ (Agresti, 2002). Thus, assuming that Y_{ij} has a probability distribution belonging to the exponential family in the form of Equation 3.1, where the weight of the observation is taken as 1, the mean

$$\mu_{ij} = E(Y_{ij})$$

is related to a linear combination of the explanatory variables via the link function

$$\eta_{ij} = g(\mu_{ij}) = \boldsymbol{x}'_{ij}\boldsymbol{\beta}$$

Therefore it follows, similar to the GLM, the marginal variance is given by

$$Var(Y_{ij}) = \phi\, \upsilon(\mu_{ij})$$

where $\upsilon(\mu_{ij}) = b''(\theta_{ij}) = \upsilon_{ij}$ is the variance function and ϕ is the dispersion parameter. Another way in which the score Equation 3.10 for the GLM can be represented is (Agresti, 2002)

$$\sum_i \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \upsilon_i^{-1}(y_i - \mu_i) = 0 \tag{4.1}$$

Now, for the case where the outcome $\boldsymbol{y}_i$ is an $n_i \times 1$ vector of correlated outcomes for cluster i, the score equation above becomes

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_i \frac{\partial \boldsymbol{\mu}'_i}{\partial \boldsymbol{\beta}} \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) \\ &= \sum_i \boldsymbol{F}'_i \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \boldsymbol{0} \end{aligned} \tag{4.2}$$

where $\boldsymbol{F}_i = \dfrac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}$ is an $n_i \times (p+1)$ matrix, $\boldsymbol{\mu}_i = E(\boldsymbol{y}_i)$ and $\boldsymbol{V}_i$ is referred to as the working covariance matrix for $\boldsymbol{y}_i$ defined by

$$\boldsymbol{V}_i = \boldsymbol{A}_i^{\frac{1}{2}} \boldsymbol{R}_i(\boldsymbol{\alpha}) \boldsymbol{A}_i^{\frac{1}{2}} \tag{4.3}$$

where $\boldsymbol{A}_i = \text{diag}(Var(Y_{ij})) = \text{diag}(\phi\, \upsilon_{ij})$ and $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is the working correlation matrix which depends on a vector of unknown parameters $\boldsymbol{\alpha}$. The GEE estimates of $\boldsymbol{\beta}$ is the solution to the estimating equations in Equation 4.2.

If $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is the true correlation matrix for $\boldsymbol{y}_i$, then $\boldsymbol{V}_i = cov(\boldsymbol{y}_i)$. In the case of independent observations, $\boldsymbol{R}_i(\boldsymbol{\alpha})$ is replaced by the identity matrix $\boldsymbol{I}_{n_i}$ and thus the working covariance matrix reduces to $\boldsymbol{V}_i = \boldsymbol{A}_i = \text{diag}(\phi\, \upsilon_{ij})$. Using this value of $\boldsymbol{V}_i$ in the estimating equations in Equation 4.2 would result in the ordinary GLM estimate of $\boldsymbol{\beta}$.

Iterative procedures are required to solve the estimating equations in Equation 4.2. However, these solutions for $\widehat{\boldsymbol{\beta}}$ depend on the parameters ϕ and $\boldsymbol{\alpha}$. Therefore, Liang & Zeger (1986) proposed the method of moments to estimate these unknown parameters at a given iteration using a function of the current standardized Pearson residuals

$$\widehat{e}_{ij} = \frac{y_{ij} - \widehat{\mu}_{ij}}{\sqrt{\upsilon(\widehat{\mu}_{ij})}} \tag{4.4}$$

Note: this value of $\widehat{e}_{ij}$ depends on the current value of the estimate for $\boldsymbol{\beta}$ (Liang & Zeger, 1986). The dispersion parameter can be estimated by

$$\widehat{\phi} = \frac{1}{N - p - 1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \widehat{e}^{\,2}_{ij} \tag{4.5}$$

where $N = \sum\limits_{i=1}^{m} n_i$ is the total number of observations. The next section discusses common estimates for $\boldsymbol{\alpha}$ depending on the choice of the working correlation structure for the observations.

4.1.1 Specifying The Working Correlation Structure

The correlation structure of the data is not of primary interest, however, it is essential for valid inferences. Different estimates for both ϕ and $\boldsymbol{\alpha}$ are available. Table 4.1 on the next page gives the moment based estimates that the statistical program SAS uses for each of the common choices of the working correlation structure. In the case ϕ is unknown, it is replaced by its estimate given in Equation 4.5. The closer the chosen working correlation structure is to the true correlation, the more efficient the estimate of both $\boldsymbol{\beta}$ and $\boldsymbol{V}_i$, however, for any choice of the working correlation structure, both the estimates will be consistent (Liang & Zeger, 1986). Thus, it is common to choose the structure with the smallest number of parameters to estimate, therefore the exchangeable (compound symmetry) correlation structure with only two parameters is often the choice.

Table 4.1: Moment based estimators for the common choices of the working correlation structure.

Working Correlation Structure		Estimator
Independent	$\text{Corr}(Y_{ij}, Y_{ik}) = \begin{cases} 1 & j = k \\ 0 & j \neq k \end{cases}$	—
M-dependent	$\text{Corr}(Y_{ij}, Y_{i;j+t}) = \begin{cases} 1 & t = 0 \\ \alpha_t & t = 1, \ldots, M \\ 0 & t > M \end{cases}$	$\widehat{\alpha}_t = \frac{1}{(m_t - p - 1)\phi} \sum_{i=1}^{m} \sum_{j \leq n_i - 1} \widehat{e}_{ij}\, \widehat{e}_{i;j+1}$ where $m_t = \sum_{i=1}^{m} (n_i - t)$
Exchangeable	$\text{Corr}(Y_{ij}, Y_{ik}) = \begin{cases} 1 & j = k \\ \alpha & j \neq k \end{cases}$	$\widehat{\alpha} = \frac{1}{(m^* - p - 1)\phi} \sum_{i=1}^{m} \sum_{j<k} \widehat{e}_{ij}\, \widehat{e}_{ik}$ where $m^* = \frac{1}{2} \sum_{i=1}^{m} n_i(n_i - 1)$
Unstructured	$\text{Corr}(Y_{ij}, Y_{ik}) = \begin{cases} 1 & j = k \\ \alpha_{jk} & j \neq k \end{cases}$	$\widehat{\alpha}_{jk} = \frac{1}{(m - p - 1)\phi} \sum_{i=1}^{m} \widehat{e}_{ij}\, \widehat{e}_{ik}$
AR(1)	$\text{Corr}(Y_{ij}, Y_{i;j+t}) = \alpha^t,\ t = 0, 1, \ldots, n_i - j$	$\widehat{\alpha} = \frac{1}{(m_1 - p - 1)\phi} \sum_{i=1}^{m} \sum_{j \leq n_i - 1} \widehat{e}_{ij}\, \widehat{e}_{i;j+1}$ where $m_1 = \sum_{i=1}^{m} (n_i - 1)$

4.1.2 Fitting The GEE Model

The values of $\widehat{\phi}$ and $\widehat{\boldsymbol{\alpha}}$ depend on the estimated Pearson residuals, $\widehat{e}_{ij}$, which in turn depend on the estimated values of $\boldsymbol{\beta}$. Thus, Liang & Zeger (1986) proposed computing the GEE estimates of $\boldsymbol{\beta}$ using the following iterative procedure:

1. Compute the initial estimates of $\boldsymbol{\beta}$ using a GLM by assuming the observations are independent.
2. Compute the standardized Pearson residuals $\widehat{e}_{ij}$ using Equation 4.4.
3. Compute the estimates for $\boldsymbol{\alpha}$ depending on the chosen correlation structure from the table above.
4. Compute the estimate for ϕ using Equation 4.5.
5. Compute $\widehat{\boldsymbol{R}}_i(\widehat{\boldsymbol{\alpha}})$ according to the chosen correlation structure.

6. Compute $\widehat{\boldsymbol{V}}_i(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\phi}) = \boldsymbol{A}_i(\widehat{\boldsymbol{\beta}})^{\frac{1}{2}}\,\widehat{\boldsymbol{R}}_i(\widehat{\boldsymbol{\alpha}})\,\boldsymbol{A}_i(\widehat{\boldsymbol{\beta}})^{\frac{1}{2}}$

7. Update the estimate for $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}}^{(t+1)} = \widehat{\boldsymbol{\beta}}^{(t)} - \left[\sum_{i=1}^{m} \widehat{\boldsymbol{F}}_i' \widehat{\boldsymbol{V}}_i^{-1} \widehat{\boldsymbol{F}}_i\right]^{-1} \left[\sum_{i=1}^{m} \widehat{\boldsymbol{F}}_i' \widehat{\boldsymbol{V}}_i^{-1} (\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_i)\right]$$

   where $\widehat{\boldsymbol{F}}_i = \dfrac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}$ evaluated at $\widehat{\boldsymbol{\beta}}$.

8. Repeat steps 2 to 7 until convergence is reached.

The resulting GEE estimate $\widehat{\boldsymbol{\beta}}$ is asymptotically normally distributed. The robust or empirical estimator of the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ is given by

$$\widehat{Var}(\widehat{\boldsymbol{\beta}}) = \left[\sum_{i=1}^{m} \widehat{\boldsymbol{F}}_i' \widehat{\boldsymbol{V}}_i^{-1} \widehat{\boldsymbol{F}}_i\right]^{-1} \left[\sum_{i=1}^{m} \widehat{\boldsymbol{F}}_i' \widehat{\boldsymbol{V}}_i^{-1} Var(\boldsymbol{y_i}) \widehat{\boldsymbol{V}}_i^{-1} \widehat{\boldsymbol{F}}_i\right] \left[\sum_{i=1}^{m} \widehat{\boldsymbol{F}}_i' \widehat{\boldsymbol{V}}_i^{-1} \widehat{\boldsymbol{F}}_i\right]^{-1}$$

It is common to replace $Var(\boldsymbol{y}_i)$ by $(\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_i)(\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_i)'$. The above is a consistent estimator of the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$, even if the working correlation structure has been misspecified (Liang & Zeger, 1986). If the correlation structure has been correctly specified, the estimator of the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ reduces to

$$\widehat{Var}(\widehat{\boldsymbol{\beta}}) = \left[\sum_{i=1}^{m} \widehat{\boldsymbol{F}}_i' \widehat{\boldsymbol{V}}_i^{-1} \widehat{\boldsymbol{F}}_i\right]^{-1}$$

This estimator is commonly referred to as a model-based estimator (SAS Institute Inc., 2009).

4.1.3 Model Selection

The GEE method is not a likelihood-based method, therefore likelihood ratio tests and information criteria such as Akaike's Information Criterion (AIC) cannot be used for model selection. Thus, Pan (2001) proposed a modification to AIC for the GEE, where the likelihood function is replaced by the quasi-likelihood function and the penalty term is adjusted. This modified information criterion is referred to as the 'Quasi-Likelihood Under the Independence Model Criterion' (QIC).

Using McCullagh & Nelder's (1989) quasi-likelihood (log) function defined as

$$Q(\mu, \phi; y) = \int_{y}^{\mu} \frac{y - t}{\phi\, v(t)} dt$$

where $Var(y) = \phi\, v(\mu)$ is the specified relationship between the variance and the mean for the distribution of y, the QIC is defined as

$$QIC(R) = -2\, Q(\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{V}}}, \boldsymbol{I}) + 2\, trace(\widehat{\boldsymbol{\Omega}}^{-1}\, \widehat{\boldsymbol{V}}_{\boldsymbol{R}})$$

where $Q(\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{V}}}, \boldsymbol{I})$ is the quasi-likelihood calculated using the independent working correlation structure, $\boldsymbol{I}$, but with the parameter estimates $\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{V}}}$, determined using the hypothesized correlation structure. $\widehat{\boldsymbol{\Omega}}$ is the model-based estimated variance-covariance matrix for $\widehat{\boldsymbol{\beta}}$ under the assumption of an independent working correlation structure and $\widehat{\boldsymbol{V}}_{\boldsymbol{R}}$ is the robust variance-covariance estimator with working correlation structure $\boldsymbol{R}$. The QIC can be used to determine the best fitting working correlation structure by selecting the structure whose model has the smallest QIC (Pan, 2001).

This $QIC(R)$ can be approximated by $QIC_u(R)$, which can be used in variable selection. This is given by

$$QIC_u(R) = Q(\widehat{\boldsymbol{\beta}}_{\widehat{\boldsymbol{V}}}, \boldsymbol{I}) + 2d$$

where d is the number of parameters fitted to the model. The value $2d$ serves as a penalty factor for increasing the number of parameters in the model. The model with the smallest value of $QIC_u(R)$ will be the optimal model. However, this $QIC_u(R)$ cannot be used to determine the best fitting working correlation structure (Pan, 2001).

4.2 Generalized Linear Mixed Models

For complex survey designs that use stratified cluster sampling methods, the design of the study is such that the clusters included in the sample represent only a random sample from a population of clusters. Thus, if one is interested in including the effect of clustering in the model, it would be included as a random effect. As discussed in Chapter 3, the GLM is used when modeling a non-normal response variable. When a random effect is included in a GLM, the resulting model is referred to as a generalized linear mixed model (GLMM).

4.2.1 The GLMM

Similar to the the case of the linear mixed model discussed in chapter 2, Y_{ij} is the j^{th} response, $j = 1, \ldots, n_i$, from the i^{th} cluster, $i = 1, \ldots, m$, and thus, $\boldsymbol{y}_i$ is the $n_i \times 1$ vector of responses for the i^{th} cluster. In the GLMM, responses Y_{ij} in $\boldsymbol{y}_i$ are assumed to be conditionally independent given a vector of random effects, $\boldsymbol{\gamma}_i$ which are normally distributed. It is also assumed that all Y_{ij} have a density belonging to the exponential family with the following form

$$f(y_{ij}|\theta_{ij}, \phi) = exp\left\{\frac{y_{ij}\,\theta_{ij} - b(\theta_{ij})}{\phi} + c(y_{ij}, \phi)\right\}$$

μ_{ij} is the conditional mean of Y_{ij} that is modeled through a linear predictor, η_{ij}, containing fixed regression parameters $\boldsymbol{\beta}$, as well as subject-specific parameters $\boldsymbol{\gamma}_i$, thus

$$\begin{aligned} \eta_{ij} = g(\mu_{ij}) &= g\left[E(y_{ij}|\boldsymbol{\gamma}_i)\right] \\ &= \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\gamma_i \end{aligned} \tag{4.6}$$

or in matrix form

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} \tag{4.7}$$

where $g(.)$ is the known link function that links the conditional mean of $\mathbf{y}$ and the linear form of predictors. $\mathbf{X}, \boldsymbol{\beta}, \mathbf{Z}$ and $\boldsymbol{\gamma}$ are defined as those in Equation 2.7 from Chapter 2. Thus, it is assumed $\boldsymbol{\gamma} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{G})$ where $\boldsymbol{G}$ depends on unknown variance components.

There are two approaches used to estimate the parameters in a GLMM: the Bayesian approach and the maximum likelihood approach. The method of maximum likelihood is the most commonly used method of estimation and has a variety of optimality properties (Searle et al., 2006). Thus, this method of estimation will be focused on.

4.2.2 Maximum Likelihood Estimation

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

$$f_i(y_{ij}\,|\boldsymbol{\beta}, \boldsymbol{G}, \phi) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}\,|\boldsymbol{\gamma}_i, \boldsymbol{\beta}, \phi) f(\boldsymbol{\gamma}_i\,|\,\boldsymbol{G})\, d\boldsymbol{\gamma}_i \tag{4.8}$$

where $f(\boldsymbol{\gamma}_i\,|\,\boldsymbol{G})$ is the distribution of the random effects.

Therefore, the complete likelihood function for $\boldsymbol{\beta}, \boldsymbol{G}$ and ϕ is given by

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{G}, \phi) &= \prod_{i=1}^{m} f_i(y_{ij}\,|\boldsymbol{\beta}, \boldsymbol{G}, \phi) \\ &= \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}\,|\boldsymbol{\gamma}_i, \boldsymbol{\beta}, \phi) f(\boldsymbol{\gamma}_i\,|\,\boldsymbol{G})\, d\boldsymbol{\gamma}_i \end{aligned} \tag{4.9}$$

In the case of normality assumptions, the method of maximum likelihood for the estimation of the fixed effects in the GLMM becomes the same as that for the LMM. However, for many cases of the GLMM, the likelihood function typically does not have a closed-form expression (Jiang, 2007). This is due to the likelihood involving high-dimensional integrals that cannot be evaluated analytically. Thus, approximations are required to evaluate the likelihood function given in Equation 4.9. There have been a number of proposed methods of approximation (Hedeker, 2005), however, there are three basic approaches:

- Approximation of the integrand.
- Approximation of the integral itself.
- Approximation of the data.

Methods based on each of the above approaches are discussed in the following sections.

4.2.3 Laplace Approximation

When the exact likelihood function is difficult to evaluate, a common method used for an approximation is the Laplace approximation, which is based on an approximation of the integrand (Jiang, 2007). Suppose one wishes to approximate an integral in the form

$$\int e^{Q(\boldsymbol{x})} dx \tag{4.10}$$

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

$$Q(\boldsymbol{x}) \approx Q(\widehat{\boldsymbol{x}}) + \frac{1}{2}(\boldsymbol{x} - \widehat{\boldsymbol{x}})' Q''(\widehat{\boldsymbol{x}})(\boldsymbol{x} - \widehat{\boldsymbol{x}}) \tag{4.11}$$

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

This yields an approximation to Equation 4.10:

$$\int e^{Q(\boldsymbol{x})} dx \approx (2\pi)^{\frac{q}{2}} |Q''(\widehat{\boldsymbol{x}})|^{-\frac{1}{2}} \, e^{-\, Q'(\widehat{\boldsymbol{x}})} \tag{4.12}$$

The approximation to this integral uses as many different estimates of $\widehat{\boldsymbol{x}}$ as necessary according to the different modes of function Q. Since the $\boldsymbol{\gamma} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{G})$, it can be shown that the integral in the likelihood Equation 4.9 is proportional to the integral in Equation 4.10, where the function Q is given by

$$Q(\boldsymbol{\gamma}) = \phi^{-1} \sum_{j=1}^{n_i} \left[y_{ij}(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{\gamma}) - b(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{\gamma}) \right] - \frac{1}{2}\,\boldsymbol{\gamma}'\boldsymbol{G}\,\boldsymbol{\gamma} \tag{4.13}$$

such that Laplace's method can be applied. This approximation method tends to be better for large cluster sizes and can be improved by adding higher-order terms to the Taylor series expansion.

4.2.4 Gaussian Quadrature

Laplace approximation is based on a linearization method of the integrand. An alternative to this is an approximation of the integral or numerical integration. Two such methods are the Gauss-Hermite Quadrature and the Adaptive Gauss-Hermite Quadrature which, due to their relation with Gaussian densities, give approximations to an integral in the following form (Liu & Pierce, 1994)

$$\int h(x) e^{-x^2} dx \tag{4.14}$$

In order to apply these two methods, the likelihood contribution for the i^{th} cluster in Equation 4.8 must be represented in the form of the integral in Equation 4.14. This is done by standardizing the random effects such that they have an identity variance-covariance matrix $\boldsymbol{I}$. Let $\delta_i = \boldsymbol{G}^{-\frac{1}{2}}\,\gamma_i$. Thus, $\boldsymbol{\delta}_i$ has a normal distribution with mean $\boldsymbol{0}$ and variance-covariance matrix $\boldsymbol{I}$. The linear predictor therefore becomes $\theta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\,\boldsymbol{G}^{\frac{1}{2}}\,\boldsymbol{\delta}_i$, which now contains the variance components in $\boldsymbol{G}$. Therefore, the likelihood contribution for the i^{th} cluster is given by

$$f_i(y_{ij}\,|\boldsymbol{\beta}, \boldsymbol{G}, \phi) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}\,|\boldsymbol{\gamma}_i, \boldsymbol{\beta}, \phi) f(\boldsymbol{\gamma}_i\,|\,\boldsymbol{G})\, d\boldsymbol{\gamma}_i \tag{4.15}$$

$$= \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}\,|\boldsymbol{\delta}_i, \boldsymbol{\beta}, \boldsymbol{G}, \phi) f(\boldsymbol{\delta}_i) d\boldsymbol{\delta}_i, \tag{4.16}$$

Thus, this equation is now in the form of Equation 4.14 and therefore can be approximated using the Gauss-Hermite quadrature or adaptive Gauss-Hermite quadrature.

In Gauss-Hermite quadrature, the integral in Equation 4.14 is approximated by

$$\int h(x)e^{-x^2}dx \approx \sum_{i=1}^{L} w_i\, h(x_i) \tag{4.17}$$

where the nodes or quadrature x_i are the solutions of the L^{th} order to the Hermite polynomial with corresponding weights w_i. The values of x_i and w_i for $i = 1, 2, \ldots, 20$ are found in tables given by Abramowitz & Stegun (1972). Increasing L improves the approximation, however when the sum is taken from 1 to L, this Gauss-Hermite quadrature gives exact solutions for all polynomials of degree $2L-1$ (McCulloch & Searle, 2001). A disadvantage with this method of approximation is the quadrature points x_i are chosen independently of the function $h(x)$ and thus may result in x_i not lying in the region of interest (Pinheiro & Bates, 1995). This method can also involve summation over a large number of points, especially as the number of random effects in the model is increased (Hedeker, 2005).

To overcome the problems with the Gauss-Hermite quadrature discussed above, the quadrature points are rescaled and shifted such that the integrand in Equation 4.14 is sampled in a suitable range (Liu & Pierce, 1994). This method is referred to as the adaptive Gauss-Hermite quadrature and is based on centering the quadrature points with respect to the mode of the function being integrated and scales them according to the estimated curvature at that mode (Hartzel et al., 2001). This method requires significantly less quadrature points in order to obtain the same level of accuracy as the Gauss-Hermite quadrature. However, this adaptive Gauss-Hermite quadrature is much more time consuming to compute as the mode and curvature is calculated for each cluster in the data set (Hartzel et al., 2001). The adaptive Gauss-Hermite quadrature reduces to the Laplace Approximation when $L = 1$.

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

4.2.5 Penalized Quasi-Likelihood

This approach is based on the decomposition of the data into the mean and an appropriate error term using the Taylor series expansion of the mean. Since the mean is the inverse of the link function, it is a non-linear function of the linear predictor (Molenberghs & Verbeke, 2005). Consider the decomposition

$$\begin{aligned} Y_{ij} &= \mu_{ij} + \epsilon_{ij} \\ &= h(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\,\gamma_i) + \epsilon_{ij} \end{aligned} \tag{4.18}$$

where $h(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\,\gamma_i) = g^{-1}(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\,\gamma_i)$ is the inverse of the link function. The error terms are assumed to follow a distribution with a mean of zero and variance equal to $Var(Y_{ij}) = \phi\, v(\mu_{ij})$. Assuming the natural or canonical link function, $v(\mu_{ij}) = h'(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\,\gamma_i)$, where h' is the derivative with respect to μ_{ij}. In order to obtain an approximation of the mean, and therefore the parameters, the Taylor series expansion of Equation 4.18 is carried out. When this is done about current estimates $\widehat{\boldsymbol{\beta}}$ and $\widehat{\gamma}_i$, the method is referred to as Penalized Quasi-Likelihood (PQL) (Goldstein & Rasbash, 1996). This gives

$$\begin{aligned} Y_{ij} &\approx h(\boldsymbol{x}'_{ij}\widehat{\boldsymbol{\beta}} + \boldsymbol{z}'_{ij}\,\widehat{\gamma}_i) \\ &\quad + h'(\boldsymbol{x}'_{ij}\widehat{\boldsymbol{\beta}} + \boldsymbol{z}'_{ij}\,\widehat{\gamma}_i)\,\boldsymbol{x}'_{ij}\,(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \\ &\quad + h'(\boldsymbol{x}'_{ij}\widehat{\boldsymbol{\beta}} + \boldsymbol{z}'_{ij}\,\widehat{\gamma}_i)\,\boldsymbol{z}'_{ij}\,(\gamma_i - \widehat{\gamma}) + \epsilon_{ij} \\ &= \widehat{\mu}_{ij} + v(\widehat{\mu}_{ij})\,\boldsymbol{x}'_{ij}\,(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + v(\widehat{\mu}_{ij})\,\boldsymbol{z}'_{ij}\,(\gamma_i - \widehat{\gamma}) + \epsilon_{ij} \end{aligned} \tag{4.19}$$

where $\widehat{\mu}_{ij}$ is equal to its current predictor $h(\boldsymbol{x}'_{ij}\widehat{\boldsymbol{\beta}} + \boldsymbol{z}'_{ij}\,\widehat{\gamma}_i)$ for the conditional mean $E(Y_{ij}|\gamma_i)$. More compactly in vector form, this becomes

$$\boldsymbol{y}_i \approx \widehat{\boldsymbol{\mu}}_i + \widehat{\boldsymbol{V}}_i\,\boldsymbol{X}_i\,(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + \widehat{\boldsymbol{V}}_i\,\boldsymbol{Z}_i\,(\gamma_i - \widehat{\gamma}) + \boldsymbol{\epsilon}_i \tag{4.20}$$

where $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are appropriate design matrices and $\widehat{\boldsymbol{V}}_i$ is the diagonal matrix with elements $v(\widehat{\mu}_{ij}) = h'(\boldsymbol{x}'_{ij}\widehat{\boldsymbol{\beta}} + \boldsymbol{z}'_{ij}\,\widehat{\gamma}_i)$. Rearranging the terms in the above equation and multiplying by $\widehat{\boldsymbol{V}}_i^{-1}$ gives

$$\boldsymbol{y}^*_i \equiv \widehat{\boldsymbol{V}}_i^{-1}\,(\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_i) + \boldsymbol{X}_i\,\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}_i\,\widehat{\gamma} \approx \boldsymbol{X}_i\,\boldsymbol{\beta} + \boldsymbol{Z}_i\,\gamma + \boldsymbol{\epsilon}^*_i \tag{4.21}$$

where $\boldsymbol{\epsilon}^*_i = \widehat{\boldsymbol{V}}_i^{-1}\,\boldsymbol{\epsilon}_i$ still has a mean of zero. Equation 4.21 can be viewed as a linear mixed model for the pseudo response $\boldsymbol{y}^*_i$. Thus, methods of fitting LMMs become available in order to obtain updated estimates for $\boldsymbol{\beta}, \boldsymbol{G}$ and ϕ.

Specifically, estimates can be obtained by optimizing the quasi-likelihood function that includes a penalty term on the random effects of the form

$$\frac{1}{2}\boldsymbol{\gamma}'\boldsymbol{G}\,\boldsymbol{\gamma}$$

This results in optimizing a penalized quasi-likelihood function below

$$L_{PQL} = \sum Q_i - \frac{1}{2}\boldsymbol{\gamma}'\boldsymbol{G}\,\boldsymbol{\gamma} \tag{4.22}$$

where Q_i is McCullagh & Nelder's (1989) quasi-likelihood function. Breslow & Clayton (1993) give more information on this procedure.

4.2.6 Marginal Quasi-Likelihood

The marginal quasi-likelihood (MQL) method of approximation is very similar to the PQL method, however the Taylor series expansion of the mean in Equation 4.18 is carried out about the current estimate of $\widehat{\boldsymbol{\beta}}$ for the fixed effects but about $\widehat{\gamma}_i = \mathbf{0}$ for the random effects. Thus, the current predictor of $\widehat{\mu}_{ij}$ will be of the form $h(\boldsymbol{x}'_{ij}\widehat{\boldsymbol{\beta}})$. Therefore, the pseudo data in Equation 4.21 can be represented as

$$\boldsymbol{y}_i^* \equiv \widehat{\boldsymbol{V}}_i^{-1}(\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_i) + \boldsymbol{X}_i\,\widehat{\boldsymbol{\beta}} + \approx \boldsymbol{X}_i\,\boldsymbol{\beta} + \boldsymbol{Z}_i\,\boldsymbol{\gamma} + \boldsymbol{\epsilon}_i^* \tag{4.23}$$

which satisfies the approximate linear mixed model. The same procedure of obtaining the updated estimates of $\boldsymbol{\beta}, \boldsymbol{G}$ and ϕ for the PQL method can be followed for the MQL method, however, the resulting estimates will be referred to as marginal quasi-likelihood estimates (Breslow & Clayton, 1993).

Both the PQL and MQL methods may result in estimates of the fixed effects and variances components that are biased towards zero (Hedeker, 2005). Various methods of dealing with these bias estimates have been proposed. Beslow & Lin (1995) and Lin & Breslow (1996) proposed the inclusion of bias correction terms and Kuk (1995) proposed the use of iterative bootstrap. Goldstein & Rasbash (1996) showed that including a second order term in the Taylor series expansion improves the accuracy of the approximations.

4.2.7 Model Selection

In the case where the fixed effect parameter estimates are obtained using the numerical approximations discussed in Sections 4.2.3 and 4.2.4, inferences about these parameters can be done using the likelihood ratio test and Wald test for GLMs, or alternatively the approximate Wald-test (F-test) discussed for the survey logistic regression model in Section 3.4.6. The likelihood ratio test can be used to compare two nested models with different mean structures but with the same covariance structure. However, in the case where the estimates are obtained using the PQL and MQL methods, the likelihood ratio test cannot be used for model selection as these methods are not likelihood-based (Hedeker, 2005).

Similarly, the likelihood ratio test can also be used for comparing nested models with different covariance structures but with the same mean, and inferences on the variance components will also be valid for approximate Wald tests. However, if the variance parameter to be tested takes values on the boundary of the parameter space, the normal approximation fails and thus the test statistics for these tests will not have the traditional chi-square distribution under the null hypothesis (Zhang & Lin, 2008). Self & Liang (1987), Stram & Lee (1994) and Zhang & Lin (2008) have shown that testing the null hypothesis of no random effects can be carried out using a mixture of chi-squared distributions rather that the classical single chi-squared distribution.

In the next chapter, the results of the survey logistic regression model, generalized estimating equations and generalized linear mixed model applied to the MIS data are given.

Chapter 5

Analysis of MIS Data

5.1 Survey Logistic Regression Applied to MIS Data

The socio-economic, demographic and environmental variables discussed in the exploratory data analysis in Section 1.5 were used to model the outcome of the malaria microscopy test, however, the variables age of the child in months, age of the caregiver in years, number of household members, total number of mosquito nets used in the household and cluster altitude in metres were included in the model as continuous variables. The sampling weights were adjusted for non-response and to represent only those households included in the data set used in this thesis, where only the households that had children under the age of five years old tested for malaria were included in the sample. The analyses in this thesis were done using SAS version 9.3. Some of these SAS codes can be seen in Appendix B on page 105. Specifically, the procedure PROC SURVEYLOGISTIC was used to fit the survey logistic regression model to the data. This procedure allowed the clusters, strata and sampling weights to be specified in the analysis, where the 10 different regions of Uganda discussed in Chapter 1 represented the strata.

Selection procedures (forward, backward and stepwise) used to fit an ordinary logistic regression model can also be used to fit a survey logistic regression model. However, at this time, these selection procedures have not yet been included in SAS 9.3 for PROC SURVEYLOGISTIC. Thus, the SLR model was fitted to the MIS data using similar steps suggested by Hosmer & Lemeshow (2000) as follows:

- Perform bivariate analyses of the relationship of malaria result with the explanatory variables one at a time.

- Select the explanatory variables that have a bivariate association with malaria result at p-values less than 0.1 to go into the multivariate survey logistic regression model.

- Exclude insignificant variables from the model one at a time based on the Wald test (given in the Type III Analysis of Effects in SAS) and observe the contribution of the remaining variables to the deviance reduction.
- Continue the above step until only significant main effects are left in the model.
- Check for scientifically significant interactions among the remaining significant explanatory variables.

In addition to these steps above, information criteria such as Akaike's Information Criteria (AIC) and Schwarz Criterion (SC) can be used to compare the goodness-of-fit of two nested models. The Archer and Lemeshow goodness-of-fit test (an extension of the Hosmer-Lemeshow goodness-of-fit test) used to assess the overall goodness-of-fit of a model fitted to complex survey data, discussed in Section 3.4.6, is also not available in PROC SURVEYLOGISTIC for SAS 9.3. However, the model's predictive accuracy can be assessed using statistics such as the concordance index (*c*), Somers' D (SD), Goodman-Kruskal Gamma (GKG), and Kendall's Tau-a (KT), which are produced in the output for PROC SURVEYLOGISTIC and calculated as follows:

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

where n_c is the number of concordant pairs (a pair of observations with different observed responses is concordant if the observation with the lower ordered response value, $y = 0$, has a lower predicted mean score than the observation with the higher ordered response value, $y = 1$), n_d is the number of discordant pairs (the opposite to concordant pairs), N is the sum of observation frequencies in the data and t is the total number of pairs. The paired observations with different responses that are neither concordant nor discordant are said to be tied and is given by $t - n_c - n_d$. The concordance index *c* is equal to the area under the receiver operating characteristic (ROC) curve and ranges from 0 to 1. A value of 0 implies that there is no association. The predictive accuracy is poor if *c* is between 0.5 and 0.6, moderate between 0.6 and 0.7, acceptable between 0.7 and 0.8 and excellent if *c* is greater than 0.8. Somers' D statistic is used to determine the strength and direction of relation of the pairs of observations (UCLA: Statistical Consulting Group, 2014). It ranges from -1 (all pairs disagree) to 1 (all pairs agree). The Goodman-Kruskal Gamma statistic is calculated similarly to Somers' D, however it does not take into consideration the tied pairs. It also ranges from -1 (no association) to 1 (perfect association).

The last statistic, Kendall's Tau-a, is a modification to Somers' D and takes into account the difference between the total number of paired observations and the number of paired observations with different responses.

Table 5.1 on the next page displays the p-values and odds ratios with their 95% confidence intervals for each of the variables fitted in the bivariate analyses. The gender of the child, the number of household members and incidence of indoor residual spraying were insignificant. These three variables were therefore excluded in the selection procedure in determining the multivariate SLR model.

After performing a backward selection procedure, in order to determine the final SLR model given in Table 5.2, the remaining significant main effects (at a 5% significance level) were age of the child in months, caregiver's education level, cluster altitude in metres, type of place of residence, source of drinking water, toilet facility, main floor material and total number of mosquito nets in household. None of the availability of the household items (radio, bicycle, television, refrigerator and electricity) were significant. The caregiver's knowledge of malaria was also found to be insignificant as well as their age in years. Furthermore, main wall and roof material were found to be insignificant when included in the multivariate model. The use of a mosquito bednet was also found to be insignificant, however this variable could be explained by the inclusion of the total number of mosquito nets in the household. It was then determined if any interaction terms needed to be incorporated into the model. Two-way interaction terms of the remaining variables were fitted one at a time. Significant interaction terms that led to a large decrease in the deviance were selected. Three two-way interaction terms were selected and included in the model at the same time in order to assess the change in deviance. However, with all three included, the deviance reduced by very little. Thus, the deviances of all the combinations of two of the interaction terms included in the model were assessed. The inclusion of the interaction between source of drinking water and main floor material as well as the interaction between caregiver's education level and type of place of residence resulted in the smallest deviance with a reduction of 882.34. It was further checked if higher order interaction terms should be included in the model, however none resulted in a large enough change in the deviance to warrant being included in the model. Thus, the final SLR model is displayed in Table 5.2 on page 75.

The Taylor series approximation method was used for variance estimation of the SLR model, which is the default for SAS PROC SURVEYLOGISTIC. The concordance index (c) of the final model was 0.754, indicating that, in predicting the probability of a positive malaria result, 75.4% of the cases were predicted correctly. Thus, the predictive accuracy of the final SLR model is in an acceptable range.

Table 5.1: Unadjusted odds ratios from bivariate survey logistic regression.

Variable (p-value)	Odds Ratio (95% CI)
Age in Months (<0.0001)	1.023 (1.017, 1.030)
Gender (0.2609)	
Female	1.096 (0.934, 1.286)
Male	1
Number of Household Members (0.1706)	1.024 (0.990, 1.058)
Caregiver's Age in Years (0.0235)	1.013 (1.002, 1.025)
Caregiver's Education Level (<0.0001)	
No Education	3.035 (2.023, 4.553)
Primary	2.764 (1.973, 3.871)
Secondary	1
Higher	0.317 (0.113, 0.888)
Knowledge that mosquito bites can cause malaria (0.0015)	
No	1.493 (1.166, 1.911)
Yes	1
Knowledge of ways of avoiding malaria (0.0709)	
No	1.401 (0.972, 2.019)
Yes	1
Cluster Altitude in Metres (<0.0001)	0.997 (0.996, 0.997)
Type of Place of Residence (<0.0001)	
Rural	5.474 (2.402, 12.471)
Urban	1
Household had a Radio (0.0895)	
No	1.237 (0.968, 1.581)
Yes	1
Household had a Television (<0.0001)	
No	5.775 (3.163, 10.544)
Yes	1
Household had a Bicycle (0.0002)	
No	0.577 (0.433, 0.769)
Yes	1
Household had a Refrigerator (0.0002)	
No	4.379 (2.003, 9.574)
Yes	1
Household had Electricity (<0.0001)	
No	5.300 (2.826, 9.938)
Yes	1

Continued on next page

Table 5.1 – *Continued from previous page*

Variable (p-value)	Odds Ratio (95% CI)
Source of Drinking Water (<0.0001)	
Other	0.897 (0.083, 9.655)
Tap Water	0.161 (0.092, 0.282)
Unprotected Water	0.788 (0.536, 1.158)
Protected Water	1
Toilet Facility (<0.0001)	
No Toilet Facility	5.317 (2.668, 10.598)
Uncovered Pit Latrine	3.653 (1.893, 7.051)
Covered Pit Latrine	2.745 (1.497, 5.036)
Flush Toilet	<0.001 (<0.001, <0.001)
Other	4.280 (1.284, 14.266)
VIP Latrine	1
Main Floor Material (<0.0001)	
Cement	0.300 (0.222, 0.406)
Earth/Sand	1
Earth and Dung	0.924 (0.681, 1.254)
Other	0.392 (0.175, 0.882)
Main Wall Material (<0.0001)	
Thatch/Straw	2.890 (0.609, 13.705)
Unburnt Bricks	2.279 (1.493, 3.478)
Mud and Poles	1.117 (0.672, 1.858)
Cement Blocks	0.596 (0.264, 1.344)
Other	0.534 (0.264, 1.344)
Burnt Bricks	1
Main Roof Material (<0.0001)	
Thatch	13.811 (6.758, 28.225)
Iron Sheets	5.481 (2.605, 11.531)
Other	2.166 (0.550, 8.532)
Tiles	1
Incidence of Indoor Residual Spraying (0.2166)	
No	0.657 (0.409, 1.057)
Not Known	0.944 (0.230, 3.872)
Yes	1
Use of a bednet (<0.0001)	
No	1.650 (1.300, 2.096)
Yes	1
Number of Mosquito Nets in Household (<0.0001)	0.794 (0.712, 0.884)

Table 5.2: Type III analysis of effects for the final SLR model.

Effect	DF	Chi-Square	P-Value
Age in Months	1	160.4844	<0.0001
Caregivers Education Level	3	242.5110	<0.0001
Cluster Altitude in Metres	1	61.1079	<0.0001
Type of Place of Residence	1	219.4395	<0.0001
Source of Drinking Water	3	68.5035	<0.0001
Toilet Facility	5	387.9419	<0.0001
Main Floor Material	3	261.0907	<0.0001
Number of Mosquito Nets in Household	1	18.1391	<0.0001
Caregiver's Education Level $*$ Type of Place of Residence	3	188.7966	<0.0001
Main Floor Material $*$ Source of Drinking Water	7	421.3516	<0.0001

The parameter estimates, as well as the adjusted odds ratios (aOR) with their 95% confidence intervals, and the p-values are given in Table 5.3 on the next page. When controlling for the other variables in the final multivariate SLR model, as seen in Table 5.3, the risk of malaria still remained higher for children that are a month older (aOR = 1.033, p-value $<$ 0.0001). Similarly, the risk decreases as the cluster altitude in metres increases (aOR = 0.996, p-value $<$ 0.0001) and the number of mosquito nets in the household increases (aOR = 0.812, p-value $<$ 0.0001). However, compared to the unadjusted odds ratios, there was a large decrease in the odds ratios associated with toilet facility when controlling for other variables, where children in households with no toilet facility no longer had the highest risk of malaria, but rather those in households with uncovered pit latrines. Compared to those in households with VIP latrines, children in households with uncovered pit latrines were 1.877 times more likely to have malaria, with the 95% confidence interval for the odds ratio ranging from 1.101 to as high as 3.198. Not far behind were those children in households with covered pit latrines, where their odds of having malaria were 1.736 times the odds for those in households with VIP latrines, with the 95% confidence interval ranging from 1.002 to 3.008. The odds of malaria for those with either covered pit latrines or uncovered pit latrines were significantly different from the odds of those with VIP latrines, however the odds of those with no toilet facility was not significantly different. This differed substantially from the results of the bivariate analysis with the odds ratio associated with no toilet facility decreasing from 5.317 in the bivariate analysis to 1.529 when controlling for other factors. The odds of malaria for children with flush toilets remained significantly different from that for children with VIP latrines, however this could be a result of no children in households with flush toilets testing positive for malaria. This was also evident by the very small odds ratio and confidence interval (aOR $<$ 0.00001).

Table 5.3: Estimates and adjusted odds ratios (aOR) with 95% confidence intervals for the final SLR model.

Parameter	**Estimate**	**Std. Error**	**aOR**	**95% C.I. (aOR)**		**P-Value**
				Lower	**Upper**	
Intercept	1.845	0.736				0.0122
Age in Months	0.032	0.003	1.033	1.028	1.038	<0.0001
Caregiver's Education Level (ref = Secondary)						
No Education	-1.464	1.133	0.231	0.025	2.132	0.2310
Primary	1.222	0.319	3.392	1.817	6.332	0.0001
Higher	-10.915	0.629	<0.0001	<0.0001	<0.0001	<0.0001
Cluster Altitude in Metres	-0.004	0.001	0.996	0.995	0.997	<0.0001
Type of Place of Residence (ref = Urban)						
Rural	1.810	0.283	6.111	3.509	10.646	<0.0001
Source of Drinking Water (ref = Prot. Water)						
Other	0.514	1.218	1.671	0.154	18.191	0.6732
Tap Water	-0.603	0.358	0.547	0.271	1.103	0.0918
Unprotected Water	-0.410	0.225	0.664	0.427	1.031	0.0680
Toilet Facility (ref = VIP Latrine)						
No Toilet Facility	0.425	0.362	1.529	0.752	3.107	0.2406
Uncovered Pit Latrine	0.630	0.272	1.877	1.101	3.198	0.0206
Covered Pit Latrine	0.558	0.280	1.736	1.002	3.008	0.0491
Flush Toilet	-13.031	0.737	<0.0001	<0.0001	<0.0001	<0.0001
Other	-0.062	1.309	0.940	0.072	12.235	0.9625
Main Floor Material (ref = Earth/Sand)						
Cement	-0.585	0.252	0.557	0.340	0.914	0.0205
Earth and Dung	-0.251	0.176	0.778	0.551	1.098	0.1535
Other	-0.753	0.649	0.471	0.132	1.682	0.2464
Number of Mosquito Nets in Household	-0.208	0.049	0.812	0.738	0.894	<0.0001
Main Floor Material ∗ **Source of Drinking Water** (ref = Earth/Sand and Prot. Water)						
Cement and Tap Water	-0.509	0.475	0.601	0.237	1.525	0.2839
Cement and Unprotected Water	0.609	0.477	1.838	0.721	4.683	0.2022
Earth and Dung and Tap Water	0.344	0.646	1.411	0.398	5.005	0.5940
Earth and Dung and Unprotected Water	0.647	0.262	1.909	1.142	3.193	0.0137
Earth and Dung and Other	-17.047	0.883	<0.0001	<0.0001	<0.0001	<0.0001
Other and Tap Water	0.314	1.143	1.369	0.146	12.860	0.7832
Other and Unprotected Water	1.684	0.833	5.386	1.052	27.586	0.0433
Caregiver's Education Level ∗ **Type of Place of Residence** (ref = Secondary and Urban)						
No Education and Rural	1.802	1.132	6.060	0.660	55.684	0.1113
Primary and Rural	-0.932	0.383	0.186	0.833	6.332	0.0149
Higher and Rural	10.695	0.791	>1000	>1000	>1000	<0.0001

Even though it is unknown what the category 'other' for source of drinking water, type of toilet facility and main floor material was, it could not be left out the analysis as it would reduce the sample size by a significant amount. However, for the purpose of exploring the interaction effects, this category will not be reported. Thus, the figure below shows the estimated probability of testing positive for malaria, determined using the final SLR model's estimates, associated with main floor material and source of drinking water, excluding the categories 'other'. From the categories reported in this figure, it is clear that children residing in households with cement as the main floor material had the lowest risk of malaria, for all sources of drinking water. Children using unprotected sources of drinking water had the highest risk, however excluding those in households with just earth/sand as the main floor material. These children instead had a higher risk when using protected sources of drinking water, which is not a result one would expect. Children in households with either of the three floor materials and tap water had the lowest risk of malaria.

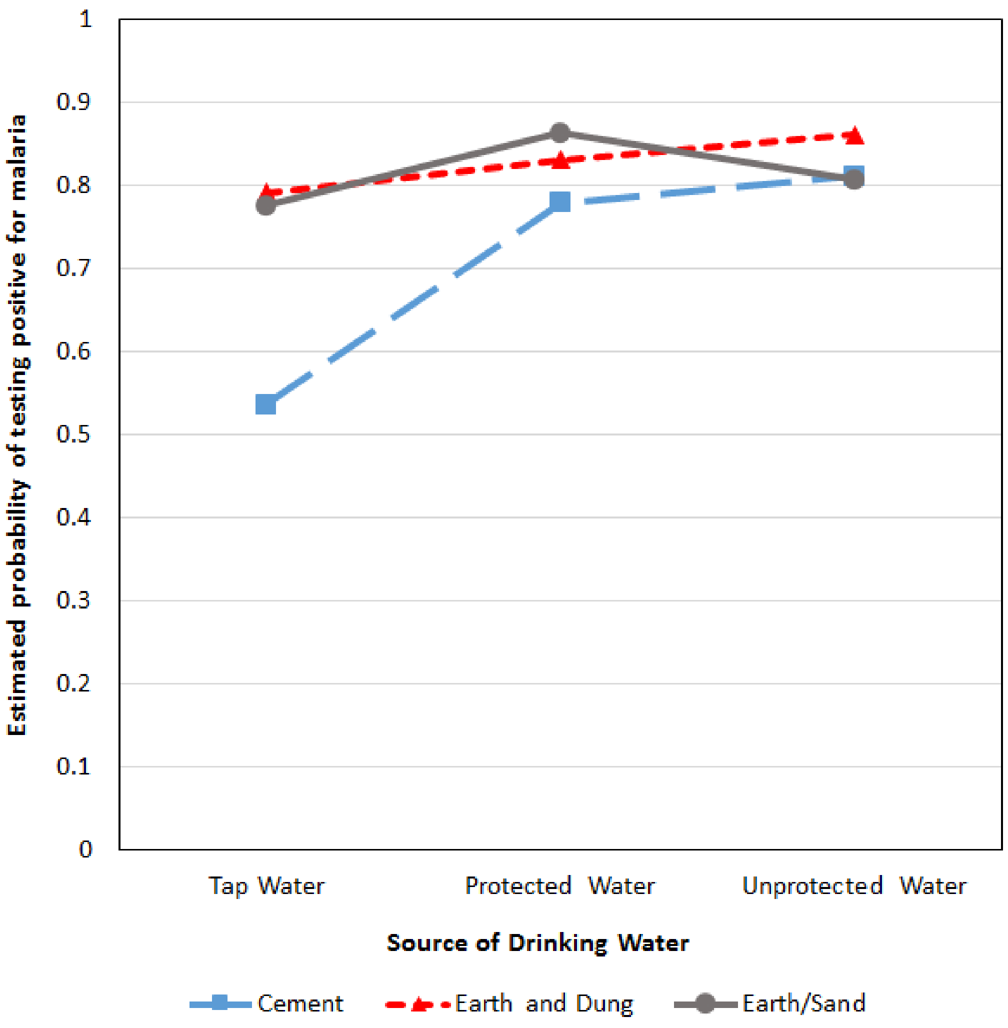

Figure 5.1: Estimated probability of testing positive for malaria associated with the interaction of main floor material and source of drinking water.

The next figure displays the estimated probability of testing positive for malaria associated with the caregiver's education level and type of place of residence. For all levels of education, children residing in rural areas had a much higher risk of malaria. The figure also reveals that children in rural areas who had caregiver's with no education were most at risk, with a slight decrease in risk as education level increased. However, in urban areas, children who had caregivers with only primary education had the highest risk of malaria. This may be a result of the very low frequency of children in urban areas in the sample who had caregivers with no education, as seen in the exploratory data analysis. Furthermore, in the exploratory data analysis it was seen that, in urban areas, no children who had caregivers with higher education tested positive for malaria, which explains the estimated probability for these children being calculated as zero.

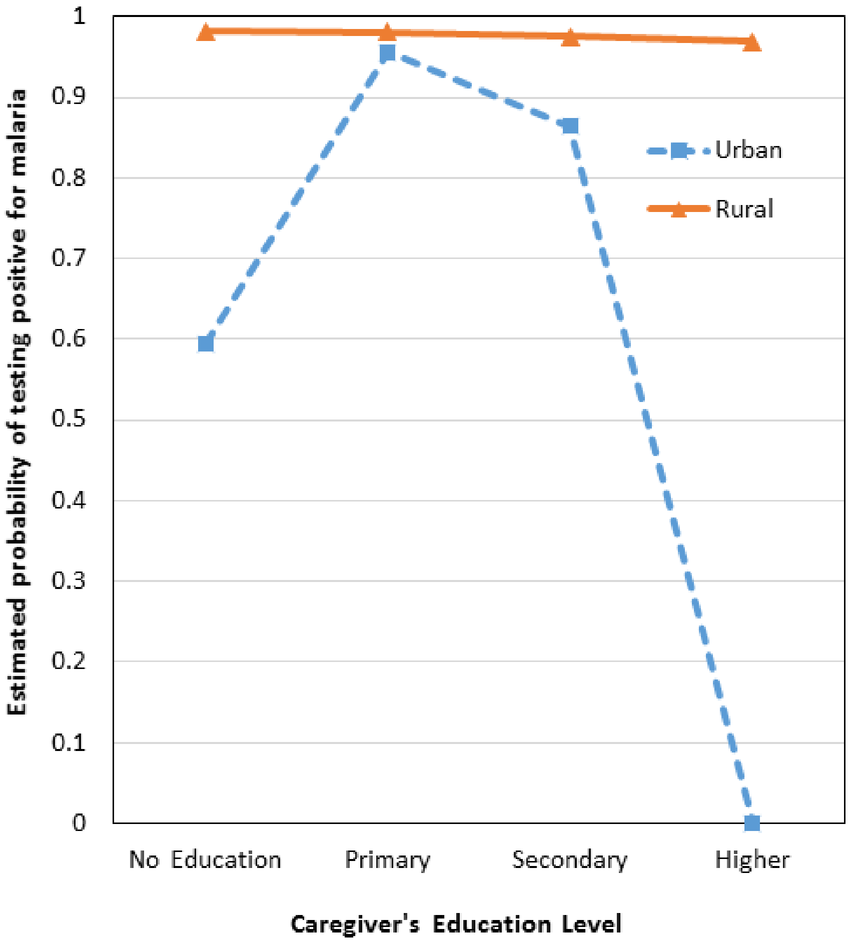

Figure 5.2: Estimated probability of testing positive for malaria associated with the interaction of caregiver's education level and type of place of residence.

The SLR model takes into consideration the design of the study. However, it does not account for effects of clustering where children within the same cluster may be more alike compared to those from different clusters. Thus, the next two analyses will take this possible clustering effect into consideration.

5.2 GEE Applied to MIS Data

The SAS procedure PROC GENMOD with the logit link function was used to fit a GEE model to the MIS data, where the REPEATED statement was used to specify the cluster in order to model the cluster effect. The full model was fitted with different correlation structures in order to determine which one best suited the data. However, region of Uganda was included as a variable in the full model, unlike the SLR model where the regions were specified in the strata. The four correlation structures fitted were Independent, Unstructured, Exchangeable and AR(1). Table 5.4 gives the QIC for each of the different correlation structures fitted. The best correlation structure for the GEE model was AR(1) as it gave the smallest QIC, even though there was not a large difference between the QIC for all the correlation structures.

Table 5.4: QIC Goodness-of-Fit Statistic for GEE.

Correlation Structure	QIC
Independent	4200.68
Unstructured	4245.91
Exchangeable	4199.68
AR(1)	4199.07

As very briefly mentioned in Section 4.1.2: Fitting the GEE Model, the variance, and hence the standard errors, of the parameter estimates can be based on two methods; an empirical method of estimation and a model-based method of estimation. The model-based standard errors are determined under the assumption that the stated correlation structure is correct. However, the empirical standard errors are based on the empirical covariance structure that is estimated directly using the data. If the correlation structure has been correctly specified, there will be very little difference between the empirical and model-based standard errors.

The GEE model was selected using a backward selection procedure according to their p-values in the Type III analysis given in the SAS output. Each time a variable was dropped from the model, the $\mathrm{QIC_u}$ was observed. The variables that minimized the $\mathrm{QIC_u}$ were selected as the main effects.

After these variables were selected, all two-way interaction effects were examined, one at a time. However, a limitation of GEEs is that if a category, after cross-classification of all the levels of the two variables, has no observations where the event of interest has occurred (e.g. there were no children that tested positive for malaria in urban areas with caregivers that had higher education), then the odds of that event occurring for that category cannot be modeled. Thus, with the inclusion of the interaction between caregiver's education level and type of place of residence, SAS produced an error in estimation. Similarly, the effect of the interaction between type of place of residence and main floor material, as well as that between type of place of residence and main wall material, could not be assessed. This may be a result of the very small sample in urban areas. In general, GEEs produce better results for large samples, therefore the results of this analysis should be interpreted with caution. Consequently, none of the other two-way interactions were significant. Thus, the final GEE model is represented in Table 5.6.

This model produced the lowest $\mathrm{QIC_u}$. Even though the variables caregiver's knowledge that mosquito bites can cause malaria and their knowledge that there are ways of avoiding malaria, were insignificant, excluding these variable from the model resulted in a large increase in the $\mathrm{QIC_u}$. Similar to the results of the SLR model, the variables age of the child in months, caregiver's education level, cluster altitude in metres, type of place of residence, source of drinking water, main floor material and total number of mosquito nets in the household were all significant. In addition to these significant variables, region of Uganda, the household had electricity, and main wall material were also found to be significant.

Table 5.5: Score statistics for Type III GEE analysis.

Effect	**DF**	**Chi-Square**	**P-Value**
Age in months	1	216.30	<0.0001
Caregivers Education Level	3	10.76	0.0124
Knowledge that mosquito bites can cause malaria	1	1.82	0.1685
Knowledge of ways of avoiding malaria	1	1.93	0.1893
Region of Uganda	9	115.40	<0.0001
Cluster Altitude in Metres	1	81.51	<0.0001
Type of Place of Residence	1	23.98	<0.0001
Household had Electricity	1	5.21	0.0070
Source of Drinking Water	3	14.52	0.0020
Main Floor Material	3	13.51	0.0027
Main Wall Material	5	12.34	0.0262
Number of Mosquito Nets in Household	1	21.48	<0.0001

The AR(1) correlation structure of the final model was once again checked. However, this correlation structure still produced the lowest QIC value compared to the other three structures. The empirical standard errors were compared with the model-based standard errors and both gave similar estimates, thus suggesting AR(1) was the correct correlation structure. SAS has the option to obtain the estimated working correlation matrix. In doing so, the non-zero elements produced in this matrix ranged from 0.0006 to 0.2257, thus indicating there existed a slight positive correlation between observations which occurred in the same cluster. This may possibly be due to similar cultural beliefs and practices that are prevalent in regions closer together.

The next table displays the results of the variables in the final GEE model, based on the empirical standard errors. The adjusted odds ratio for an increase in the child's age by a month was 1.034, which is almost in line with that produced by the SLR model (aOR = 1.033). The confidence interval for the aOR was however slightly narrower for the GEE model, (1.029, 1.038) compared to (1.028, 1.038) for the SLR model. This was also the case for cluster altitude in metres [aOR = 0.998 with 95% C.I. (0.997, 0.998) for the GEE and aOR = 0.996 with 95% C.I. (0.995, 0.997) for the SLR model] and number of mosquito nets in the household [aOR = 0.841 with 95% C.I. (0.785, 0.902) for the GEE and aOR = 0.812 with 95% C.I. (0.738, 0.894) for the SLR model]. Compared to children who had caregivers with a secondary education, those who had caregivers with no education were most at risk for malaria (aOR = 1.669), followed by those who had caregivers with only primary education (aOR = 1.453). Even though the odds of malaria for children who had caregivers with higher education were 1.445 times that for children who had caregivers with a secondary education, there was no significant difference between them (p-value = 0.3619). As one would expect, the odds of malaria were higher for children with caregivers who did not know mosquito bites can cause malaria and/or that there are ways of avoiding malaria (aOR = 1.175 and aOR = 1.174, respectively). However, there was no significant difference between those with caregivers who had and did not have that knowledge. Children in rural areas were largely more at risk for malaria compared to those in urban areas (aOR = 3.557), with the 95% confidence interval for the odds ratio ranging from 2.164 to as high as 5.849. Compared to children who resided in the South Western region of Uganda, which had the second lowest observed prevalence of malaria, children who resided in all the other regions of Uganda were more at risk, with the odds ratios ranging from 1.216 in the North East (which had the third lowest observed prevalence) to 5.619 in the East Central (which had the highest observed prevalence). Even though Kampala was solely made up of urban areas and had the lowest observed prevalence, it was still associated with a higher risk of malaria (aOR = 1.720) compared to the South Western, however there was not a significant difference between these two regions.

Table 5.6: Estimates and adjusted odds ratios (aOR) with 95% confidence intervals for the final GEE model.

Parameter	**Estimate**	**Std. Error**	**aOR**	**95% C.I. (aOR)**		**P-Value**
				Lower	**Upper**	
Intercept	-1.186	0.572				0.0396
Age in Months	0.033	0.002	1.034	1.029	1.038	<0.0001
Caregiver's Education Level (ref = Secondary)						
No Education	0.512	0.156	1.669	1.228	2.269	0.0011
Primary	0.374	0.135	1.453	1.114	1.895	0.0059
Higher	0.368	0.478	1.445	0.655	3.192	0.3619
Mosquito bites can cause malaria (ref = Yes)	0.162	0.117	1.175	0.935	1.478	0.1668
There are ways of avoiding malaria (ref = Yes)	0.161	0.122	1.174	0.923	1.494	0.1914
Region of Uganda (ref = South Western)						
Central 1	0.776	0.239	2.150	1.339	3.453	0.0015
Central 2	1.193	0.236	3.298	2.073	5.246	<0.0001
East Central	1.726	0.235	5.619	3.537	8.924	<0.0001
Kampala	0.542	0.503	1.720	0.618	4.782	0.2989
Mid Eastern	1.062	0.226	2.893	1.867	4.483	<0.0001
Mid Northern	0.624	0.254	1.870	1.137	3.075	0.0137
Mid Western	0.488	0.233	1.629	1.038	2.558	0.0339
North East	0.195	0.250	1.216	0.742	1.993	0.4386
West Nile	0.390	0.247	1.478	0.912	2.394	0.1128
Cluster Altitude in Metres	-0.002	0.00003	0.998	0.997	0.998	<0.0001
Type of Place of Residence (ref = Urban)						
Rural	1.269	0.238	3.557	2.164	5.849	<0.0001
Household had Electricity (ref = Yes)	0.681	0.274	1.976	1.168	3.342	0.0111
Source of Drinking Water (ref = Protected Water)						
Other	-0.649	0.703	0.523	0.069	3.987	0.5315
Tap Water	-0.636	0.182	0.530	0.378	0.743	0.0002
Unprotected Water	0.012	0.100	1.012	0.833	1.230	0.9018
Main Floor Material (ref = Earth/Sand)						
Cement	-0.540	0.151	0.583	0.432	0.787	0.0004
Earth and Dung	0.047	0.108	1.048	0.848	1.295	0.6662
Other	0.012	0.417	1.012	0.446	2.297	0.9771
Main Wall Material (ref = Burnt Bricks)						
Thatch/Straw	-0.266	0.385	0.767	0.372	1.583	0.4728
Unburnt Bricks	0.274	0.135	1.315	1.000	1.729	0.0496
Mud and Poles	-0.176	0.126	0.839	0.652	1.080	0.1726
Cement Blocks	0.500	0.403	1.648	0.787	3.451	0.1852
Other	-0.316	0.428	0.729	0.296	1.796	0.4921
Number of Mosquito Nets in Household	-0.173	0.035	0.841	0.785	0.902	<0.0001

Children in households with no electricity had a higher risk of malaria (aOR = 1.976). There was also a significant difference between those in households with and without electricity with the 95% confidence interval for the odds ratio ranging from 1.168 to 3.342. Tap water was associated with a lower risk of malaria with an odds of just over half of that for protected water sources (aOR = 0.530). Cement as the main floor material was also associated with a lower risk and was significantly different from just earth/sand as the main floor material. The only significantly different wall material from burnt bricks was unburnt bricks, which was associated with a higher odds of malaria (aOR = 1.315). Even though children who resided in households with cement blocks as the main floor material had the highest odds of malaria, which is a surprising result as it is associated with a higher socio-economic status compared to the other wall materials, it was not significantly different from burnt bricks.

5.3 GLMM Applied to MIS Data

In SAS, the procedure PROC GLIMMIX allows a GLMM to be fitted to the data. The RANDOM statement specifies the random effects to be included in the model. In order to account for any heterogeneity between clusters in the MIS data, an intercept term that varied at cluster level was included in the model, thus resulting in a random intercept model. Once again the logit link function was used with a binary distribution specified. The model was fitted using the Laplace approximation as this method is likelihood based and therefore allows for the comparison of models using model selection criteria such as AIC and BIC. This method was also computationally less demanding. The need for a random intercept was assessed by testing if its corresponding covariance parameter equaled zero. This was done using the COVTEST statement in SAS, which produces likelihood ratio tests for covariance parameters. Since the parameter under the null hypothesis fell on the boundary of the parameter space, the p-value for the test was determined using a linear combination of central Chi-Square probabilities. The table below shows the result of this test when fitting the full GLMM to the data. This result indicates that the null hypothesis of the covariance parameter equal to zero was rejected, thus suggesting the random cluster effect was highly significant in the model.

Table 5.7: Test of covariance parameters based on the likelihood.

Label	**DF**	**-2Log Likelihood**	χ^2	**P-value**
No G - side effects	1	4093.84	122.73	<0.0001

Notice the label SAS gives the covariance parameter of the random effect: "G-side effects". SAS distinguishes between two types of random effects, that included in the linear predictor and/or that modeling correlations among the data directly. Covariance parameters for the random components in the model that are contained in the variance-covariance matrices $\boldsymbol{G}$ and $\boldsymbol{R}$ are termed G-side and R-side effects, respectively. Thus, the inclusion of a random intercept according to the different clusters resulted in the inclusion of G-side effects in the model. R-side effects are also called "residual" effects. Models with only these effects are also known as marginal or population-averaged models, equivalent to GEE models. Inclusion of R-side effects in the model, via the SAS RANDOM_RESIDUAL_ statement, adds an overdispersion effect to the model that acts as a multiplier on the variance function, thus lifting the restriction of the dispersion parameter $\phi = 1$.

Before model selection of the fixed effects, the model was fitted with different covariance structures for $\boldsymbol{G}$ in order to determine which one best suited the data. The first structure fitted was SAS's default, VC or Variance Components, which is a simple diagonal matrix that gives a different variance component for each random effect (SAS Institute Inc., 2013). Other structures fitted were AR(1), CS and UN (unstructured). However, VC and UN produced the lowest AIC values, both of which were the same. As VC is the simplest of the two to fit, it was selected. In order to obtain the final GLMM, a backward selection procedure was carried out where insignificant fixed effects, according to the p-values of the fixed effects in the Type III analysis (determined using the Wald F-test), were removed from the model one at a time until only significant fixed effects were left. All two-way and higher order interactions were explored, however PROC GLIMMIX produced none that were significant.

Table 5.8 gives the final GLMM. The denominator's degrees of freedom was calculated as 3560. Once again, the age of the child in months, caregiver's education level, cluster altitude in metres, type of place of residence, source of drinking water, main floor material and number of mosquito nets in household were all significant, which was consistent with both the SLR and GEE model. Similar to the GEE results, region of Uganda, availability of electricity in the household, and main wall material were also found to be significant. The Pearson Chi-Square statistic over its degrees of freedom was 0.90, which is close to 1, thus indicating the variability in the data was properly modeled, and hence there was no residual overdispersion. The variance component for the random cluster effect was estimated as 0.5654 with a standard error of 0.1141. This estimate is relatively far from zero, thus confirming again the need for this random effect in the model.

Table 5.8: Type III analysis of fixed effects for the final GLMM.

Effect	Numerator DF	F-Value	P-Value
Age in months	1	225.50	<0.0001
Caregivers Education Level	3	4.12	0.0063
Region of Uganda	9	4.62	<0.0001
Cluster Altitude in Metres	1	27.1	<0.0001
Type of Place of Residence	1	11.90	0.0006
Household had Electricity	1	5.62	0.0177
Source of Drinking Water	3	3.37	0.0178
Main Floor Material	3	4.92	0.0021
Main Wall Material	5	2.27	0.0448
Number of Mosquito Nets in Household	1	33.13	<0.0001

Table 5.9 on the next page presents the parameter estimates, adjusted odds ratios with their 95% confidence intervals, and p-values for the fixed effects of the final GLMM . The GLMM produced very similar estimates to the GEE model, however, it also produced many higher standard errors, and hence wider confidence intervals. As a result, some of the levels of the variables found to be significantly different in the GEE model, were not found in the GLMM, specifically for the region of Uganda. The GEE model found a significant difference between the South Western region of Uganda and 6 out of the other 9 regions. Whereas the GLMM only found a significant difference between the South Western region and 3 of the other regions. However, like the GEE model, the GLMM also resulted in East Central having the highest risk of malaria compared to the South Western region, followed by Central 2 and Mid Eastern. In terms of the adjusted odds ratios, the South Western region remained least at risk for malaria compared to each of the other regions. The results of the GLMM for age in months, cluster altitude in metres, and number of mosquito nets in the household were very similar to that of the GEE model. Furthermore, children with caregivers who had no education remained most at risk for malaria compared to those with caregivers who had a secondary education. Again, those in households with no electricity had a higher risk of malaria. Both models gave an adjusted odds ratio of just over 1 for unprotected water sources compared to protected water sources (aOR = 1.012 for the GEE model and aOR = 1.009 for the GLMM), however neither model produced a significant difference between the two sources of water. Children in households with cement as the main floor material remained less at risk for malaria compared to those in households with just earth/sand. For the main wall material of a household, unburnt bricks, which was associated with a higher risk of malaria, was the only significantly different material compared to burnt bricks, as seen with the GEE model.

Table 5.9: Estimates and adjusted odds ratios (aOR) with 95% confidence intervals for the fixed effects in the final GLMM with one random effect.

Parameter	Estimate	Std. Error	aOR	95% C.I. (aOR) Lower	Upper	P-Value
Intercept	-1.203	0.877				0.1703
Age in Months	0.037	0.002	1.038	1.033	1.043	<0.0001
Caregiver's Education Level (ref = Secondary)						
No Education	0.554	0.160	1.740	1.273	2.380	0.0005
Primary	0.414	0.138	1.513	1.155	1.982	0.0026
Higher	0.369	0.491	1.447	0.552	3.790	0.4524
Region of Uganda (ref = South Western)						
Central 1	0.726	0.383	2.067	0.976	4.380	0.0580
Central 2	1.234	0.382	3.424	1.624	7.261	0.0013
East Central	1.714	0.387	5.552	2.600	11.857	<0.0001
Kampala	0.526	0.645	1.693	0.478	5.996	0.4145
Mid Eastern	0.974	0.365	2.648	1.295	5.413	0.0076
Mid Northern	0.583	0.412	1.792	0.798	4.022	0.1572
Mid Western	0.476	0.383	1.609	0.759	3.412	0.2143
North East	0.168	0.397	1.183	0.543	2.577	0.6725
West Nile	0.380	0.418	1.463	0.644	3.321	0.3632
Cluster Altitude in Metres	-0.002	0.0005	0.998	0.997	0.998	<0.0001
Type of Place of Residence (ref = Urban)						
Rural	1.317	0.382	3.732	1.765	7.888	0.0006
Household had Electricity (ref = Yes)	0.663	0.279	1.940	1.122	3.355	0.0178
Source of Drinking Water (ref = Protected Water)						
Other	-0.381	0.765	0.684	0.152	3.065	0.6191
Tap Water	-0.637	0.206	0.529	0.353	0.792	0.0020
Unprotected Water	0.009	0.112	1.009	0.809	1.257	0.9393
Main Floor Material (ref = Earth/Sand)						
Cement	-0.553	0.154	0.575	0.425	0.778	0.0003
Earth and Dung	0.021	0.111	1.022	0.822	1.270	0.8470
Other	0.036	0.436	1.036	0.441	2.434	0.9345
Main Wall Material (ref = Burnt Bricks)						
Thatch/Straw	-0.151	0.401	0.860	0.392	1.886	0.7062
Unburnt Bricks	0.273	0.138	1.314	1.001	1.723	0.0489
Mud and Poles	-0.165	0.135	0.848	0.651	1.105	0.2218
Cement Blocks	0.541	0.393	1.717	0.795	3.707	0.1685
Other	-0.224	0.445	0.800	0.335	1.912	0.6151
Number of Mosquito Nets in Household	-0.219	0.038	0.804	0.746	0.866	<0.0001

By adding cluster into the model as a random effect, the heterogeneity between clusters is accounted for, however there may be an extra source of variation between households within clusters. It may be the case that children within the same household are more homogeneous than those from different households, and with some households having up to six children tested for malaria, it may be necessary to account for possible correlations that may exist within the households. Therefore, households nested within clusters were further added as random effect. In fitting the full GLMM with this additional random effect, the test of covariance parameters once again concluded the G-side effects were significant (p-value < 0.0001). The AIC for this model was also lower than that for the full GLMM with only the cluster random effect. Thus suggesting inclusion of both the cluster and household nested within the cluster as random effects may be useful. The same selection procedure was carried out for the new GLMM, again using the Laplace approximation method of estimation. The resulting model was the same as the previous GLMM, however main floor material was no longer significant at 5%, but rather at a 10% significance level. A test of covariance parameters still showed the G-side effects were significant. The variance components for the cluster effect and household by cluster effect were estimated at 0.6239 and 0.5886, respectively, as seen in Table 5.10 below.

Table 5.10: Covariance parameter estimates for GLMM with two random effects.

Covariance Parameter	**Subject**	**Estimate**	**Standard Error**
Intercept	Cluster	0.6239	0.1261
Intercept	Household(Cluster)	0.5886	0.1707

The Pearson Chi-Square statistic over its degrees of freedom was 0.69, thus significantly reducing from 0.90 in the GLMM with only one random effect. Therefore, by including an extra random effect, there is a concern for underdispersion in the model, however, this is not as serious as overdispersion. The next table gives the results of the fixed effects for the final GLMM with two random effects. This new GLMM produced slightly higher standard errors compared to the GLMM with only one random effect, which is expected as it is accounting for an extra source of variation. This explains why main wall material was no longer significant at a 5% level of significance. However, both models produced similar parameter estimates. Thus, very similar conclusions can be drawn from this new model.

The last chapter in this thesis discusses the results of all the different models applied to the MIS data. This chapter also discusses the limitations to the study as well as recommendations for future research.

Table 5.11: Estimates and adjusted odds ratios (aOR) with 95% confidence intervals for the fixed effects in the final GLMM with two random effects.

Parameter	Estimate	Std. Error	aOR	95% C.I. (aOR)		P-Value
				Lower	Upper	
Intercept	-0.988	0.953				0.3015
Age in Months	0.041	0.003	1.042	1.036	1.048	<0.0001
Caregiver's Education Level (ref = Secondary)						
No Education	0.576	0.183	1.779	1.244	2.546	0.0016
Primary	0.405	0.156	1.499	1.104	2.038	0.0095
Higher	0.293	0.556	1.340	0.450	3.988	0.5988
Region of Uganda (ref = South Western)						
Central 1	0.813	0.415	2.255	0.998	5.094	0.0504
Central 2	1.276	0.415	3.583	1.589	8.080	0.0021
East Central	1.956	0.422	7.070	3.087	16.191	<0.0001
Kampala	0.492	0.705	1.636	0.410	6.526	0.4853
Mid Eastern	1.218	0.398	3.379	1.549	7.371	0.0022
Mid Northern	0.733	0.448	2.081	0.864	5.008	0.1020
Mid Western	0.332	0.414	1.393	0.618	3.140	0.4235
North East	0.126	0.431	1.134	0.487	2.641	0.7702
West Nile	0.368	0.452	1.444	0.595	3.505	0.4161
Cluster Altitude in Metres	-0.003	0.001	0.997	0.996	0.998	<0.0001
Type of Place of Residence (ref = Urban)						
Rural	1.703	0.416	5.488	2.428	12.405	<0.0001
Household had Electricity (ref = Yes)	0.791	0.312	2.204	1.194	4.069	0.0115
Source of Drinking Water (ref = Protected Water)						
Other	-0.376	0.879	0.687	0.122	3.850	0.6688
Tap Water	-0.573	0.231	0.564	0.359	0.886	0.0131
Unprotected Water	0.085	0.129	1.089	0.845	1.402	0.5103
Main Floor Material (ref = Earth/Sand)						
Cement	-0.522	0.177	0.593	0.419	0.839	0.0032
Earth and Dung	-0.022	0.128	0.978	0.761	1.257	0.8409
Other	-0.088	0.492	0.916	0.349	2.406	0.8584
Main Wall Material (ref = Burnt Bricks)						
Thatch/Straw	-0.124	0.467	0.884	0.354	2.209	0.7913
Unburnt Bricks	0.258	0.160	1.294	0.947	1.770	0.1058
Mud and Poles	-0.194	0.155	0.824	0.608	1.116	0.2106
Cement Blocks	0.638	0.452	1.893	0.779	4.597	0.1586
Other	-0.295	0.503	0.745	0.278	1.997	0.5580
Number of Mosquito Nets in Household	-0.238	0.044	0.788	0.723	0.859	<0.0001

Chapter 6

Discussion and Conclusion

The main objective of this thesis was to identify significant risk factors associated with malaria infection in children under the age of five in Uganda. This was done using three statistical approaches applied to the MIS data. The survey logistic regression model was used to account for the design of the study, where sampling weights were included in the analysis. Both generalized estimating equations and generalized linear mixed models were used in order to account for intracluster correlation that may have existed, with the generalized linear mixed model extended to account for possible correlations within households, nested within clusters. The GEE method is a population averaged approach that determines a working correlation structure within the subjects. This working correlation is assumed to be the same for all subjects, which in this case was represented by the clusters. However, the GLMM is a subject-specific approach and therefore allows the within-subject correlation to vary from one subject to another, by means of the inclusion of a random effect in the model (Carrière & Bouyer, 2002). With the GEE applied to the MIS data, where the clusters were specified as the repeated subjects, the best fitting correlation structure was AR(1). The working correlation matrix indicated that there was a slight positive correlation between children in the same clusters. Even though three two-way interaction effects in the GEE model were unable to be explored, no interaction effects were found to be significant in the GLMM, which leads one to have more confidence in the GEE results. In contrast, the SLR model produced two significant two-way interactions that resulted in a substantial reduction in the model's deviance. This may be an effect of taking the sampling weights into consideration, where observations are either up-weighted or down-weighted depending on their associated sampling weight.

The GEE model and GLMM with one random effect produced very similar results, where possible differences seen in the parameter estimates of the two models could be attributed to the two extra variables (caregiver's knowledge of malaria) maintained in the GEE model, which was selected by minimizing the QIC_u.

Both these models found the child's age in months, the caregiver's education level, region of Uganda, cluster altitude, type of place of residence, availability of electricity in the household, source of drinking water, main floor and wall material, and the number of mosquito nets in the household to be significant. With the addition of another random effect in the GLMM, in order to account for a possible correlation among children within the same household, nested within a cluster, the variable main wall material was no longer significant at 5%, but rather at a 10% level of significance. This could be seen as a result of the new GLMM's higher standard errors due to an extra source of variation being modeled. However, this GLMM with two random effects produced results relatively consistent with that of the GEE model and the GLMM with only one random effect. The survey logistic regression model, however, produced a slightly different result where, in addition to the significant two-way interactions between caregiver's education level and type of place of residence as well as that between main floor material and source of drinking water, the type of toilet facility was found to be significant, which was not found in any of the other three models fitted. Furthermore, the variables main wall material of a household and availability of electricity were not significant in the SLR model, in contrast to the GEE model and GLMM.

Despite these differences in the models, the same conclusions can be drawn. An older child was associated with a higher risk of malaria. However, their risk decreased with an increase in cluster altitude and as the number of mosquito nets in the household increased. Compared to children with caregivers who had secondary education, those with caregivers who had no education were most at risk for malaria, followed by those with caregivers who had only primary education. These results for the age of the child and caregiver's education level are consistent with those in the literature (Ghebremeskel et al., 2000; Gahutu et al., 2011). According to the SLR model, a child's risk of malaria was substantially greater in rural areas, regardless of their caregiver's education level. This higher risk of malaria for children in rural areas was also evident in the other statistical models fitted to the MIS data. The GEE model and both GLMMs produced significant results for the region of Uganda, with children in the East Central region having the highest risk of malaria, followed by the Central 2 and Mid Eastern regions. Both tap water and cement as the main floor material were associated with a lower risk of malaria, which is in line with other research (Ayele et al., 2012). Thus, poor housing conditions were associated with a higher risk of malaria. Furthermore, children in households with no electricity had a higher risk of malaria. While electricity in a household could be related to one's socio-economic status, it also contributes to an individual's way of life, where those in households with no electricity may be required to go outside more often and therefore may be more susceptible to mosquito bites, and thus malaria.

While none of the models found access to household items, such as a radio, television, bicycle and refrigerator, significant, many of these variables could be explained by the inclusion of electricity access in the model. Gender was found to be highly insignificant, with the distribution of malaria prevalence being almost the same for both males and females. This was also the case for the number of members in the household. Incidence of indoor residual spraying within the last 12 months was also found to be insignificant, however this could be as a result of the very low percentage of households that were sprayed. While the use of a bednet by the child was not significant in any of the models, which is a contrast to the findings of Gahutu et al. (2011), Baragatti et al. (2009) and Ayele et al. (2012), this variable could be explained by the number of nets in the household, which was highly significant in all the models.

The small sample sizes within the different levels of some variables were found to be very limiting. This could possibly account for some of the variables not producing significant results. This is specifically seen with the variable main roof material, where other studies have shown that it is an important contributing factor towards an individual's malaria status (Ghebremeskel et al., 2000; Ayele et al., 2012). The Malaria Indicator Survey was a cross-sectional study, therefore it was not taken into consideration whether children who tested negative for malaria had previously been infected. A future study that possibly takes this into consideration could be useful.

The results of this study largely agree with those in the literature, where a lower socio-economic status was associated with a higher risk of malaria. This could particularly be why the results revealed that children in rural areas were largely more at risk for malaria compared to those in urban areas. This study also revealed the extent of the under-development of Uganda, and how there was a great lack of knowledge of the causes of malaria, as well as possible ways of avoiding it. Mosquito net usage and incidence of indoor residual spraying were very low. Although the government of Uganda has adopted various strategies for malaria control, there is still a considerable way to go before a significant reduction in malaria can be seen. The extent of the under-development of the country presents a great challenge in the efforts of malaria reduction, especially as poor housing conditions are experienced by a vast majority of the population. As resources for malaria control in Uganda are very limited, and the different regions of the country have been shown to be unequally at risk, it is of great importance to identify the geographical areas that are most at risk through spatial modeling. This aids in the production of malaria maps, which have been recognized as an important tool for malaria control where they can effectively guide the allocation of the limited resources and interventions. This presents the future direction of this study.

References

Abramowitz, M., & Stegun, I. A. (1972). *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*. New York: Dover.

Against Malaria Foundation (2013). About malaria. http://www.againstmalaria.com/faq_malaria.aspx#statistics. [Online; accessed July-2013].

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley & Sons, Inc., 2nd ed.

Agresti, A. (2007). *An introduction to categorical data analysis*. John Wiley & Sons, Inc., 2nd ed.

Ahmad, T. (2014). Jackknife and bootsrap methods of variance estimation. In H. Chandra, U. C. Sud, K. Aditya, V. K. Gupta, & A. Bharadwaj (Eds.) *Recent Advances in Sample Survey and Analysis of Survey Data using Statistical Softwares*. Online E-Book. http://sample.iasri.res.in/ssrs/Home.htm [Online; accessed July-2014].

An, A. B. (2002). Performing logistic regression on survey data with the new SURVEYLOGISTIC procedure. *Proceedings of SAS Users Group International (SUGI)*. Paper 258-27.

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

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

Atieli, H. E., Zhou, G., Afrane, Y., Lee, M.-C., Mwanzo, I., Githeko, A. K., & Yan, G. (2011). Insecticide-treated net (ITN) ownership, usage, and malaria transmission in the highlands of western Kenya. *Parasites & Vectors*, *4:113*.

Ayele, D., Zewotir, T., & Mwambi, H. (2012). Prevalence and risk factors of malaria in Ethiopia. *Malaria Journal*, *11:195*.

Ayele, D., Zewotir, T., & Mwambi, H. (2013). The risk factor indicators of malaria in Ethiopia. *International Journal of Medicine and Medical Sciences*, *5*(7), 335–347.

Baragatti, M., Fournet, F., Henry, M.-C., Assi, S., Ouedraogo, H., Rogier, C., & Salem, G. (2009). Social and environmental malaria risk factors in urban areas of Ouagadougou, Burkina Faso. *Malaria Journal*, *8*, 13.

Bates, D. (2010). Generalized linear models. Unpublished. `http://www.math.ust.hk/~majing/GLMH.pdf` [Online; accessed August-2014].

Beslow, N., & Lin, X. (1995). Bias correction in generalized linear mixed models with single component of dispersion. *Biometrka*, *82*, 81–91.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*. 51: 279-292.

Binka, F., Kubaje, A., Adjuik, M., Williams, L., Lengeler, C., Maude, G., Armah, G., Kajihara, B., Adiamah, J., & Smith, P. (1996). Impact of permethrin impregnated bednets on child mortality in Kassena-Nankana district, Ghana: a randomized controlled trial. *Tropical Medicine and International Health*, *1*(2), 147–154.

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

Carrière, I., & Bouyer, J. (2002). Choosing marginal or random-effects models for longitudinal binary responses: application to self-reported disability among older persons. *BMC medical research methodology*, *2*(1), 15.

CDC (2012). Global health - division of parasitic diseases and malaria. `http://www.cdc.gov/malaria/malaria_worldwide/impact.html`. [Online; accessed July-2013].

Chandra, H. (2014). Logistics regression for sample surveys. In H. Chandra, U. C. Sud, K. Aditya, V. K. Gupta, & A. Bharadwaj (Eds.) *Recent Advances in Sample Survey and Analysis of Survey Data using Statistical Softwares*. Online E-Book. `http://sample.iasri.res.in/ssrs/Home.htm` [Online; accessed July-2014].

Clark, T. D., Greenhouse, B., Njama-Meya, D., Nzarubara, B., Maiteki-Sebuguzi, C., Staedke, S. G., Seto, E., Kamya, M. R., J.Rosenthal, P., & Dorsey, G. (2008). Factors determining the heterogeneity of malaria incidence in children in Kampala, Uganda. *The Journal of Infectious Diseases*, *198*(3), 393–400.

Der, G., & Everitt, B. (2006). *Statistical analysis of medical data using SAS*. Taylor & Francis Group, LLC.

Fox, J. (2002). Linear mixed models. *Appendix to R and S-PLUS Companion to Applied Regression*. Sage Publications.

Gahutu, J., Steininger, C., Shyirambere, C., Zeile, I., Cwinya-Ay, N., Danquah, I., Larsen, C., Eggelte, T., Uwimana, A., Karema, C., Musemakweri, A., Harms, G., & Mockenhaupt, F. (2011). Prevalence and risk factors of malaria among children in southern highland Rwanda. *Malaria Journal*, *10*, 134.

Gallup, J., & Sachs, J. (2001). The Economic Burden of Malaria. *American Society of Tropical Medicine and Hygiene*, *64:1*.

Ghebremeskel, T., Haile, M., Witten, K., Getachew, A., Yohannes, M., Lindsay, S., & 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*, 17–21.

Ghosh, S., & Pahwa, P. (2006). Design-based versus model-based methods: A comparative study using longitudinal survey data. In *Proceedings of Statistical Society of Canada Survey Methods Section, London, Ontario*.

Goldstein, H., & Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society*, *A159*(3), 505–513.

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

Greenwood, B. (1989). Impact of culture and environmental changes on epidemiology and control of malaria and babesiosi: the microepidemiology of malaria and its importance to malaria control. *Transactions of the Royal Society of Tropical Medicince and Hygiene*, *83*, 25–29.

Guthmann, J., Hall, A., Jaffar, S., Palacios, A., Lines, J., & Llanos-Cuentas, A. (2001). Environmental risk factors for clinical malaria: a case-control study in the Grau region of Peru. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, *95*, 577–583.

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

Hedeker, D. (2005). Generalized linear mixed models. In B. Everitt, & D. Howell (Eds.) *Encyclopedia of Statistics in Behavioral Science*. John Wiley & Sons, Inc.

Heeringa, S. G., West, B. T., & Berglund, P. A. (2010). *Applied Survey Data Analysis*. Statistics in the Social and Behavioral Sciences Series. Chapman & Hall/CRC. Taylor & Francis Group, LLC.

Hochman, S., & Kim, K. (2009). The impact of HIV and malaria co-infection: What is known and suggested venues for further study. *Interdisciplinary Perspectives on Infectious Diseases, Vol. 2009*.

Hosmer, D., & Lemeshow, S. (2000). *Applied Logistic Regression*. John Wiley & Sons, Inc., 2nd ed.

Hosmer, D. W., & Lemeshow, S. (1980). Goodness-of-fit tests for the multiple logistic regression model. *Communications in Statistics, Theory and Methods*, *A10*, 1043–1069.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression*. John Wiley & Sons, Inc, 3rd ed.

Idro, R., Aloyo, J., Mayende, L., Bitarakwate, E., John, C., & Kivumbi, G. (2006). Severe malaria in children in areas of low, moderate and high transmission intensity in Uganda. *Tropical Medicine and International Health*, *11:115-124*.

Idro, R., Bitarakwate, E., Tumwesigye, E., & Chandy, C. (2005). Clinical manifestations of severe malaria in the highlands of southwestern Uganda. *American Journal of Tropical Medicine and Hygiene*, *72:561-567*.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. Springer Science + Business Media, LLC.

Johansson, E. W., Newby, H., Renshaw, M., & Wardlaw., T. (2007). Malaria and children. `http://www.unicef.org/spanish/health/files/Malaria_Oct6_for_web(1).pdf`. [Online; accessed July-2013].

Jones, H. (1974). Jackknife estimation of functions of stratum means. *Biometrika*, *61*, 343–348.

Ker, H. W. (2014). Application of hierarchical linear models/linear mixed-effects models in school effectiveness research. *Universal Journal of Educational Research*, *2*(2), 173–180.

Kiggundu, V. L., O'Meara, W. P., Musoke, R., Nalugoda, F. K., Baghendaghe, G. K. E., Lutalo, T., Achienge, M. K., Reynolds, S. J., Makumbi, F., Serwadda, D., Gray, R. H., & Wools-Kaloustian, K. K. (2013). High prevalence of malaria parasitemia and anemia among hospitalized children in Rakai, Uganda. *PLoS ONE*, *8*(12), e82455.

Kish, L., & Frankel, M. (1974). Inference from complex samples. *Journal of the Royal Statistical Society*, *B36*, 1–37.

Koram, K. A., Bennett, S., Adiamah, J. H., & Greenwood, B. M. (1995). Socio-economic risk factors for malaria in a peri-urban area of The Gambia. *Transactions of the Royal Society of Tropical Medicince and Hygiene*, *89*, 146–150.

Kovar, J., Rao, J., & Wu, C. (1988). Bootstrap and other methods to measure errors in survey estimates. *Canadian Journal of Statistics*, *16 Suppl.*, 25–45.

Krewski, D., & Rao, J. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, *9*(5), 1010–1019.

Kuk, A. (1995). Asymptotically unbiased estimation in generalized linear models with random effects. *Journal of the Royal Statistical Society*, *B57*(2), 395–407.

Kutner, M., Nachtsheim, C., Neter, J., & Li, W. (2005). *Applied linear statistical models*. McGraw-Hill/Irwin, 5th ed.

Lee, K. (1973). Variance estimation in stratified sampling. *Journal of the American Statistical Association*, *68*, 336–342.

Liang, K., & Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22.

Lin, X., & Breslow, N. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, *91*(435), 1007–1016.

Lipsitz, S. R., Dear, K. B. G., & Zhao, L. (1994). Jackknife estimators of variance for parameter estimates from estimating equations with applications to clustered survival data. *Biometrics*, *50*(3), 842–846.

Liu, Q., & Pierce, D. A. (1994). A note on Gauss-Hermite Quadrature. *Biometrika*, *81*(3), 624–629.

Malaria Control Programme (2005 - 2010). Uganda malaria control strategic plan 2005/6 - 2009/10. `http://health.go.ug/mcp/index2.html`. [Online; accessed August-2013].

Malaria Foundation International (2013). About malaria. `http://www.malaria.org/index.php?option=com_content&task=section&id=8&Itemid=32`. [Online; accessed July-2013].

Malaria.com (2013). Uniting against malaria. `http://www.malaria.com/`. [Online; accessed July-2013].

Mayah, E. (2011). Africa: climate change exacerbates malaria menace. *All Africa*.

McCarthy, P. (1969). Pseudoreplication: half-samples. *Review of the International Statistical Institute*, *37*, 239–264.

McCullagh, P. (1983). Quasi-likelihood functions. *Annals of Statistics*, *11*(1), 59–67.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models*. Chapman & Hall, London.

McCulloch, C., & Searle, S. (2001). *Generalized, linear, and mixed models*. John Wiley & Sons, Inc.

MEASURE DHS, MEASURE Evaluation, Presidents Malaria Initiative, Roll Back Malaria and United Nations Childrens Fund (2013). Malaria indicator survey: basic documentation for survey design and implementation. Calverton, Maryland: MEASURE Evaluation.

Mendez, F., Carrasquilla, G., & Munoz, A. (2000). Risk factors associated with malaria infection in an urban setting. *Transactions of the Royal Society of Tropical Medicince and Hygiene*, *94*, 367–371.

Miller, J., & Haden, P. (2006). *Statistical analysis with the general linear model*. University of Otago, Department of Psychology, New Zealand.

Miller, R. (1974). An unbalanced jackknife. *Annals of Statistics*, *2*, 880–891.

Minakawa, N., Dida, G. O., Sonye, G. O., Futami, K., & Njenga, S. M. (2012). Malaria vectors in Lake Victoria and adjacent habitats in Western Kenya. *PLoS ONE*, *7*(3), e32725.

Ministry of Health Online (2013). Malaria control programme in Uganda. `http://health.go.ug/mcp/index2.html`. [Online; accessed August-2013].

Molenberghs, G., & Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer-Verlag: New York.

Nadimpalli, V., & Hubbell, K. (2012). Simplifying the analysis of complex survey data using the SAS survey analysis procedures. *SAS Global Forum 2012*. Paper 348-2012.

Nahum, A., Erhart, A., Maye, A., Ahounou, D., van Overmeir, C., Menten, J., van Loen, H., Akogobeto, M., Coosemans, M., Massougbodji, A., & D'Alessandro, U. (2010). Malaria incidence and prevalence among children living in a peri-urban area on the coast of Benin, West Africa: a longitudinal study. *American Society of Tropical Medicine and Hygiene*, *83:3*.

Namanya, D. B. (2013). Malaria risk factors facing Uganda's Batwa population. *The African Portal Backgrounder*, (No. 56).

National Institute of Allergy and Infectious Diseases (2007). Understanding malaria. *NIH Publication*, (07-7139).

Ndyomugyenyi, R., & Magnussen, P. (2001). Malaria morbidity, mortality and pregnancy outcome in areas with different levels of malaria transmission in Uganda: a hospital record-based study. *Royal Society of Tropical Medicine and Hygiene*, *95:463-468*.

Ndyomugyenyi, R., & Magnussen, P. (2004). Trends in malaria-attributable morbidity and mortality among young children admitted to Ugandan hospitals, for the period 1990-2001. *Annals of Tropical Medicine and Parasitology*, *98:315-327*.

Nelder, J., & Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, *A135*, 370–384.

Nevill, C., Some, E., Mung'ala, V., Mutemi, W., New, L., Marsh, K., Lengeler, C., & Snow, R. (1996). Insecticide-treated bednets reduce mortality and severe morbidity from malaria among children on the Kenyan coast. *Tropical Medicine and International Health*, *1*(2), 139–146.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, *57*, 120–125.

Patterson, H., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*, 545–554.

Peterson, I., Borrell, L., El-Sadr, W., & Teklehaimanot, A. (2009). Individual and household level factors associated with malaria incidence in a highland region of Ethiopia: a multilevel analysis. *The American Society of Tropical Medicine and Hygiene*, *80*(1), 103–111.

Pinheiro, J., & Bates, D. (1995). Approximations to the log-likelihood function in the nonlinear mixed effects model. *Journal of Computational and Graphical Statistics*, *4*, 12–35.

Population Secretariat (POPSEC) of Uganda (2012). The state of Uganda population report. Ministry of Finance, Planning and Economic Development of Uganda.

Pullan, R., Bukirwa, H., Staedke, S., Snow, R., & Baker, S. (2010). Plasmodium infection and its risk factors in Eastern Uganda. *Malaria Journal*, *9:2*.

Quenouille, M. H. (1949). Problems in plane sampling. *Annals of Mathematical Statistics*, *20*, 355–375.

Rao, J., & Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, *79*, 811–822.

Rao, J., & Wu, C. (1985). Inference from stratified samples: Second order analysis of three methods for nonlinear statistics. *Journal of the American Statistical Association*, *80*, 620–630.

Rao, J. N. K. (1997). Developments in sample survey theory: An appraisal. *Canadian Journal of Statistics*, *25*, 1–21.

Rencher, A., & Schaalje, G. (2008). *Linear models in statistics*. John Wiley & Sons, Inc.

Reyburn, H., Mbakilwa, H., Mwerinde, R. M. O., Olomi, R., Drakeley, C., & Whitty, C. J. (2007). Rapid diagnostic tests compared with malaria microscopy for guiding outpatient treatment of febrile illness in Tanzania: randomised trial. *BMJ*, *334*(7590), 403–406.

Roberts, G., Rao, J., & Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, *74*(1), 1–12.

Robins, J., Rotnizky, A., & Zhao, L. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, *90*, 106–121.

SAS Institute Inc. (2009). *SAS/STAT ® 9.2 User's Guide, Second Edition*. Cary, NC: SAS Institute Inc.

SAS Institute Inc. (2013). *SAS/STAT ® 13.1 User's Guide*. Cary, NC: SAS Institute Inc.

Schofield, C., & White, G. (1984). Engineering against insect-borne diseases in the domestic environment: house design and domestic vectors of disease. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, *78*, 285–292.

Searle, S. R., Casella, G., , & McCulloch, C. E. (2006). *Variance Components*. John Wiley & Sons, Inc.

Self, S. G., & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, *82*(398), 605–610.

Shackman, G. (2001). Sample size and design effect. NYS DOH. Presented at Albany Chapter of American Statistical Association on March 24, 2001.

Shao, J. (1992). Jackknifing in generalized linear models. *Annals of the Institute of Statistical Mathematics*, *44*(4), 673–686.

Sintasash, D., Ghebremeskel, T., Lynch, M., Kleinau, E., Bretas, G., Shililu, J., Brantly, E., Graves, P., & Beier, J. (2005). Malaria prevalence and associated risk factors in Eritrea. *The American Society of Tropical Medicine and Hygiene*, *75*(6), 682–687.

Stram, D. O., & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, *50*(4), 1171–1177.

The World Bank (2013). World Development Indicators. `http://data.worldbank.org/country/uganda#cp_wdi`. [Online; accessed July-2013].

Tukey, J. (1958). Bias and confidence in not-quite large samples (abstract). *Annals of Mathematical Statistics*, *29*, 614.

UCLA: Statistical Consulting Group (2014). Introduction to SAS. [Online; accessed November-2014].
URL `http://www.ats.ucla.edu/stat/sas/output/sas_logit_output.htm`

Uganda Bureau of Statistics (UBOS) and ICF International Inc. (2012). Uganda demographic and health survey 2011. Kampala, Uganda: UBOS and Calverton, Maryland: ICF International Inc.

Uganda Bureau of Statistics (UBOS) and ICF Macro (2010). Uganda malaria indicator survey 2009. Calverton, Maryland, USA: UBOS and ICF Macro.

Wedderburn, R. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, *61*, 439–447.

Weil, D. (2010). The impact of malaria on African development over the longue Dure. Brown University and NBER.

West, B. T., Welch, K. B., & Galecki, A. T. (2007). *Linear mixed models: a practical guide using statistical software*. Taylor & Francis Group, LLC.

WHO (2008). World Malaria Report 2008. Geneva, Switzerland: WHO.

WHO (2013). Malaria report by the secretariat at the sixty-sixth world health assembly. Tech. rep.

Wu, Z. (2005). Generalized linear models in family studies. *Journal of Marriage and Family*, *67*(4), 1029–1047.

Yung, W., & Rao, J. N. K. (1996). Jackknife linearization variance estimators under stratified multi-stage sampling. *Survey Methodology*, *22*, 23–31.

Yung, W., & 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.

Zhang, D., & Lin, X. (2008). Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics. In D. Dunson (Ed.) *Random Effect and Latent Variable Model Selection*, vol. 192 of *Lecture Notes in Statistics*, (pp. 19–36). Springer New York.

Appendix A

A.1 Sample Size Calculation

The primary sampling units were the census enumeration areas of the 2002 National Housing and Population Census. The sample was designed to provide malaria prevalence estimates for each of the 10 regions with a relative standard error (RSE) of approximately 12%. The determination of the sample size for this MIS was based on an estimated prevalence of plasmodia parasitemia in Uganda of 20% among children under the age of 5 (Uganda Bureau of Statistics (UBOS) and ICF Macro, 2010). Since the survey used a cluster design, the design effect was taken into consideration in the sample size calculation. Design effect is the loss of effectiveness when cluster sampling is used instead of simple random sampling (Shackman, 2001). This design effect was assumed to be 1.44. A response rate of 98% was expected.

The following formula was used to estimate the required sample size (n) of children under the age of 5 for each region:

$$n = \frac{p(1-p)}{se^2} \times \frac{DEFF}{R}$$

where

$p = 0.2$ (estimated 20% malaria prevalence)

$se = 12\% \times 0.2 = 0.024$ (estimated sampling error)

$DEFF = 1.44$ (design effect)

$R = 0.98$ (estimated 98% response rate)

Using the above in the formula, a total of 408 children were required for each region. Assuming the number of children under the age of 5 per household was just less than 1 (0.96 to be exact), then the required number of households per region was 408 ÷ 0.96 = 425. Thus, a total of 4,250 households for the 10 regions of the country was required.

A total of 17 EAs/clusters per region was chosen resulting in 170 clusters selected in total. Therefore, since 4,250 households were required from 170 clusters, a total of 25 households per cluster needed to be selected. It was decided that selecting 28 households per cluster would ensure that about 25 would be interviewed, which would result in the required number of children under the age of 5.

A.2 Sample Probabilities and Weights

Since the sample was not spread geographically in proportion to the population, but rather equally across the regions, sampling weights are required in the analysis of the MIS data. Since the MIS sample was obtained using a stratified two-stage cluster design, sampling weights were calculated based on sampling probabilities for each stage of the sampling. The sampling probability for the first stage, which involved selecting 17 EAs/clusters for each of the 10 regions proportional to size, is given by

$$P_{1hi} = \frac{b \times m_{hi}}{\sum m_{hi}}$$

where

$b = 17$ is the fixed number of clusters selected in each region

m_{hi} is the number of households according to the sampling frame in the i^{th} cluster for the h^{th} region, with $i = 1, ..., 17$ and $h = 1, ..., 10$

$\sum m_{hi}$ is the total number of households in the h^{th} region

In other words, P_{1hi} represents the sampling probability associated with the i^{th} cluster in the h^{th} region.

The sampling probability for the second stage, which involved obtaining a systematic sample of 28 households for each cluster, is given by

$$P_{2hi} = \frac{c}{L_{hi}}$$

where

$c = 28$ is the fixed number of households selected in each cluster

L_{hi} is the total number of households found in the listing process for the i^{th} cluster in the h^{th} region

Thus, P_{2hi} represents the sampling probability associated with a household in the selected i^{th} cluster in the h^{th} region.

Therefore, the overall selection probability of each household in the i^{th} cluster in the h^{th} region is:

$$P_{hi} = P_{1hi} \times P_{2hi}$$

And the sampling weight for each household in the i^{th} cluster in the h^{th} region is the inverse of its overall selection probability:

$$W_{hi} = \frac{1}{P_{1hi}}$$

Appendix B

SAS Codes

The SAS codes for the models used in the analysis of the MIS data are given below:

```
/******************** Final SLR Model ********************/
proc surveylogistic data = Datamalaria;
stratum  X7 / list;
cluster  Cluster;
class  X4b  X9  X10a  X11b  X17c / param=glm;
model  Y (descending)  =  X1a  X4b  X8a  X9  X11b  X10a  X17c  X10a*X17c  X9*X4b
X22a / clparm;
weight  SamplingWeight;
run;

/******************** Final GEE Model ********************/
proc genmod data = Datamalaria descending;
class  X4b  X6a  X6b  X7b  X9  X10a  X16  X17c  X18c  Cluster / param=glm;
model  Y  =  X1a  X4b  X6a  X6b  X7b  X8a  X9  X10a  X16  X17c  X18c  X22a / link=logit
dist=bin  type3 ;
repeated subject = Cluster  /  type=ar(1)  modelse  corrw ;
run;

/*************** Final GLMM with cluster random effect ***************/
proc glimmix data = Datamalaria  method=laplace;
class  X4b  X7b  X9  X10a  X16  X17c  X18c  Cluster ;
model  Y (descending)  =  X1a  X4b  X7b  X8a  X9  X10a  X16  X17c  X18c  X22a  /
link=logit dist=binary oddsratio solution ;
random intercept / subject=Cluster type=VC ;
covtest zerog ;
run;
```

```
/********** Final GLMM with cluster & household(cluster) random effects **********/
proc glimmix data = Datamalaria  method=laplace;
class  X4b  X7b  X9  X10a  X16  X17c  X18c  Cluster  HHNumber;
model  Y (descending)  =  X1a  X4b  X7b  X8a  X9  X10a  X16  X17c  X18c  X22a  /
link=logit dist=binary oddsratio solution ;
random intercept / subject=Cluster ;
random intercept / subject=HHNumber(Cluster) ;
covtest zerog ;
run;
```

where:

X1a = age of child in months
X4b = caregiver′s education level
X6a = caregiver′s knowledge that mosquito bites can cause malaria
X6b = caregiver′s knowledge that there are ways of avoiding malaria
X7 = region of Uganda
X8a = cluster altitude in metres
X9 = type of place of residence
X10a = source of drinking water
X11b = type of toilet facility
X16 = household had electricity
X17c = main floor material
X18c = main wall material
X22a = number of mosquito nets in the household