Factors associated with teenage pregnancy in Malawi

UNIVERSITY OF
KWAZULU - NATAL

INYUVESI
YAKWAZULU-NATALI

Sandile Innocent Gumede

December, 2020

Factors associated with teenage pregnancy in Malawi

by

Sandile Innocent Gumede

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

Thesis Supervisor: Ms Danielle Roberts

Thesis Co-supervisor: Ms Arusha Desai

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

Declaration - Plagiarism

I, Sandile Innocent Gumede, 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.

Sandile Innocent Gumede (Student)	Date
Ms Danielle Roberts (Supervisor)	Date
Ms Arusha Desai (Co-supervisor)	Date

Disclaimer

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

Abstract

Teenage pregnancy is a challenge that society at large is faced with. This challenge is experienced primarily in developing countries, where an estimated 21 million girls aged between 15 and 19 years old become pregnant, with approximately 12 million giving birth in 2020. In 2018, the estimated average adolescent birth rate globally was 44 births per 1000 girls aged 15 to 19 years old. However, this rate in Malawi is significantly higher at 141. There are high health, social and economic costs of teenage pregnancy, and childbearing can lead to short and long term adverse consequences for the teen parents, the child and the community. Teenage pregnancies are more likely to occur in marginalized communities, commonly driven by poverty and a lack of education and employment opportunities.

This study aimed at investigating the factors associated with pregnancy among young sexually active girls between the ages of 15 and 19 years old in Malawi. The study made use of data from a nationally representative survey, which resulted in an observed prevalence of pregnancy of 57.7% among the sexually active teenagers. Three statistical approaches were applied, namely a survey logistic regression model, a generalised linear mixed model and a spatial generalised linear mixed model. These approaches accounted for the complex survey design that was implemented during the data collection.The findings of the study outlined that age, the event of hearing of family planning on the radio, union type, socio-economic status, contraceptive use, and education level, among others, had a significant association with teenage pregnancy in Malawi. Such insight into the factors associated with and contribut-

ing to teenage pregnancy in Malawi can help all stakeholders develop policies and interventions that will address this challenge.

Acknowledgements

I would like to express gratitude to my supervisor Ms Roberts for her continuous support and motivation throughout the journey. I also thank my co-supervisor Ms Desai for her input. To my Family and every individual who has made this journey possible, I appreciate your contribution indeed. I thank God for giving me strength throughout the journey.

Contents

List of Figures

List of Tables

Abbreviations

AIDS	Acquired Immunodeficiency Syndrome
AIC	Alkaike's Information Criterion
BIC	Bayesian Information Criterion
BRR	Balanced Repeated Replication
DHS	Demographic and Health Survey
GLM	Generalised Linear Model
GLMM	Generalised Linear Mixed Model
GKG	Goodman Kruskal Gamma
GIS	Geographic Information System
HIV	Human Immunodeficiency Virus
IRWLS	Iteractively Rewreighted Least Squares
JRR	Jackknife Repeated Replication
KT	Kendalls Tau-a
MDHS	Malawi Demoghraphics Heathly Survey
MPC	Malawi Population and Household Census
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimation
MQL	Marginal Quasi-Likelihood
NGO	Non-Govermental Organisation
NSO	National Statistics Office (Mlawi)
OLR	Ordinary Logistic Regression
OR	Odds Ratio
REML	Restricted Maximum Likelihood
ROC	Reciever Operating Characteristic
PML	Pseud Maximum Likelihood
PSU	Primary Sampling Unit

PQL	Penalized Quasi-Likelihood
SLR	Survey Logistic Regression
STI	Sexually Transmitted Infections
SGLMM	Spatial Generalised Linear Mixed Model
S D	Somers D
TCM	Traditional Contraceptives Methods
VC	Variance Components
QL	Quasi-Likelihood
QIC	Quasi-Likelihood Under Independent Model Criterion
Quad	Gaussian Hermite Quadrature
UN	Unstructured

Chapter 1

Introduction

Teenage pregnancy, also known as adolescent pregnancy, occurs in females under the age of 20 (Organization et al., 2004). After her first menstrual period, a female can get pregnant by involving herself in unprotected sexual intercourse. In well-nourished females, menstrual periods often take place around the age of 12 to 13 years (Gita D., 2009). Teenage pregnancy is a global problem faced by high-, middle-, and low-income countries. However, this challenge is experienced primarily in developing countries, where an estimated 21 million girls aged between 15 and 19 years old become pregnant, with approximately 12 million giving birth (WHO, 2020). In 2018, the estimated average adolescent birth rate globally was 44 births per 1000 girls aged 15 to 19 years old (UNICEF, 2019). However, Malawi's rate is significantly higher at 141 due to the high rate of pregnancies among teenage girls, currently at just over 29% (Office/Malawi & ICF, 2015-16). The high rate of teenage pregnancies and adolescent births contribute to Malawi's persistently high fertility rates. Coupled with decreasing mortality, it creates a large portion of youth dependents and a high population growth in the country (International Bank for Reconstruction and Development / The World Bank, 2016).

There are high health, social and economic costs associated with teenage pregnancy and childbearing, as they can lead to short and long-term adverse consequences for the teen parents, the child, and the community. Globally, the second leading cause of mortality among 15 to 19-year-old females is pregnancy and childbirth complications (WHO, 2020). Teenagers who fall pregnant are also less likely to complete their secondary education. While having a child poses numerous challenges to school continuation, in Malawi, these are exacerbated by policies enacted in 1993 by the Ministry of Education, Science and Technology. Girls who become pregnant and boys who are found to impregnate a girl are required to withdraw from school, and they are only allowed to seek re-admission six months after the child's birth and re-enroll one year after the birth (Chalasani et al., 2012). However, the teen girl is often required to drop out of school to care for their baby, thus contributing to youth unemployment. In addition to being unable to work and contribute to the economy, the teen may also become reliant on a child support grant, therefore costing the government. Teenage pregnancy is also associated with higher risks of HIV infection. Malawi is considered an HIV hotspot with one of the world's highest HIV rates (Avert, 2019).

Numerous factors contribute to teenage pregnancies. In many cultures, girls are often subjected to the pressures of early marriage and childbearing. Limited knowledge and financial resources, as well as misconceptions on how to obtain and use contraceptives correctly, hinder sexually active adolescents from avoiding unplanned pregnancies (WHO, 2020). In some societies, teen girls may choose to become pregnant as they have limited educational and employment prospects. The limited studies on teenage pregnancy have found various individual-, household- and community-level factors associated with it. However, teenage pregnancy rates vary across different regions, and thus these factors may differ for different countries (Habitu et al., 2018). Hence, it is crucial to understand these factors in a local context.

1.1 Literature Review

This section gives an overview of common factors associated with and contributes to teenage pregnancy based on other research findings. Such factors include knowledge of contraceptive use among teenagers, sex education, sexual behaviour, first sex debut (age at first sex), peer influence, socio-economic status, cultural influence and religion, relationship dynamics among teenagers, and substance use.

Contraceptive Use

Many reasons influence teenagers to not use contraceptives, which includes the fear that parents will discover they involve themselves in sexual activities. Many teenagers do not want to admit that they are sexually active and therefore do not use contraceptives (Maharaj, 2006). A study found that health facilities had an adequate stock of family planning contraceptive supplies in rural Malawi, however, they were not being used. Reasons include a lack of contraceptive knowledge, beliefs, and attitude. Family planning methods were perceived to have side effects, such as prolonged menstruation, men's concerns about impotence and genital sores, weight loss or gain, and infertility (Maharaj, 2006). Traditional family planning methods were used for infertility problems. Studies show that despite knowing different types of family planning methods and awareness of their availability, use is low because considerable misinformation still prevails regarding contraceptive methods' side effects (Chipeta et al., 2010). Traditional contraceptive methods (TCMs) have been used by their ancestors for a long time in child spacing before the advent of the modern contraceptive methods. However, even with the introduction of modern methods, some women prefer and are still using TCMs, according to Rabiu et al. (2018).

A study examining factors associated with teen mothers' use of modern contraceptives after giving birth in Malawi found that 54.8% of teenage mothers are still at risk of having a repeat teenage pregnancy due to their non-use of contraceptives. This implies that less than 50% of teen mothers use contraceptives after experiencing teen

birth. It is noted that health care factors such as the use of antenatal care, awareness of pregnancy complications, attainment of primary education, and exposure to media predict teen mothers' use of modern contraceptives (Machira & Palamuleni, 2017).

A study conducted on more than 5500 males and females aged 12-24 years in Nigeria on sexual behaviour, reproductive knowledge, and contraceptive use among urban Nigerians demonstrated that sexual intercourse appears to be sporadic and unstable. The majority of these young people, mainly males, had more than one sexual partner. Only 15% of these young adults used contraceptives. They also had little information about reproductive biology. About 3 in 5 never knew that pregnancy could occur at the first sexual intercourse, and few knew that a woman's pregnancy risk varies during the menstrual cycle (Makinwa-Adebusoye, 1992).

A study on factors affecting British teenagers and contraceptive use at the first intercourse states that the age of a man at first intercourse only significantly impacts the odds of using a modern method at first intercourse. A young man's ability to communicate about contraception significantly increased those odds of contraceptive use (Stone & Ingham, 2002).

Sexual Behaviour

Being sexually active starts for most men and women in the later teenage years. For women, the median age at first intercourse is low, in which early marriage is the norm and high in Latin America and some countries of the Middle East and Southeast Asia. For men, age at first intercourse, in general, is not linked to the age at marriage in most African and Asian countries. Men start to have sex later than women (Wellings et al., 2006). Gender differences are most pronounced in the less industrialised countries.

Factors that determine variations and trends in sexual behaviour are environmental. They include shifts in poverty, education and employment, demographic trends such as the changing age structure of population and trends towards later marriage, increased migration between and within countries, globalisation of mass media advances in contraceptives, access to family planning services and public-health, and sexually transmitted disease prevention strategies (Blanc & Way, 1998).

Sex Education

A study that was conducted on evaluating the need for sex education in developing countries found that key characteristics like school attendance and literacy are crucial consideration in providing adequate knowledge that will protect teenage girls' sexual health (Singh et al., 2005). It also influences the choice of the most effective means or the channels through which such education can be delivered and therefore has relevance for the design and implementation of an intervention. The study further concludes that there was evidence on key behavioural indicators. It showed that a high proportion of young people initiate sexual activity during their teenage years, and there were gaps in knowledge about contraceptives and other protective behaviours. A substantial number of young people engage in risky behaviours. The findings provide insight into the extent of the need for comprehensive sex education starting in the early teenage years. Contextual factors such as the relatively high proportion of young people in sub-Saharan Africa who do not attend school, harder to reach, and the substantial proportions who do not have media exposure, have implications for determining how sex education can reach and benefit young people (Singh et al., 2005).

A study in South Africa points out that sex education is essential in schools to delay an early sexual debut, which contributes to early pregnancy among young people. Teenage girls who can complete school without being involved in sexual activities can make rational decisions about their sexual behaviour and possibly delay preg-

nancy. Also, girls who have been in school but choose not only to delay pregnancy but also their first sexual debut, can protect themselves. Studies show that sex education is imperative in the curriculum as it introduces the topic of sexuality, delays an early sexual debut, and promotes safer sex. Early pregnancy disrupts schooling, whereby a teenage mother will spend more time on antenatal care visits. Comprehensive sex education can delay the first sex debut and early pregnancy and promote the use of contraceptives in teenagers, according to (Mjwara, 2014).

First Sex Debut

Studies in South Africa and some countries in Africa have discovered that young people involve themselves in sexual activities at an early age (Mjwara, 2014). The age at which a young person has sex for the first time is essential because it usually marks the beginning of exposure to the risk of getting pregnant. Also, it increases the chance of getting STIs and HIV/AIDs. Studies globally have shown that young people are engaging in risky sexual behaviour. Risky sexual behaviours start to show at the young ages among teenagers, where sexual debut is often unprotected (Mjwara, 2014). A study focusing on community factors shaping early age at first sex among adolescents in Burkina Faso, Ghana, Malawi, and Uganda mentions that nearly 60% of young women in sub-Saharan Africa and 45% of young men have had sex before the age of 18 years. There is a widening gap between initiation of sexual act and marriage, resulting in a more extended period of pre-marital sexual activity. Early initiation of the sexual act and an increased period of sexual activity before marriage can lead to increased risk of both HIV and unintended pregnancies among the youths (Stephenson et al., 2014).

Peer Influence

It is believed that young people have inadequate sexual knowledge due to incorrect information or misconception about sexual relations. Peer influence and peer pressure are often cited as the most influential factors affecting adolescents' sexual

decisions. Teenagers happen to share information that is not always accurate about sex, thus, the transfer of inadequate knowledge could lead to inaccurate information, which contributes to early pregnancy. Previous research suggested that their peers' perceptions have a substantially consistent impact on young people's sexual behaviour (Mjwara, 2014).

Cultural Influence and Religion

A study in Malawi focused on the role of stigma and investigated the social consequences of unwanted pregnancy and unsafe abortion (Levandowski et al., 2012). It had the following findings of cultural influence: Malawian women in all social sectors experience social implications of unwanted pregnancy and unsafe abortion. Unwanted pregnancy often occurs in women who have limited access to contraceptives and safe abortion, negatively influencing them and their families. It was found that the impact of unwanted pregnancy and unsafe abortion was high on young women. The role of initiation ceremonies in introducing sexuality education to young people in rural areas was linked to unwanted pregnancy among young people. The Malawian census identified about 85% of Malawi as rural, showing the significance of adolescent initiation ceremonies.

The messages of engaging in sexual practices after initiation conflict with cultural taboos of pregnancy outside marriage, creating a challenging environment for young people to negotiate as they find their role as adults in the community. Abortion stigma has been found as a negative attribute ascribed to women who seek to terminate a pregnancy that marks them internally or externally inferior to ideal womanhood (Levandowski et al., 2012).

Socio-Economic Status

A study conducted in Malawi on risk factors for unwanted teenage pregnancy in Zomba points out that teenage girls' socio-economic status may lead them to involve

themselves in sexual intercourse. It revealed that 66% of adolescents had accepted money or gift in exchange for sex (Kaphagawani, 2006). In some cases, it is believed that parents may encourage their daughter into relationships with men for consumer goods, or a girl may go out with a man because her parents cannot give her the basic needs. Teenagers with unplanned pregnancies are more likely to come from a low socio-economic status than those with planned pregnancies. The parents' education level also plays a significant role, especially a mother, as she plays a role in being a role model to a teenage girl (Kaphagawani, 2006).

A study conducted in Zimbabwe on factors contributing to teenage pregnancy in a rural community mentions that financial inadequacies and social customs induce girls to stay out of school and enter into early sexual relationships. The study's findings revealed that the socio-economic background is a significant factor contributing to teenage pregnancies in the rural community of Zimbabwe (Mutara, 2015). This finding agrees with the observation by Jovbert (2008), who discusses the impact of economic challenges in developing countries. Those who live in poverty are often exposed to more "live" sexual activities due to living in small houses where there is a distinct lack of privacy for parents. Children exposed to this situation can easily engage themselves in sexual activity as soon as they enter the puberty stage (Mutara, 2015).

Relationship Dynamics

A study conducted in South Africa on relationship dynamics and teenage pregnancy focused on their association with pregnancy risk (Jewkes et al., 2001). Both groups of teenagers had been in a relationship for about two and half years and were still with their first sexual partner. The pregnant teenagers' partners were significantly older, less likely to be in school, and less likely to have other girlfriends. The pregnant teenagers were significantly more likely to have experienced forced sexual initiation and were beaten more often. They were not able to confront their partners when

they discovered that he had other girlfriends.

A study revealed a relationship between forced sexual initiation and unwillingness to confront an unfaithful partner, strongly associated with pregnancy. The gender power imbalance is a vital issue for many challenges that a woman usually faces in the relationship. The low level of control that young women have over their own lives has a critical consequence for their reproductive and sexual health. Discrimination against teenage girls places them at a disadvantage in deciding the relationship regarding contraceptives and childbearing and how their earnings must be spent. In such circumstances, the teenagers may find it difficult to express their interest and views in the relationship, hence pregnancy prevention strategies often fail (Jewkes et al., 2001).

Substance Use

The use of substances including alcohol, marijuana, and other drugs are risk-taking behaviours associated with unplanned teenage pregnancy. Teenagers drink alcohol to get drunk, to forget about their stress, and to feel good. Drug users tend to be at a greater risk for unwanted teenage pregnancy than non-drug users as they may be more sexually active, less likely to use contraceptives, and fail to make the right decisions about sex. Furthermore, teenagers who use drugs are more exposed to becoming pregnant, and are four times more likely than those who have never taken drugs, specifically in Malawi (Cavazos-Rehg et al., 2011).

1.2 Problem Statement

Teenage pregnancy is associated with multiple adverse consequences, such as higher health risks, lower educational attainment, and a lower socio-economic status. The current teenage pregnancy rate in Malawi exacerbates the country's already high birth rate, contributing to its rapid population growth. This puts additional strain

on the economy and perpetuates the cycle of poverty. More insight into the factors associated with and contributing to teenage pregnancy in Malawi can help all stakeholders develop policies and interventions that will address this challenge.

1.3 Thesis Objectives

This thesis aimed at investigating the factors associated with pregnancy in young sexually active girls between the ages 15 and 19 years old in Malawi, herein referred to as teenage pregnancy. The specific objective are:

- to investigate the prevalence of teenage pregnancy according the different regions of Malawi as well as according to various factors of interest.
- to determine the factors that are significantly associated with teenage pregnancy in Malawi, and to determine which factors contribute to an increased likelihood of teenage pregnancy using the most recent data available.

1.4 Thesis Structure

Chapter 1 provides an introduction to the thesis, outlines the significance of the study and presents the aims and objectives. Chapter 2 introduces the variables of interest and describes the data set used in this thesis, as well as presents some exploratory data analysis. Chapter 3 gives an overview of the generalised linear model, which is then extended to the survey logistic regression model. The chapter also presents the results of the application of the survey logistic regression model. Chapter 4 further extends the generalised linear model to a generalised linear mixed model and provides the results of it applied to the data used in this thesis. Chapter 5 briefly describes the spatial generalised linear mixed model with its application. Finally, chapter 6 discusses the results of the three approaches, gives conclusions and recommendations for future studies, as well as presents the limitations of the study.

Chapter 2

Data Description and Exploration

This chapter provides an overview of the study area and introduces the data set and variables of interest as well as explore the data.

2.1 Study Area

This study focuses on Malawi, a landlocked country in sub-Saharan Africa. Malawi is divided into three regions: Northern region, Central region and Southern region, which are further divided into 28 districts (Kauye & Mafuta, 2007). The country is bordered by Mozambique to the South and East, Tanzania to the East and North, and Zambia to the West (Figure 2.1). It has total area of 119,140 Kilometres Squared, of which 20% is accounted for by water bodies (Moyo & Sill, 2014).

Malawi's population stood at 17.2 million in 2016. The average household size is 4.5 members, and in three in ten households, women are head of the family. Nearly half of the Malawian population is under age 15, making it one of the region's youngest populations (National Statistical Office , Malawi; NSO). Also, nearly half of the population lives below the poverty line, most of whom live in rural areas, where 90% of households depend on rain-fed subsistence farming.

Figure 2.1: Map of Malawi. Source: http://hdl.handle.net/2263/20903

2.2 Data Description

The thesis's data set is based on the Malawi Demographic and Health Survey (MDHS) performed between October 2015 to February 2016. The survey's primary objective was to provide up to date estimates of fundamental demographic and health indicators. The survey was nationally represented and implemented a complex design for data collection, which involved a stratified multi-stage cluster sampling technique.

Specifically, the 2015-16 MDHS sample was stratified and selected in two stages. Each of the districts were stratified into urban and rural areas, from which the standard enumeration areas (SEAs) were selected with a probability proportional to their size. This was the first stage of selection. In the second stage of selection, a fixed number of 30 households per urban cluster and 33 per rural cluster were selected with an equal probability systematic selection. All women aged 15 to 49 years old who were either permanent residents of the selected households or visitors who stayed in the household the night before the survey were eligible to be interviewed. The data for MDHS 2015-16 was collected using questionnaires that were based on the DHS program's standard Demographic and Health Surveys (Office/Malawi & ICF, 2015-16). The MDHS 2015-16 had a response rate of 99%, from all households that participated in the survey.

2.3 Study Variables

Based on the MDHS data, all females between the ages of 15 and 19 years old who had ever been pregnant or who were currently pregnant were categorised as having experienced teenage pregnancy. This forms the response variable which is binary, indicating whether or not the female experienced teenage pregnancy.

The independent variables considered in this thesis comprise of a range of individual, household and geographical level factors. These variables are based on that in the literature as well as the availability in the data. These variables are as follows

- Region of Malawi
- Age
- Type of residence (Rural or Urban)
- Age at first sex

- Total number of partners
- Heard of family planning on the radio
- Union type
- Use of contraceptives
- Socio-economic status
- Education level
- Religion
- Head of household's gender
- Tested for HIV

2.4 Data Exploration

Exploratory data analysis provides a better understanding of data before any advanced statistical modelling. This section provides information regarding the sample as well as some graphical displays of the independent variables in relation to the response.

A total of 5251 females between the ages 15 to 19 years old were interviewed during the 2015-2016 MDHS, of which 29.10% had experienced pregnancy. However, 50.4% of this sample were sexually active at the time of the survey, and thus this study will only consider the event of teenage pregnancy among these sexually active females. Therefore, the final data set used in the thesis comprises of 2648 females between the ages of 15 and 19 years old. Table 2.1 displays how this sample is distributed according to the different independent variables of interest. The majority of the sample resided in the Southern region of Malawi (51.6%) and were from rural places of residence (80.7%).

Table 2.1: Distribution of the sample according to the independent variables of interest

Variable		%
Regions	Northern Region	17.4
	Central Region	31.0
	Southern Region	51.6
Type of Residence	Rural	80.7
	Urban	19.3
Age	15 years	9.5
	16 years	12.9
	17 years	18.3
	18 years	28.4
	19 years	30.8
Education Level	No Education	3.3
	Primary	70.7
	Secondary and Higher	26.0
Heard of Family Planning on the Radio	Yes	36.2
	No	63.8
Age at First Sex	Less than 13 years	4.2
	13 to 15 years	47.5
	16 to 19 years	48.3
Contraceptives Use	Yes	29.8
	No	70.2
Total Number of Partners	4+	4.0
	1 to 3	96.0
Union Type	Never in union	49.3
	Formerly in union	6.1
	Currently in union	44.6
Tested for HIV	Yes	73.1
	No	26.9
Socio-Economic Status	Poor	40.0
	Middle	18.9
	Rich	41.1
Religion	Catholic	18.3
	Christian	41.3
	Muslim	12.5
	Other	27.9
Head of Household's Gender	Male	67.8
	Female	32.2

The two oldest age groups (18 and 19 years) contained more than half of the sample (59.2%), while only 26% of the sample had completed a secondary or higher education level (Table 2.1). When asked whether they had heard about family planning on the radio within a few months prior to the survey, 63.8% of the sample said no. Only 4.2% of the sample had their first sexual debut under 13 years of age, with the majority (48.3%) indicating that their first sexual debut was between the ages of 16 and 19 years. An overwhelming proportion of the sample indicated they were not on any contraceptive (70.2%). However, only 26.9% indicated that they had never undergone an HIV test prior to the survey. Most of the sample (96%) indicated that they had only had between 1 and 3 sexual partners. In addition, most of the sample had never been in a union (49.3%), while 44.6% were in a union at the time of the survey. Socio-economic status was based on the household's wealth quintile, where the category 'poor' corresponds to those in the two poorest quintiles, the category 'middle' corresponds to the third quintile, and the category 'rich'; corresponds to the two richest quintiles. The majority of the sample came from poor (40.0%) or rich (41.1%) socio-economic backgrounds. The primary religion in the sample was Christianity (41.3%). Furthermore, the majority resided in households headed by males (67.8%).

The overall observed prevalence of teenage pregnancy among the sexually active females in this sample was 57.7%. Figure 2.2 shows how this observed prevalence varies according to the three regions of Malawi. The Northern region had the highest prevalence at 62.3%. Not much difference is observed in the prevalence between the Central and Southern regions. Figure 2.3 presents the observed prevalence of teenage pregnancy according to the type of place of residence. The prevalence of teenage pregnancy was highest among those residing in rural areas (59.9%). While the prevalence among those in urban areas was a lot lower, it was still alarmingly high at 48.4%.

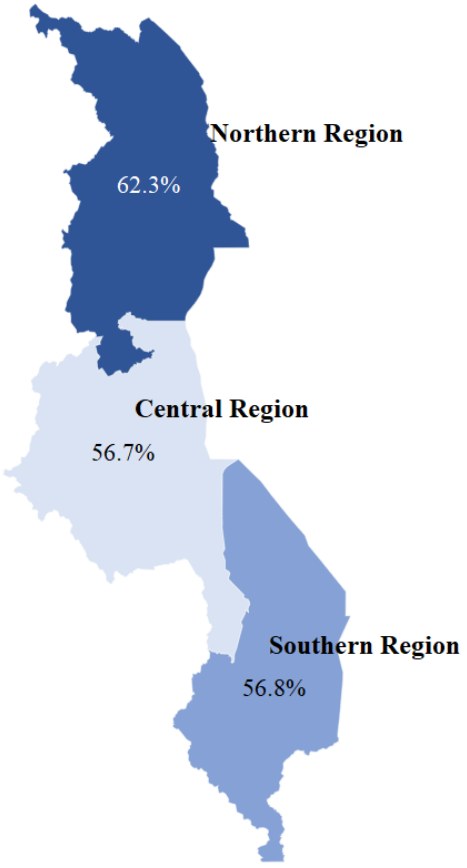

Figure 2.2: Observed prevalence of teenage pregnancy according to the regions of Malawi

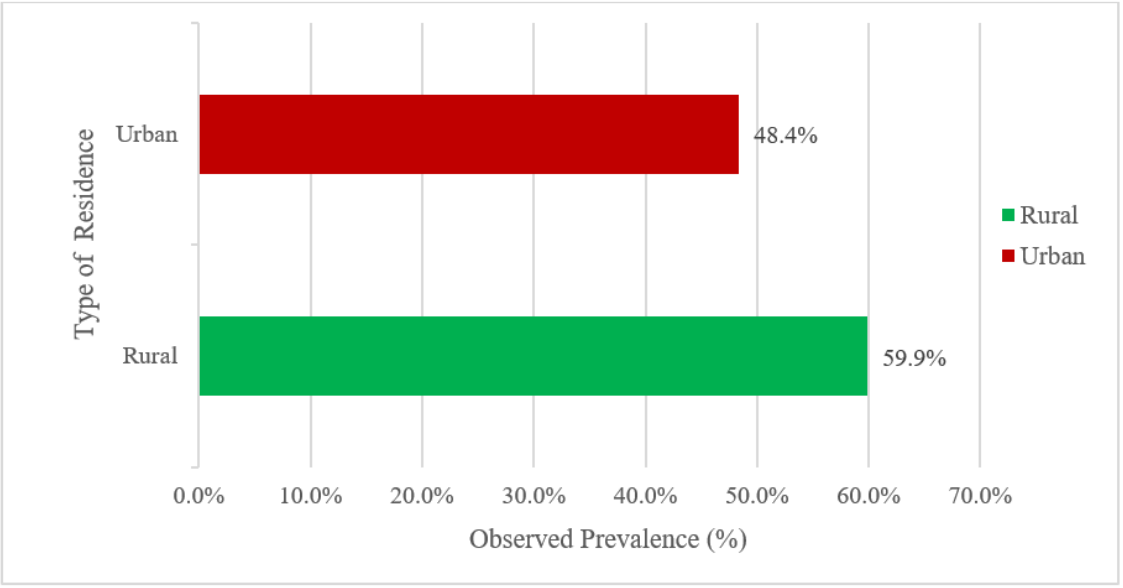

Figure 2.3: Observed prevalence of teenage pregnancy according to type of residence

Figure 2.4 below clearly demonstrates that as age increased, there was an increase in the observed prevalence of teenage pregnancy, with the oldest age group having a substantially high prevalence of 74.4%. It must be noted however that these ages represent the age of the participant at the time of the survey, and not the age at which they fell pregnant. This information was not available in the data.

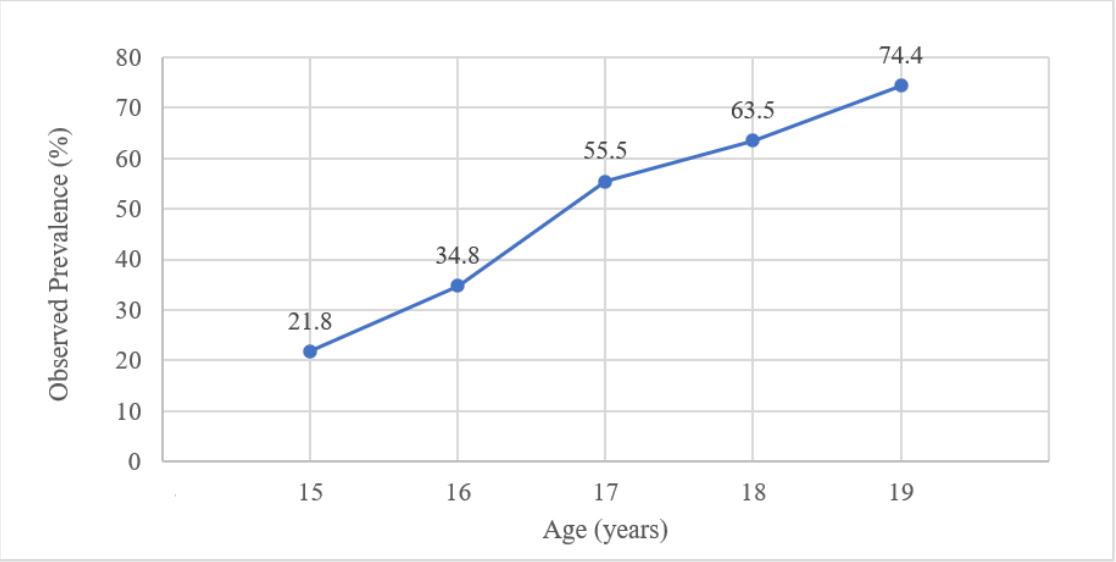

Figure 2.4: Observed prevalence of teenage pregnancy according to age in years

Figure 2.5 presents the observed prevalence of teenage pregnancy according to age at first sex, contraceptive use, total number of partners and union type. Girls whose sexual debut was between the ages of 13 and 15 years had the highest prevalence of 58.5%, which is followed by girls whose sexual debut was in the older age group (57.9%). Unsurprisingly, the observed prevalence of teenage pregnancy among those who had four or more sexual partners was highest at 61.7%. What is surprising however, is the observed prevalence among those who indicated they use contraceptives (78.5%), which is substantially higher compared to that among those who indicated they do not use contraceptives (48.9%). However, we are reminded that less than 30% of the sample indicated they were on contraceptives (Table 2.1). This high prevalence of pregnancy among those on contraceptives may indicate either a lack of knowledge of how to correctly use them or that these girls have a false sense

of security against unplanned pregnancy. The observed prevalence of teenage pregnancy among those who had been previously tested for HIV was at a staggering 71.3% compared to only 20.8% among those who had not been previously tested. However, this may be a consequence of HIV testing during antenatal care. No information regarding whether HIV testing was performed before, during or after pregnancy was available in the data. The highest prevalence was observed among those who were currently or formally in a union, at 86.8% and 91.4%, respectively. This suggests that many of these pregnancies may have been planned.

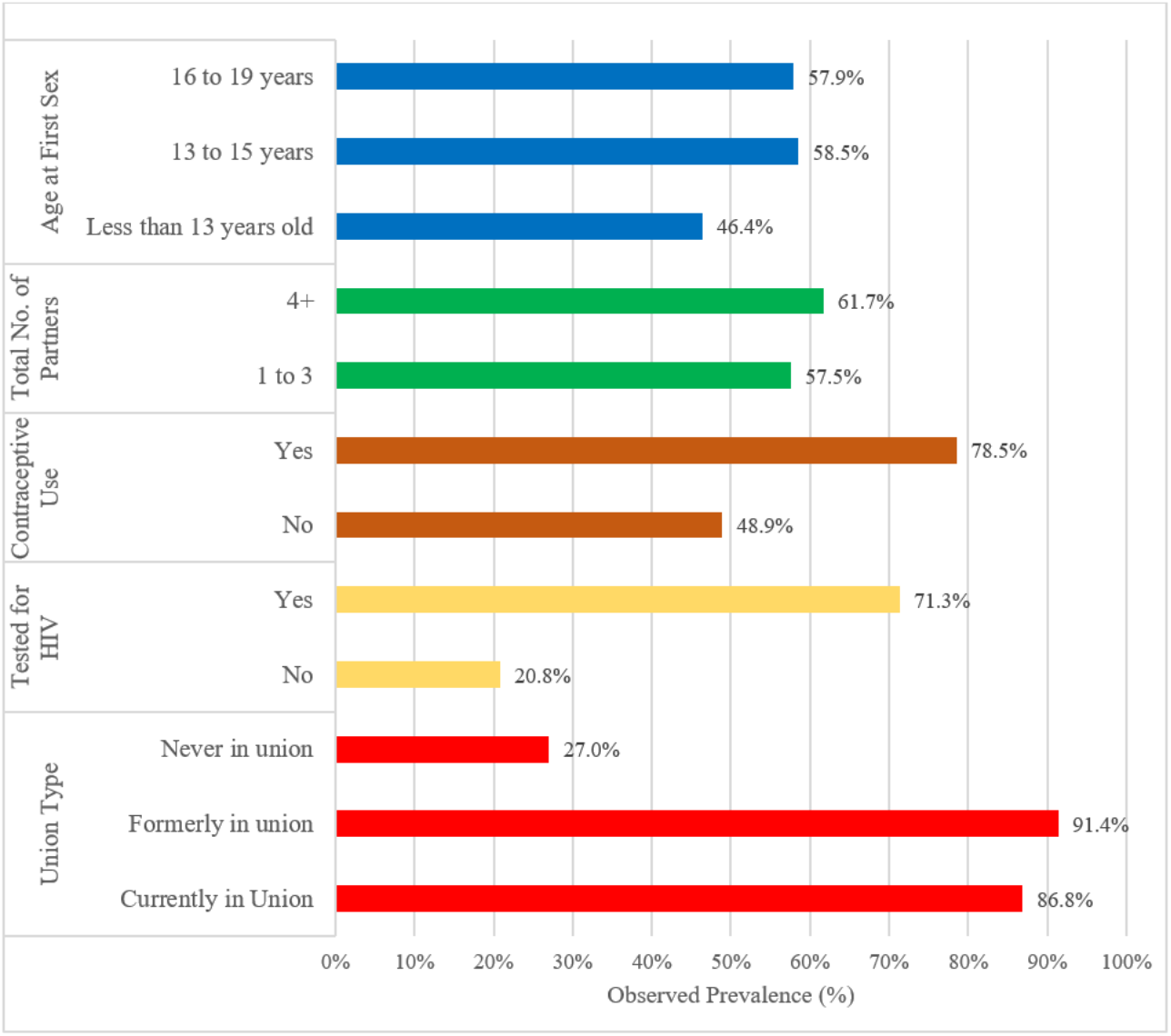

Figure 2.5: Observed prevalence of teenage pregnancy according to age at first sex, contraceptive use, total number of partners and union type

Figure 2.6 below gives the observed prevalence of pregnancy according to the individual's religion, highest education level and socio-economic status. The highest prevalence was observed among the Muslim and Christian religions at 63.9% and 60.8%, respectively. This may be due to the customs of early marriage in these religions, Although Catholics and Christians are widely known as religions with common or similar practices, in the context of the MDHS data they are classified as separate religions. Those without formal education had the highest observed prevalence of 78.2% and those who had completed secondary school or higher had a substantially lower observed prevalence of 39.3%. This indicates how education may play a role in making more informed decisions that do not lead to teenage pregnancy. The prevalence was highest among those from low socio- economic backgrounds (69.5%).

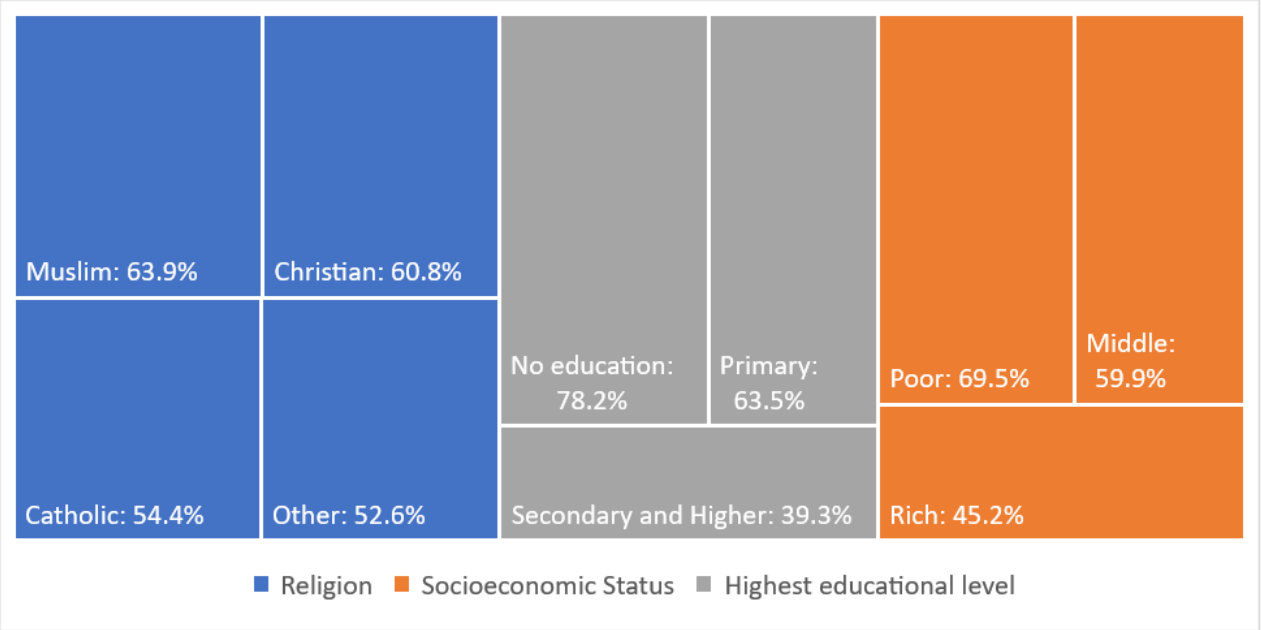

Figure 2.6: Observed prevalence of teenage pregnancy according to religion, highest education level and socio-economic status

Figure 2.7 presents the observed prevalence of teenage pregnancy according to the gender of the head of household. There was a higher prevalence among those residing in households headed by males (60.3%). Lastly, Figure 2.8 illustrates the ob-

served prevalence of teenage pregnancy according to whether or not family planning was heard on the radio. There was no substantial difference in these prevalences, however the prevalence was higher among those who had not recently heard about family planning on the radio (60.3%).

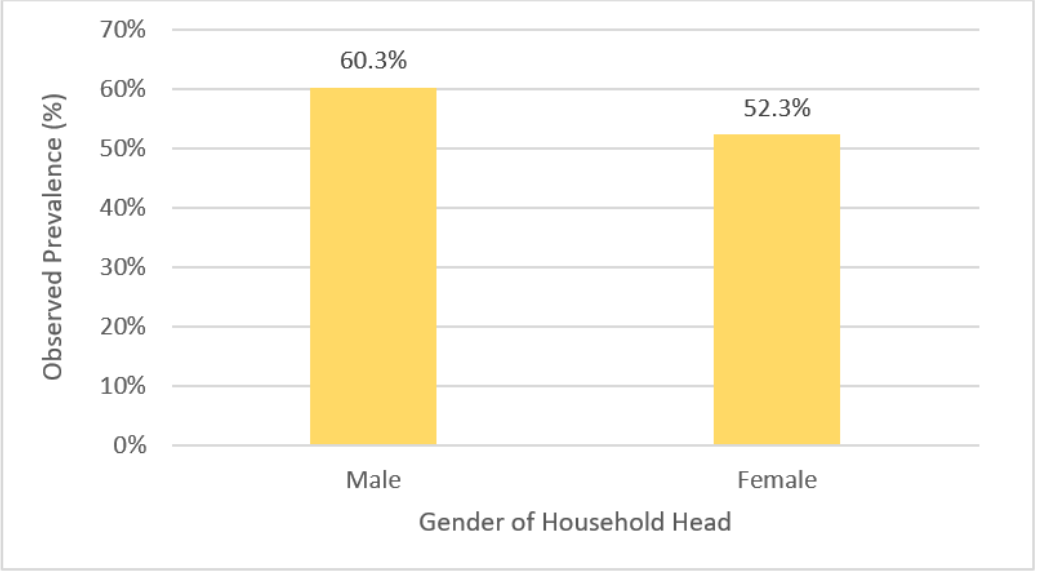

Figure 2.7: Observed prevalence of teenage pregnancy according to the gender of the head of household

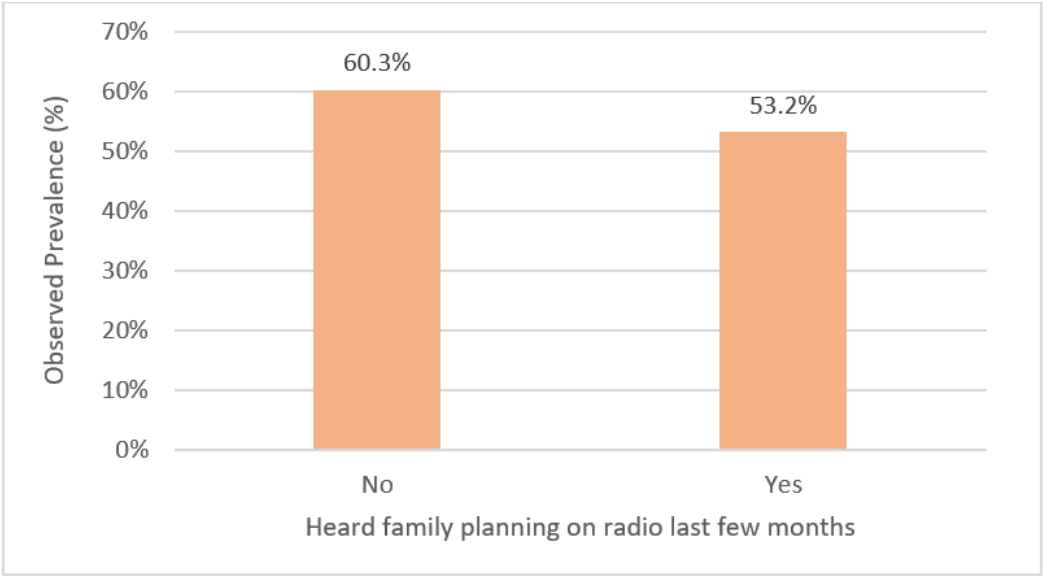

Figure 2.8: Observed prevalence of teenage pregnancy according to whether or not family planning was heard on the radio

2.5 Summary

This chapter investigated patterns in the data regarding teenage pregnancy among the sexually active females in the sample. Several individuals, household, and geographical level factors were considered. These same factors present the independent variables considered in the statistical models to follow.

The next three chapters present an overview of the statistical approaches used. Three approaches were considered: a design-based approach, a model-based approach, and a model-based approach accounting for spatial variation. All of these approaches account for the complex survey design used to obtain the data. The design-based approach uses the sampling weights to obtain the parameter estimates and their variance estimates, where the weights are equal to the inverse of the probability of selection. The model-based approach accounts for the clustering in the data, where possible correlations may exist in the observations. Individuals in the same cluster may tend to be more alike than those from different clusters. The extension of the model-based approach to account for spatial variation further accounts for possible correlations based on the observations' proximity.

Chapter 3

Generalised Linear Models

The general linear model is commonly used to model a continuous response that assumes a normal distribution. Thus, the linear model is not appropriate in the case of a discrete, binary outcome as predictions using this model can fall outside the range of the response variable. Rather, a generalised linear model (GLM) is used to model a non-normal response through a transformation function called a link function (Nelder & Wedderburn, 1972).

3.1 The 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. For a response Y_i with a distribution belonging to the exponential

family, its mean and variance are given by

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

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

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

Equation 3.3 can therefore be expressed in the form

$$Var(Y_i) = a_i(\phi)\, v(\mu_i) \tag{3.4}$$

$$= \frac{\phi}{w_i}\, v(\mu_i) \qquad \text{since } a_i(\phi) = \phi/w_i \tag{3.5}$$

This means that 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 consists of the following three components:

- *The Random Component:*

  It is assumed that $y_1, \ldots, y_n$ are samples of independent random variables $Y_1, \ldots, Y_n$, respectively. The response variable Y_i belongs to the exponential family with probability distribution in the form given in Equation 3.1.

- *The Systematic Component:*

  This component relates a vector $\boldsymbol{\eta} = (\eta_1, \eta_2, ..., \eta_n)'$ to a set of explanatory variables through a link function. Let $\boldsymbol{x}_i = (1, x_{1i}, ..., x_{pi})'$ be a $(p+1)$-dimensional vector of covariates and $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$ be a vector of the unknown regres-

sion coefficients. Then, the distribution of Y_i depends on $\boldsymbol{x}_i$ through the linear predictor, η_i, such that

$$\begin{aligned}\eta_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} \\ &= \boldsymbol{x}_i' \boldsymbol{\beta}\end{aligned}$$

- *The Link Function:*

  This component is a monotonic and differentiable function, g, which links the mean response $\mu_i = E(y_i)$ to the linear predictor $\eta_i = \boldsymbol{x_i'}\boldsymbol{\beta}$ as follows

$$\eta_i = g(\mu_i) = \boldsymbol{x}_i' \boldsymbol{\beta}$$

The exponential family comprised of numerous distributions, such as the Binomial, Poisson, Gamma and Chi-Square distribution, each of which has its own unique canonical link function. Binomial distribution has a logit link function. A GLM with a logit link is referred to as a logistic regression model, which will be discussed in Section 3.3.

3.1.1 Parameter Estimation

The method of maximum likelihood is used for the parameter estimation in GLMs. 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

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

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

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

From Equation 3.6, we get

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

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

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

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

Thus,

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

where W_i is referred to as the iterative weights, which is given by

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

where $v_i = v(\mu_i)$ is the variance function, $\dfrac{\partial \mu_i}{\partial \eta_i}$ depends on the link function of the model as $\eta_i = g(\mu_i)$. Therefore, solving for the equation below will give the ML estimate for $\boldsymbol{\beta}$

$$\sum_{i=1}^{n}(y_i - \mu_i)W_i\frac{\partial \eta_i}{\partial \mu_i}x_{ij} = 0 \tag{3.9}$$

The above equation is a non-linear function of $\boldsymbol{\beta}$. Therefore, iterative procedures such as Newton Raphson and Fisher Score are required to solve this equation.

The Newton Raphson iterative equation is given by

$$\widehat{\boldsymbol{\beta}}^{\,(t+1)} = \widehat{\boldsymbol{\beta}}^{\,(t)} - (\boldsymbol{H}^{(t)})^{\,-1}\,\boldsymbol{U}^{(t)} \tag{3.10}$$

and the Fisher Score iterative equation is given by

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

with information matrix

$$\boldsymbol{\mathcal{I}} = -E(\boldsymbol{H}) \tag{3.12}$$

$$= -E\left(\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}\,\partial \boldsymbol{\beta}'}\right) \tag{3.13}$$

$$= \boldsymbol{X}'\,\boldsymbol{W}\,\boldsymbol{X} \tag{3.14}$$

where $\boldsymbol{W}$ is known as the weight matrix with diagonal elements given in Equation 3.8. Equation 3.11 can also be represented as

$$\boldsymbol{\mathcal{I}}^{(t)}\,\widehat{\boldsymbol{\beta}}^{\,(t+1)} = \boldsymbol{\mathcal{I}}^{(t)}\,\widehat{\boldsymbol{\beta}}^{\,(t)} + \boldsymbol{U}^{(t)} \tag{3.15}$$

It can be shown that the right hand side of Equation 3.15 can be written as

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

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

evaluated at $\widehat{\boldsymbol{\beta}}^{(t)}$

$$z_i = \eta_i + (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right) \tag{3.16}$$

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

$$\widehat{\boldsymbol{\beta}}^{(t+1)} = (\boldsymbol{X}' \boldsymbol{W}^{(t)} \boldsymbol{X})^{-1} \boldsymbol{X}' \boldsymbol{W}^{(t)} \boldsymbol{z}^{(t)} \tag{3.17}$$

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 of this estimate of $\boldsymbol{\beta}$ is the inverse of the information matrix given in Equation 3.14 and can be estimated by

$$\widehat{Var}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}' \widehat{\boldsymbol{W}} \boldsymbol{X})^{-1} \tag{3.18}$$

where $\widehat{\boldsymbol{W}}$ is $\boldsymbol{W}$ 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

$$\widehat{\phi} = \frac{1}{n-p-1} \sum_{i=1}^{n} \frac{w_i \, (y_i - \widehat{\mu}_i)^2}{v(\widehat{\mu}_i)} \tag{3.19}$$

where w_i is the weight defined in Equation 3.1.

3.1.2 Measure of Fit

Measures of fit assists in assessing the goodness-of-fit of the model that is used in statistical analyses. The *deviance* is commonly used for the goodness-of-fit in a GLM. The *deviance* is a measure of discrepancy between the predicted values from the fitted model and the actual values from the data set. Suppose for the fitted model with $p+1$ parameters, $\ell(\widehat{\boldsymbol{\mu}}, \phi, \boldsymbol{y})$ is the log-likelihood function maximized over $\widehat{\boldsymbol{\beta}}$ for a fixed value of the dispersion parameter ϕ, and $\ell(\boldsymbol{y}, \phi, \boldsymbol{y})$ is the maximum log-likelihood achievable under the saturated model where the number of parameters equals the number of observations, the scaled deviance is

$$D^s = \frac{-2[\ell(\widehat{\boldsymbol{\mu}}, \phi, \boldsymbol{y}) - \ell(\boldsymbol{y}, \phi, \boldsymbol{y})]}{\phi} \tag{3.20}$$

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

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

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 often applied measure of goodness-of-fit is the *generalised Pearson's chi-square statistic* given by

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}_i)^2}{v(\widehat{\mu}_i)} \tag{3.22}$$

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 $\frac{\chi^2}{\phi}$ (Wu, 2005).

3.1.3 Likelihood Ratio Test

If the objective is to determine whether a particular variable in the model has no effect on the response variable, given the other variables in the model, we can test if the corresponding regression parameter is equal to zero. 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

$$D_{\text{reduced}} - D_{\text{full}} \tag{3.23}$$

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

$$\chi^2 = -2[\text{log-likelihood(reduced model)} - \text{log-likelihood(full model)}] \tag{3.24}$$

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 3.24 would become

$$T = \frac{-2[\text{log-likelihood(reduced model)} - \text{log-likelihood(full model)}]}{\phi} \tag{3.25}$$

When $\phi \neq 1$ and unknown, the value of ϕ can be estimated using Equation 3.19.

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

$$z_0 = \frac{\widehat{\beta}_j}{se(\widehat{\beta}_j)} \tag{3.26}$$

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 3.14. 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). Therefore, 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 needs 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). For such a case, the quasi-likelihood (QL) function can be used to estimate 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, for example 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 needed (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 $\boldsymbol{\beta}$. The following relation is used to determine the

quasi-likelihood (specifically the quasi-log likelihood) function $Q(y_i;\mu_i)$ for each observation

$$\frac{\partial Q(y_i;\mu_i)}{\partial \mu_i} = \frac{w_i\,(y_i-\mu_i)}{\phi\, v(\mu_i)} \tag{3.27}$$

where w_i is the known weight associated with observation Y_i.

Therefore, from the above equation, we can obtain

$$Q(y_i;\mu_i) = \int_{y_i}^{\mu_i} \frac{w_i\,(y_i-t)}{\phi\, v(t)}\,dt + \text{some function of } y_i \tag{3.28}$$

The maximum quasi-likelihood estimates of $\boldsymbol{\beta}$ can then be obtained from Equation 3.28 using Fisher Scoring. The estimate of ϕ can be obtained using Equation 3.19.

3.3 Ordinary Logistic Regression

Suppose we have a binary response variable given by

$$Y_i = \begin{cases} 1 & \text{if an event is observed, e.g. teenage pregnancy has been experienced} \\ 0 & \text{if an event is not observed, e.g. teenage pregnancy has not been experienced} \end{cases}$$

It then follows that Y_i has a Bernoulli distribution with $P(Y_i=1)=\pi_i$ and $P(Y_i=0)=1-\pi_i$. Therefore,

$$E(Y_i) = \pi_i \quad \text{and} \tag{3.29}$$

$$Var(Y_i) = \pi_i(1-\pi_i) \tag{3.30}$$

π_i is a probability, therefore it is limited by $0 \le \pi_i \le 1$. Thus, using a model for $E(Y_i)$ that restricts its values between 0 and 1 is required. Such a model is the logistic regression model, given by

$$logit(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i'\boldsymbol{\beta} \tag{3.31}$$

The left hand side of Equation 3.31 is referred to as the logit link, denoted by η_i in the GLM. This represents the log of the odds of an event of interest occurring, where $\frac{\pi_i}{1-\pi_i}$ is the odds of the event occurring. So, taking e^{β_j} gives the odds ratio corresponding to a one unit increase in the corresponding explanatory variable, x_{ij}, while all of the 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 provides how much more likely an event of interest is to occur when one explanatory variable changes (Kutner et al., 2005).

It then follows that $E(Y_i)$ is given by

$$\pi_i = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta})} \tag{3.32}$$

This ordinary logistic regression model 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.

3.4 Survey Logistic Regression

Ordinary logistic regression is a widely used approach to model a binary response. However, this approach is only valid for data that comes from a simple random sample. In the case of a complex survey design, ordinary logistic regression may result in overestimation of standard errors, thus leading to incorrect results (Heeringa et al., 2010). This can be avoided by making adjustments to the ordinary logistic regression model in order to account for the survey design. The resulting model is referred to as the survey logistic regression (SLR) model and is commonly used in the analysis of a binary response using data emanating from a complex survey design (Heeringa et al., 2010).

3.4.1 The Model

Consider the survey logistic regression model for a binary response where Y_{hij}, $j = 1, \ldots, n_{hi};\ i = 1, \ldots, n_h;\ h = 1, \ldots, H$ is an observation for the j^{th} individual in the i^{th} 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. Thus, the survey logistic regression model is

$$logit(\pi_{hij}) = \boldsymbol{x}'_{hij}\boldsymbol{\beta} \tag{3.33}$$

with

$$\pi_{hij} = \frac{\exp(\boldsymbol{x}'_{hij}\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}'_{hij}\boldsymbol{\beta})} \tag{3.34}$$

where $\boldsymbol{x}_{hij}$ is the row of the design matrix corresponding to the response of the j^{th} individual in the i^{th} cluster 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 given by

$$f(y_{hij}) = \pi_{hij}^{y_{hij}}\,(1 - \pi_{hij})^{1-y_{hij}} \tag{3.35}$$

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

$$\ell = \ln\ L(\boldsymbol{y}) = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} \ln\ f(y_{hij}) \tag{3.36}$$

The log-likelihood function above does not take the sampling weights into account, so the ML estimates of the model's parameters found using this function are only valid for simple random samples where observations are unweighted (Heeringa et al., 2010). In the case of more complex designs which include sampling weights and clustering, the ML estimates of the parameters and their standard errors are not consistent (Chandra, 2014). Hence, 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, although, for complex survey designs, no convenient likelihood functions are available (Chandra, 2014). Such a likelihood function that incorporates the sampling weights is called 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

The PML method requires knowledge of the distribution of the response variable, similar to the ML method, although it accounts for the sampling weights as follows:

$$P\ell = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{n_{hi}} w_{hij}\ \ln\ f(y_{hij}) \tag{3.37}$$

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

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

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

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

$$\boldsymbol{S}(\boldsymbol{\beta}) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} w_{hij}(y_{hij} - \pi_{hij})\, \boldsymbol{x}'_{hij} = \mathbf{0} \tag{3.38}$$

These estimating equations are nonlinear functions of $\boldsymbol{\beta}$, and therefore require iterative procedures such as Newton-Raphson and Fisher Scoring to be solved. It has been shown that the parameter estimates based on the PML method of estimation are consistent (Heeringa et al., 2010).

The incorporation of the sampling weights in the SLR model also adds to the complexity of obtaining the variance estimates for the estimated parameters. Commonly used methods of variance estimation for the SLR model includes Taylor series approximation, Jackknife repeated replication (JRR) and balanced repeated replication (Heeringa et al., 2010). However. only the Taylor series approximation method will be considered in this thesis.

3.4.3 Taylor Series Approximation

The estimated variances of the PML parameter estimates are no longer simply equal to the inverse of the information matrix as given in Section 3.1.1 for the GLM. This is as a result of weighting and clustering. Thus, in order to obtain these variance estimates, Binder (1983) proposed the Taylor series approximation method, which is based on a linearization technique.

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

$$\boldsymbol{S}(\widehat{\boldsymbol{\beta}}) = \mathbf{0} \tag{3.39}$$

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

is

$$\mathbf{0} = \boldsymbol{S}(\widehat{\boldsymbol{\beta}}) \simeq \boldsymbol{S}(\boldsymbol{\beta}) + \frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \tag{3.40}$$

Therefore,

$$\boldsymbol{S}(\boldsymbol{\beta}) \simeq -\frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \tag{3.41}$$

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

$$Var\left[\boldsymbol{S}(\widehat{\boldsymbol{\beta}})\right] = \left[\frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right] Var(\widehat{\boldsymbol{\beta}}) \left[\frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]' \tag{3.42}$$

or equivalently

$$Var(\widehat{\boldsymbol{\beta}}) = \left[\frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]^{-1} Var\left[\boldsymbol{S}(\widehat{\boldsymbol{\beta}})\right] \left[\frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]^{-1} \tag{3.43}$$

This leads to a sandwich-type variance estimator

$$\widehat{Var}(\widehat{\boldsymbol{\beta}}) = \left[\boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\beta}})\right]^{-1} Var\left[\boldsymbol{S}(\widehat{\boldsymbol{\beta}})\right] \left[\boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\beta}})\right]^{-1} \tag{3.44}$$

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 by making use of stratified and cluster sampling, standard formulae to estimate the variances and covariances of the estimating equations can be used (Heeringa et al., 2010).

Thus, it follows

$$Var\left[\boldsymbol{S}(\widehat{\boldsymbol{\beta}})\right] = \frac{n}{n-p-1} \sum_{h=1}^{H} (1 - f_h) \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (\boldsymbol{s}_{hi.} - \bar{\boldsymbol{s}}_{h..})'(\boldsymbol{s}_{hi.} - \bar{\boldsymbol{s}}_{h..}) \tag{3.45}$$

where

$$\boldsymbol{s}_{hi.} = \sum_{j=1}^{n_{hi}} \boldsymbol{s}_{hij} = \sum_{j=1}^{n_{hi}} w_{hij}(y_{hij} - \widehat{\pi}_{hij})\, \boldsymbol{x}'_{hij} \tag{3.46}$$

and

$$\bar{\boldsymbol{s}}_{h..} = \frac{1}{n_h} \sum_{i=1}^{n_h} \boldsymbol{s}_{hi.} \tag{3.47}$$

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 3.34. For large n, Equation 3.45 reduces to

$$Var\left[\boldsymbol{S}(\widehat{\boldsymbol{\beta}})\right] = \sum_{h=1}^{H}(1 - f_h)\frac{n_h}{n_h - 1}\sum_{i=1}^{n_h}(\boldsymbol{s}_{hi.} - \bar{\boldsymbol{s}}_{h..})'(\boldsymbol{s}_{hi.} - \bar{\boldsymbol{s}}_{h..}) \tag{3.48}$$

The variance estimator in Equation 3.44 is a consistent estimator for the asymptotic variance of $\widehat{\boldsymbol{\beta}}$ (Lipsitz et al., 1994).

3.4.4 Assessing the Model

Goodness-of-Fit

After fitting a model, a test of goodness-of-fit is used to assess whether the model used is the best for fitting the data. This test measures the discrepancy between observed values in the data and expected values based on the fitted model. Common methods for assessing the goodness-of-fit of a fitted GLM is the log-likelihood ratio (deviance) and the Pearson Chi-square statistics. These goodness-of-fit tests are based on independent observations that are identically distributed. In many cases, however, the observations are not i.i.d. Specifically, in the case of a complex survey design, observations that are in the same cluster may be more homogenous than observations from different clusters. A goodness-of-fit test in this situation is an

adaptation to the Hosmer-Lemeshow goodness-of-fit test which accounts for the design of a study, thus making it appropriate in measuring the fit of the survey logistic regression model (Archer & Lemeshow, 2006).

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

$$\widehat{C} = \sum_{k=1}^{10} \frac{(O_k - E_k)^2}{E_k \left(1 - \frac{E_k}{n_k}\right)} \tag{3.49}$$

where

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

This test statistic has a chi-square distribution with 8 degrees of freedom (Hosmer & Lemeshow, 1980). The extension of this Hosmer-Lemeshow goodness-of-fit test is referred to as the F-adjusted mean residual test, also called 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 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} cluster is calculated as follows

$$\widehat{r}_{ij} = y_{ij} - \widehat{\pi}(x_{ij}) \tag{3.50}$$

Applying the grouping strategy, 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

$$\widehat{M}_g = \frac{\sum_i \sum_j w_{ij} \widehat{r}_{ij}}{\sum_i \sum_j w_{ij}} \tag{3.51}$$

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

$$\widehat{W} = \widehat{\boldsymbol{M}}' \left[\widehat{Var}(\widehat{\boldsymbol{M}})\right]^{-1} \widehat{\boldsymbol{M}} \tag{3.52}$$

where $\widehat{Var}(\widehat{\boldsymbol{M}})$ is the variance-covariance matrix of $\widehat{\boldsymbol{M}}$, obtained using variance estimation methods such as the Taylor series approximation (Archer et al., 2007). This test statistic is approximately chi-square distributed with $g - 1 = 9$ degrees of freedom as $g = 10$ in this regard. Although, this chi-square distribution has been shown to not be an appropriate reference distribution. Instead, the F-corrected Wald test statistic has been suggested (Archer & Lemeshow, 2006). This test statistic given by

$$F = \frac{(f - g + 2)}{fg} W \tag{3.53}$$

which 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 less the number of strata and g is the number of categories. This means that,

based on this test statistic, the F-adjusted mean residual test statistic is

$$\widehat{Q}_m = \frac{(f-8)}{10f}\widehat{\boldsymbol{M}}' \left[\widehat{Var}(\widehat{\boldsymbol{M}})\right]^{-1} \widehat{\boldsymbol{M}} \tag{3.54}$$

as $g = 10$ deciles of risk.

Testing Model Parameters

Inferences about the parameters in a SLR model cannot be based on likelihood ratio tests as a pseudo-likelihood function is used for parameter estimation, which is an approximate to the true likelihood (Hosmer et al., 2013). Therefore, 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 given as follows:

$$W = (\boldsymbol{C}\widehat{\boldsymbol{\beta}})' \left[\boldsymbol{C}\,\widehat{Var}(\widehat{\boldsymbol{\beta}})\,\boldsymbol{C}'\right]^{-1} (\boldsymbol{C}\widehat{\boldsymbol{\beta}}) \tag{3.55}$$

where $\widehat{Var}(\widehat{\boldsymbol{\beta}})$ is the estimated variance-covariance matrix for $\widehat{\boldsymbol{\beta}}$ using variance estimation methods. 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 common to approximate this Wald test statistic to an F-distribution using Equation 3.53, where $g = q$.

3.5 Survey Logistic Regression Model Applied to MDHS Data

The statistical analysis for this study was performed using SAS software version 9.4. Specifically, to fit a SLR model, the SAS procedure PROC SURVEYLOGISTIC was used. The sampling weights were adjusted for non-response and to represent the teenage girls included in the data set used in this study. These sampling weights were utilised in fitting the SLR model. The Taylor series approximation method was used for variance estimation of the model. All the independent variables of inter-

est were incorporated in the SLR model to assess their association with the event of teenage pregnancy. All two-way interaction effects were explored in order to control for possible effects of confounding between the factors. Only statistically significant two-way interactions that substantially decreased the model's deviance were included in the final SLR model.

The final SLR model's predictive accuracy can be assessed by making use the Concordance Index (c) which is based on the following calculation:

$$c = [n_c - 0.5(t - n_c - n_d)]t^{-1}$$

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 also 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.

Table 3.1 provides the final SLR model. The following variables had a significant effect on the likelihood of teenage pregnancy at a 5% significance level: region of residence, age, hearing of family planning on the radio, age at first sex, union type, socio-economic status, contraceptive use, education level and tested for HIV. In addition, the interaction between region and education level as well as the interaction between age at first sex and union type had a significant effect on the likelihood of

teenage pregnancy. The final SLR model resulted in a Concordance Index (c) of 0.90, which indicates that the model has a high predictive accuracy.

Table 3.1: Type III analysis of effects for the final SLR model.

Effect	F-Value	P-Value
Region	2.84	0.0442
Age	29.63	<.0001
Type of Residence	0.15	0.7020
Family Planning via Radio	4.39	0.0365
Age at First Sex	5.19	0.0058
Total number of Partners	0.02	0.8931
Union Type	10.40	<.0001
Socio-Economic Status	3.02	0.0494
Contraceptive Use	28.44	<.0001
Head of Household gender	3.33	0.0686
Religion	1.68	0.1695
Education Level	9.11	0.0010
Tested for HIV	158.20	<.0001
Region∗Education Level	48.11	<.0001
Age at First Sex∗Union Type	1.69	0.0150

Table 3.2 presents the odds ratios and their 95% confidence intervals for the variables that were not included in the interaction effects. The significance of the factors was assessed based on the inclusion of 1 in the 95% confidence interval for the odds ratio. No significant difference in the odds of teenage pregnancy was observed for the total number of partners, type of residence and head of household gender. As a girl's age increased by one year, their odds of pregnancy significantly increases by 40.1% (95% CI: 1.240; 1.582). There was a significantly higher likelihood of teenage pregnancy for those who had not recently heard about family planning on the radio (OR = 1.382, 95% CI: 1.020; 1.870) than those who heard about family planning on the radio. In addition, those with a middle socio-economic status had a significantly higher odds of teenage pregnancy (OR = 1.609, 95% CI: 1.101; 2.353) than those with

a rich socio-economic status. There was a significantly lower odds of teenage pregnancy for those that were not on a contraceptive compared to those that were on a contraceptive (OR = 0.445, 95% CI: 0.330; 0.599). Furthermore, the likelihood of teenage pregnancy was significantly higher among the Muslim religion compared to the Christian religion (OR = 1.572; 95% CI: 1.410; 2.374). Those that had not undergone an HIV test prior to the survey were 0.111 times less likely to have experienced pregnancy compared to those who had undergone an HIV test (95% CI: 0.079; 0.157).

Table 3.2: Estimated odds ratios (OR) and corresponding 95% confidence intervals (CI) for the variables not included in interactions for the SLR model

Variables	Odds Ratio (95% CI)
Age	1.401 (1.240; 1.582)*
Type of Residence (Ref=Urban)	
Rural	0.918 (0.591; 1.426)
Family Planning via Radio (Ref=Yes)	
No	1.382 (1.020; 1.870)*
Total number of partners (Ref=4+)	
1 to 3	0.935 (0.352; 2.486)
Socio-economic status (Ref=Rich)	
Middle	1.609 (1.101; 2.353)*
Poor	1.254 (0.902; 1.745)
Contraceptive use (Ref=Yes)	
No	0.445 (0.330; 0.599)*
Head of household gender (Ref=Male)	
Female	1.323 (0.979; 1.789)
Religion (Ref=Christian)	
Muslim	1.572 (1.410; 2.374)*
Catholic	1.201 (0.792; 1.822)
Other	1.053 (0.746; 1.487)
Tested for HIV (Ref=Yes)	
No	0.111 (0.079; 0.157)*

*significant at 5% level of significance

Figures 3.1 and 3.2 present the estimated effects of the interaction between age at first sex and union type, and region and education level, respectively. Among those that were never in a union, the likelihood of teenage pregnancy was highest for those whose sexual debut was between the ages 13 and 15 years old. However, those that were formally in a union and whose sexual debut was between the ages 16 and 19 years had the highest likelihood of teenage pregnancy.

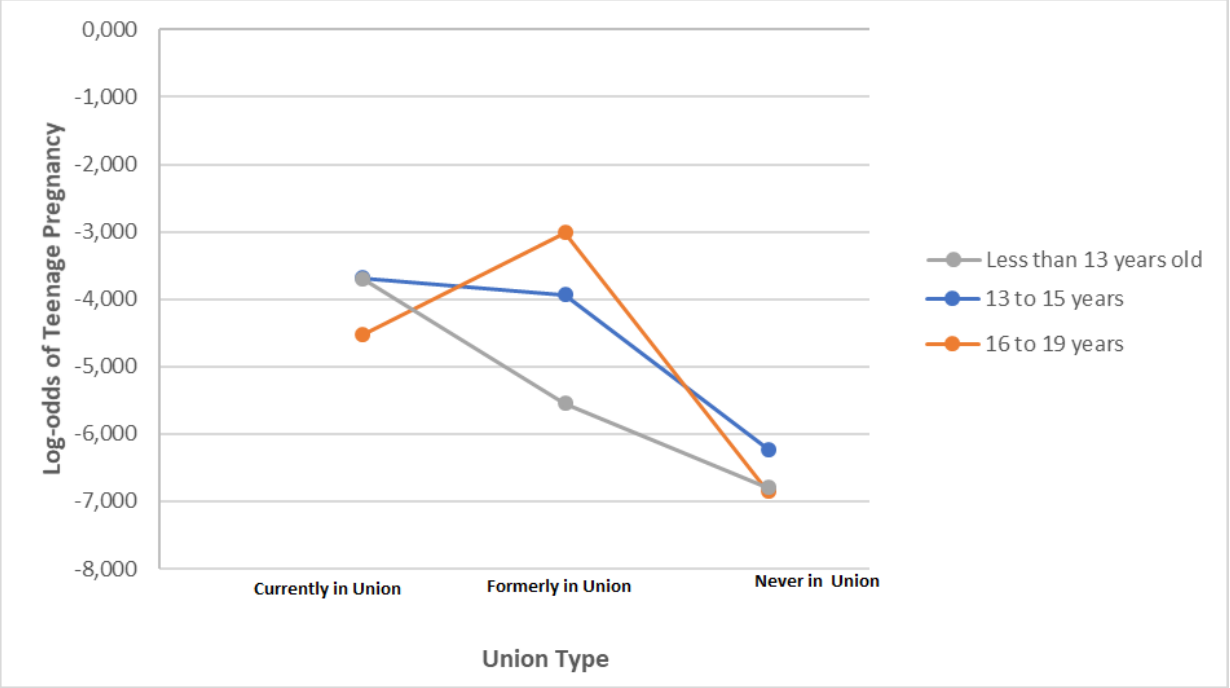

Figure 3.1: The estimated log-odds of teenage pregnancy associated with Age at first sex and Union Type for the SLR model

Considering Figure 3.2, the likelihood of teenage pregnancy was fairly similar across the different regions of Malawi for the different education levels, except for those with no education in the Northern region, where they had a substantially higher likelihood of teenage pregnancy.

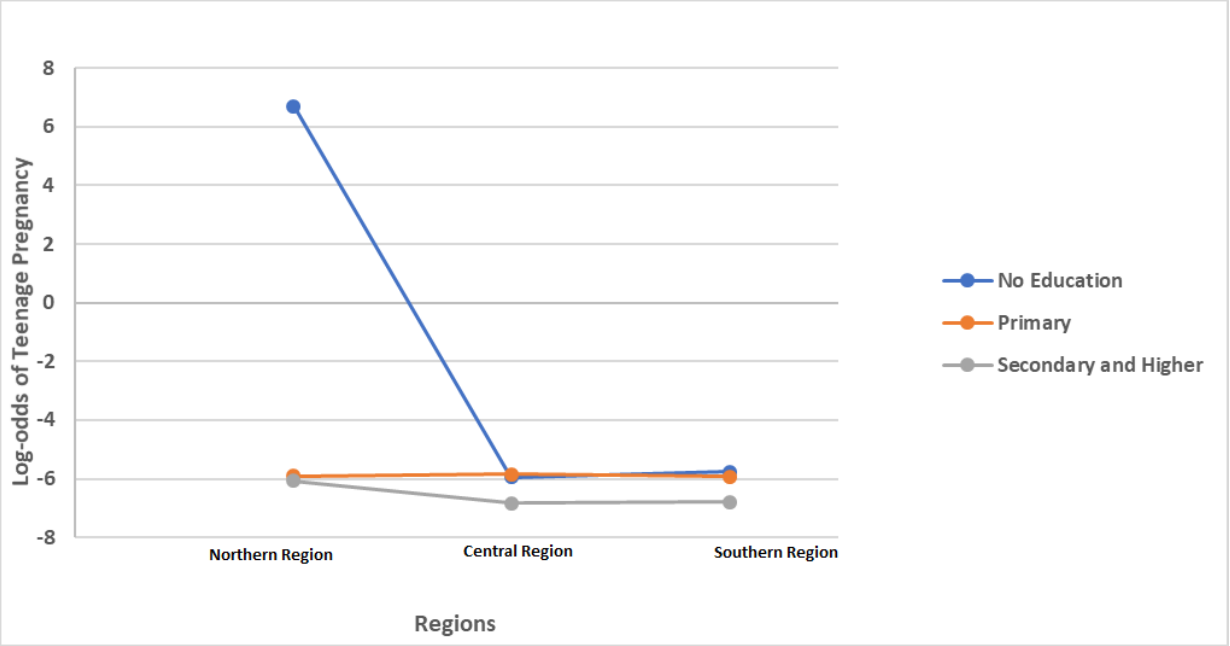

Figure 3.2: The estimated log-odds of teenage pregnancy associated with Region and Education level for the SLR model

3.6 Summary

In this chapter, an overview of the generalised linear model for modelling a non-normal response was presented. The survey logistic regression model, which is an extension to this class of models, was introduced and applied to the MDHS data. Such an approach is considered a design-based approach as it accounts for the complex survey design utilised to obtain the data, where survey weights are used in parameter estimation and inference. However, this model does not account for the effects of clustering where observations may be correlated. Thus, the next chapters considers such an approach.

Chapter 4

Generalised Linear Mixed Models

As previously mentioned in Chapter 3, the GLM assumes that observations are independent. However, cluster sampling may yield correlated observations, and thus the GLM would not be suitable. In addition, the clusters included in the sample represent only a random sample from a population of clusters. Accounting for this clustering effect can be done via the inclusion of a random effect. This leads to the generalised linear mixed model (GLMM) which is an extension of the GLM. The GLMM is a model-based approach where interest is not only on inference regarding the fixed effects, but also on estimating the proportion of variation in the response that is attributable to the multiple levels of sampling (Heeringa et al., 2010). Thus, inference on the variance components of the GLMM may also be of interest.

4.1 The Model

Suppose Y_{ij} is the j^{th} response, $j = 1, \ldots, n_i$, from the i^{th} cluster, $i = 1, \ldots, m$. 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

$$f(y_{ij}|\theta_{ij},\phi) = \exp\left\{\frac{y_{ij}\,\theta_{ij} - b(\theta_{ij})}{\phi} + c(y_{ij},\phi)\right\} \tag{4.1}$$

which follows the same form as Equation 3.1 in Chapter 3, and thus the parameters in the above equation are similar.

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

$$\begin{aligned}\eta_{ij} = g(\mu_{ij}) &= g\left[E(y_{ij}|\boldsymbol{\gamma}_i)\right] \\ &= \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{\gamma}_i \end{aligned} \tag{4.2}$$

or in matrix form

$$g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} \tag{4.3}$$

where $g(.)$ is the known link function that links the conditional mean of $\boldsymbol{y}$ and the linear form of the predictors, $\boldsymbol{X}$ is the $n \times (p+1)$ design matrix for fixed effects, $\boldsymbol{\beta}$ is a $(p+1)\times 1$ vector of regression coefficients for the fixed effects, $\boldsymbol{Z}$ is the $n \times q$ design matrix for the random effects and $\boldsymbol{\gamma}$ is a $q \times 1$ vector of random effect coefficients. It is assumed that $\boldsymbol{\gamma} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{G})$ where $\boldsymbol{G}$ depends on unknown variance components.

A Bayesian approach and a maximum likelihood approach are the two methods of estimation for a GLMM. This thesis will focus on the maximum likelihood (ML) method, which is widely used and has a variety of optimality properties (Searle et al., 2006).

4.2 Maximum Likelihood Estimation

In order to get ML estimates in the generalised linear mixed model, 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

$$f_i(y_{ij}\,|\boldsymbol{\beta}, \boldsymbol{G}, \phi) = \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}\,|\boldsymbol{\gamma}_i, \boldsymbol{\beta}, \phi) f(\boldsymbol{\gamma}_i \,|\, \boldsymbol{G})\, d\boldsymbol{\gamma}_i \tag{4.4}$$

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

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

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{G}, \phi) &= \prod_{i=1}^{m} f_i(y_{ij}\,|\boldsymbol{\beta}, \boldsymbol{G}, \phi) \\ &= \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} f_{ij}(y_{ij}\,|\boldsymbol{\gamma}_i, \boldsymbol{\beta}, \phi) f(\boldsymbol{\gamma}_i \,|\, \boldsymbol{G})\, d\boldsymbol{\gamma}_i \end{aligned} \tag{4.5}$$

Estimation of the fixed effects in a GLMM using the method of maximum likelihood is identical to that for a linear mixed model in the case of a normal response. However, for non-normal responses, the likelihood function generally 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, approximation methods are required to evaluate the likelihood function given in Equation 4.5. The various methods involve approximation of the integrand, approximation of the integral itself or approximation of the data (Hedeker, 2005). This thesis will focus on an approach to approximate the integrand, such as Laplace Approximation, which allows for model comparisons using information criteria such as Akaike's Information Criteria (AIC). Furthermore, Laplace approximation is computationally less demanding compared to other approximation methods.

4.3 Laplace Approximation

Laplace approximation is a common method of approximation of the integrand, which is used when the exact likelihood function is difficult to evaluate (Jiang, 2007). Suppose the integral in the following form is to be approximated

$$\int e^{Q(\boldsymbol{x})} dx \tag{4.6}$$

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

$$Q(\boldsymbol{x}) \approx Q(\widehat{\boldsymbol{x}}) + \frac{1}{2}(\boldsymbol{x} - \widehat{\boldsymbol{x}})' Q''(\widehat{\boldsymbol{x}})(\boldsymbol{x} - \widehat{\boldsymbol{x}}) \tag{4.7}$$

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

This yields the following approximation to Equation 4.6:

$$\int e^{Q(\boldsymbol{x})} dx \approx (2\pi)^{\frac{q}{2}} |Q''(\widehat{\boldsymbol{x}})|^{-\frac{1}{2}} e^{-Q'(\widehat{\boldsymbol{x}})} \tag{4.8}$$

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 4.5 is proportional to the integral in Equation 4.6, where the function Q is given by

$$Q(\boldsymbol{\gamma}) = \phi^{-1} \sum_{j=1}^{n_i} \left[y_{ij}(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{\gamma}) - b(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{\gamma}) \right] - \frac{1}{2}\boldsymbol{\gamma}'\boldsymbol{G}\,\boldsymbol{\gamma} \tag{4.9}$$

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.

4.4 Model Selection

The likelihood ratio test and the Wald test (F-test) described in Chapter 3 can be used for inference concerning the fixed effect parameter estimates obtained through numerical approximations, such as Laplace approximation. The likelihood ratio test can be applied to the data in comparing two nested models that have different mean structures, however consisting of the same variance-covariance structure. In a similar manner, the likelihood ratio test can be used in comparing nested models that have unique covariance structures, however consisting of the same mean, and inferences of variance-covariance components remain valid for the Wald test approximation. However, if the variance parameter that is being tested takes values on the boundary of the parameter space, the normal approximation fails. This means that the test statistics for these tests will not have the traditional Chi-square distribution under the null hypothesis (Zhang & Lin, 2008). However, when testing the null hypothesis of no random effects, the test statistic will be a mixture of Chi-square distributions, rather than the classical single Chi-square distribution (Zhang & Lin, 2008).

4.5 Generalised Linear Mixed Model Applied to MDHS Data

The GLMM was fitted to the MDHS data using the SAS procedure PROC GLIMMIX with a Laplace approximation method and a logit link function. In addition, the RANDOM statement was used to account for the effect of clustering in the data. Specifically, a random intercept that varied at cluster level was included in the model. The need for this random intercept was assessed using the COVTEST in SAS, which tests if the corresponding covariance parameter should be equal to zero. Table 4.1 below illustrates the result of this test. The null hypothesis of the covariance parameter equal to zero was rejected, hence it was necessary to account for the random cluster effect in the model.

Table 4.1: Test of covariance parameters based on the likelihood.

Label	DF	-2Log Likelihood	χ^2	P-value
No G - side effects	1	2059.61	29.17	<0.0001

Prior to obtaining the final GLMM, the model was fitted with different covariance structures for $\boldsymbol{G}$ in order to find the most suitable structure. Four common covariance structures were considered: Variance Components (VC), Unstructured (UN), Compound Symmetry (CS) and AR(1) (SAS & Guide, 1999) The VC Structure is where the correlation of errors within a subject are presumed to be zero. While CS is a covariance structure that includes within-subject correlated errors. For CS, errors are correlated between time points within the subjects and the correlations are presumed to be identical for each set of time, regardless of how distant in the time the repeated measures are made. AR(1) considers correlation to be highest for adjacent times, and systematically decreases correlation with increasing distance between time points. The UN structure estimates unique correlations for each pair of time points (PennState, 2020). The differences in these covariance structures is that CS means that all variances are equal to each other and all covariances are equal to each other, however for VS, each variance is different and covariances are equal to zero. For UN, each variance and each covariance is different and do not relate to each other (Karen Grace-Martin, 2011). In this analysis, VC yielded the lowest AIC value, thus it was selected. The final GLMM was fitted using the same variables as those in the final SLR model. The variance component for the cluster effect was estimated at 0.6534 with a standard error of 0.1808. This estimate is relatively far from zero, thus again confirming the need for this random effect in the model. It should be noted that the SAS GLIMMIX procedure can include a weight statement so that the parameter estimates are weighted, just as in the case of the SLR model. However, upon attempting to incorporate the sampling weights in the GLMM, the model was highly overdispersed. Thus, the final results are based on the unweighted GLMM, which did not suffer from residual overdispersion.

Table 4.2 presents the final GLMM. The results of the GLMM largely agreed with that of the SLR model, where the variables age, hearing of family planning on the radio, union type, socio-economic status, contraceptive use, education level and tested for HIV were significantly associated with the likelihood of teenage pregnancy at a 5% level of significance. However, the GLMM indicated that the gender of the head of household was significantly associated with teenage pregnancy at a 5% level of significance when the SLR indicated this factor was only significant at 10%. Unlike the SLR model, the GLMM did not indicate a significant association between age at first sex and the likelihood of teenage pregnancy, as well as the region of residence and the likelihood of pregnancy. Furthermore, the interaction between region and education level as well as the interaction between age at first sex and union type were no longer significant at a 5% level of significance.

Table 4.2: Analysis of effects for the final GLMM

Effect	**F-Value**	**P-Value**
Region	0.15	0.8625
Age	37.28	<0.001
Type of Residence	0.30	0.5816
Family Planning via Radio	5.85	0.0157
Age at First Sex	1.57	0.2077
Total number of Partners	0.00	0.9467
Union Type	60.69	< 0.001
Socio-Economic Status	3.93	0.0198
Contraceptive Use	28.73	<0.001
Head of Household gender	4.73	0.0298
Religion	2.86	0.0358
Education Level	7.20	0.0008
Tested for HIV	229.81	<0.001
Region$*$Education Level	1.16	0.3261
Age at First Sex$*$Union Type	2.20	0.0673

Table 4.3 presents the estimated odds ratios (OR) and their 95% confidence intervals (CI) for the variables that were not included in the interaction effects. Similar to the SLR model, the GLMM indicated that there was an increased likelihood of teenage pregnancy with an increase in age (OR = 1.436, 95% CI: 1.278; 1.613). In addition, there was a significantly higher likelihood of teenage pregnancy among those who had not heard about family planning on the radio (OR = 1.391, 95% CI: 1.064; 1.819), those with a middle socio-economic status compared to a rich socio-economic status (OR = 1.687, 95% CI: 1.170; 2.434), those residing in households headed by females (OR = 1.354, 95% CI: 1.030; 1.780), and those in the Muslim religion compared to the Christian religion (OR = 1.793, 95% CI: 1.170; 2.748). In addition, there was a significantly lower likelihood of teenage pregnancy for those who were no on a contraceptive (OR = 0.447, 95% CI: 0.333; 0.600) and well as those who had not been for an HIV test prior to the survey (OR = 0.083, 95% CI: 0.060; 0.114). No significant difference in the odds of teenage pregnancy was seen based on the type of place of residence and total number of partners.

Figure 4.1 presents the estimated log-odds of teenage pregnancy for the interaction between age at first sex and union type based on the fitted GLMM. Similarly, Figure 4.2 presents the results of the interaction between the region of residence in Malawi and education level. Both of these interaction effects display similar patterns to those in the SLR model. The likelihood of teenage pregnancy was highest for those whose sexual debut was between the ages 13 and 15 years old and were never in union (Figure 4.1). Those with no education in the Northern region of Malawi had a substantially higher likelihood of teenage pregnancy (Figure 4.2).

Table 4.3: Estimated odds ratios (OR) and corresponding 95% confidence intervals (CI) for the variables not included in interactions for the GLMM

Variables	Odds Ratio (95% CI)
Age	1.436 (1.278; 1.613)*
Type of Residence (Ref=Urban)	
Rural	0.889 (0.585; 1.351)
Family Planning via Radio (Ref=Yes)	
No	1.391 (1.064; 1.819)*
Total number of partners (Ref=4+)	
1 to 3	0.978 (0.515; 1.858)
Socio-economic status (Ref=Rich)	
Middle	1.687 (1.170; 2.434)*
Poor	1.259 (0.900; 1.760)
Contraceptives (Ref=Yes)	
No	0.447 (0.333; 0.600)*
Head of household gender (Ref=Male)	
Female	1.354 (1.030; 1.780)*
Religion (Ref=Christian)	
Muslim	1.793 (1.170; 2.748)*
Catholic	1.346 (0.943; 1.922)
Other	1.080 (0.789; 1.479)
Tested for HIV (Ref=Yes)	
No	0.083 (0.060; 0.114)*

*significant at 5% level of significance

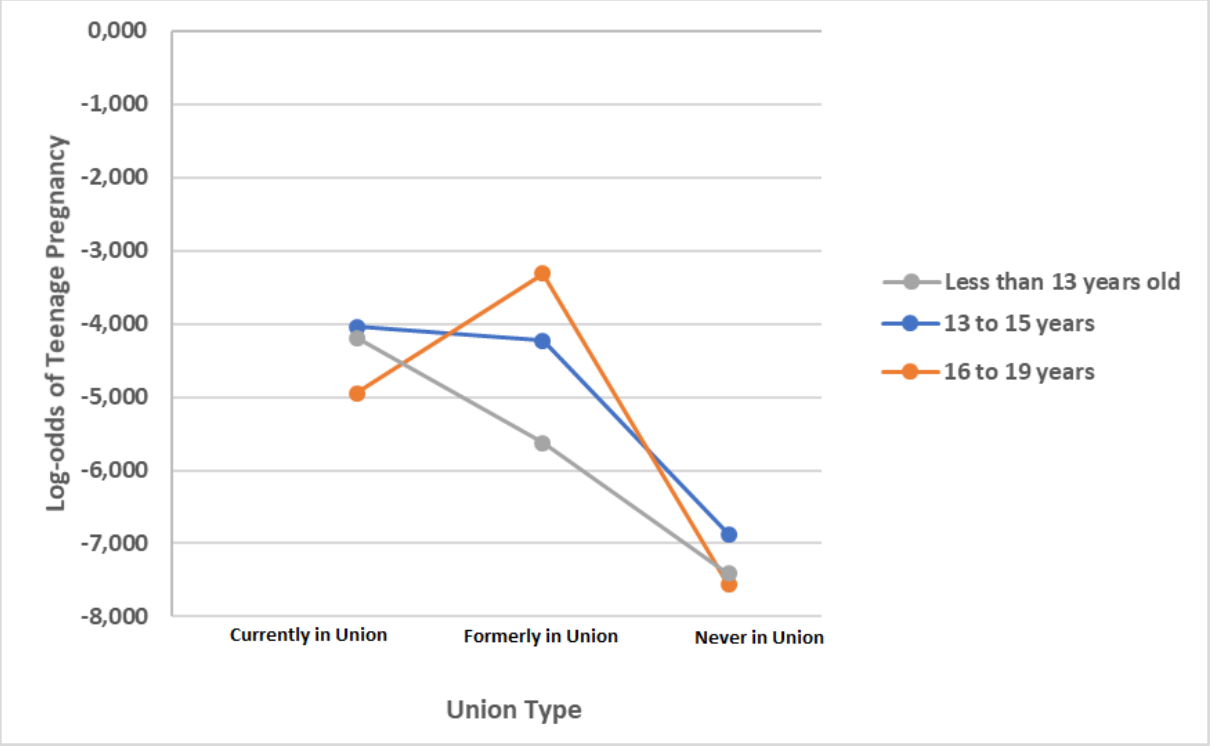

Figure 4.1: The estimated log-odds of teenage pregnancy associated with Age at first sex and Union Type for the GLMM model

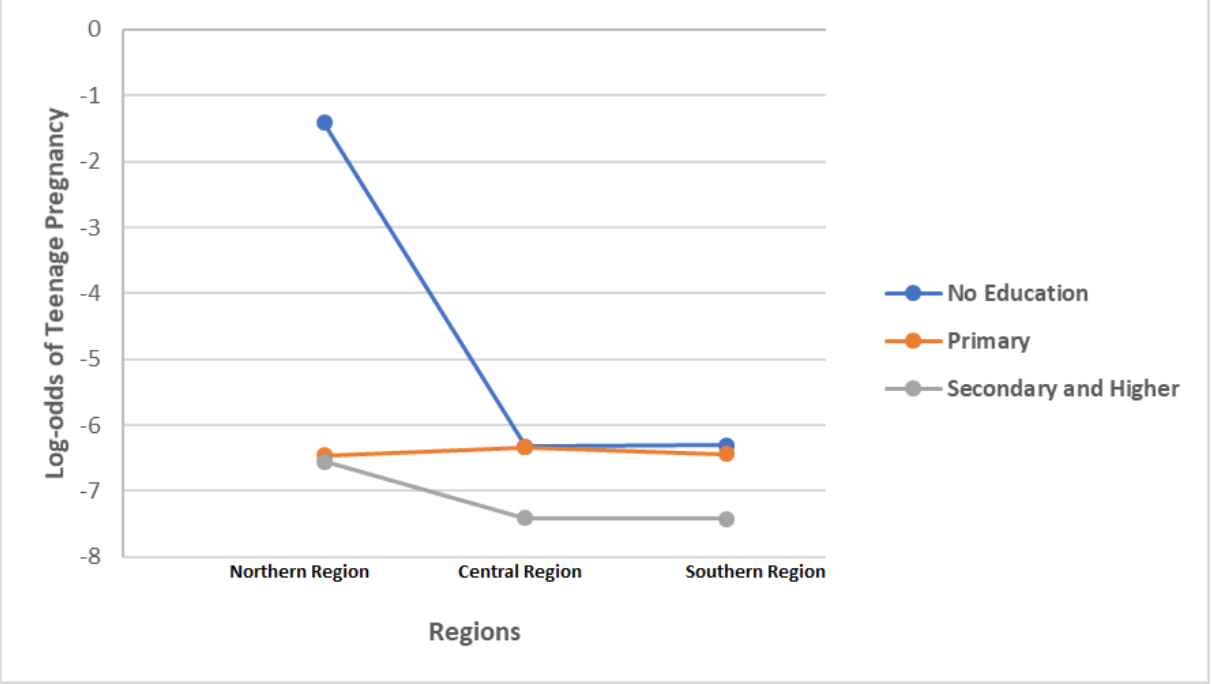

Figure 4.2: The estimated log-odds of teenage pregnancy associated with Regions and Education Level for the GLMM model

4.6 Summary

This chapter gave a brief overview of the GLMM and presented the results of its application to the MDHS data. A generalised linear mixed model extends the generalised linear model by the inclusion of a random effect. A random effect aids in accounting for possible correlations in the data that may be present due to multiple stages of sampling. In this thesis, the random effect was incorporated into the model to account for the effect of clustering. The test of covariance parameters indicated that the clustering effect was significant and necessary to account for. The results of the GLMM fitted to the MDHS data revealed similar factors associated with teenage pregnancy as what the SLR model revealed in the previous chapter. However, both the GLMM and SLR model do not account for spatial variation in the observations. The next chapter discusses the concept of spatial variation and describes the appropriate technique of accounting for it.

Chapter 5

Accounting for Spatial Variation

5.1 Introduction

Spatial variation occurs when the observations are separated or dispersed over space, where spatial dependence can be explained as objects that are close together in a space being more similar than objects which are further apart (Lloyd, 2010). In the case where data values are not spatially dependent, many forms of spatial analysis are pointless. The term positive spatial autocorrelation refers to the correlation of observations with itself and where neighbouring values tend to be similar. This means that the data are spatially dependent. In the case where neighbouring values tend to be dissimilar, this is termed negative spatial autocorrelation (Lloyd, 2010).

There are two classifications of spatially dependent data: isotropic or anistropic. Isotropy is defined as a property of either a process or data, where autocorrelation changes only with distance between two locations. Anistropy is defined as the property of either a process or data, where spatial autocorrelation depends on both the direction and the distance between two locations (Reade et al., 2016). Spatial data can be further subdivided into geostatistical data, lattice data, areal data and point patterns. Geostatistical data are collected over a continuous space, often from randomly selected sites, and using this data the aim is to predict the value of the property at

other, unsampled, locations. Lattice data pertain to equally spaced locations. Areal data involves aggregated quantities for each areal unit within some relevant spatial partition of a given region, where the neighbourhood structure of the regions is used in assessing or accounting for spatial variation. Lastly, point patterns involves the locations of events/responses of interest, and concern is usually to analyse the spatial configuration of the data/responses, rather than the values attached to them (Lloyd, 2010).

There are two processes that can be followed in dealing with spatially dependent data; one can characterize the spatial covariance parameters and describe the nature of spatial correlation, and then one can further adjust for the presence of spatial variation when modelling some event/response of interest (Reade et al., 2016). In this chapter, we consider methods for both processes and discuss their application to the MDHS data in modelling the likelihood of teenage pregnancy. For this purpose, we make use of the geographical coordinates of the clusters in the MDHS data and we assume that the data is isotropic.

5.2 Measures of Spatial Autocorrelation

In Chapter 4, the generalised linear mixed model was considered in order to account for a cluster effect where possible correlations may exist within the clusters. However, an implicit assumption behind this model is that the residuals do not vary as a function of space. Strongly correlated data reduces the statistical power of inference making a model untrustworthy. This assumption may be checked using various methods discussed below. There are two types of measures that can be used to assess spatial autocorrelation: Moran's I and Geary's C.

Moran's I

This was the first measure of spatial autocorrelation introduced by Moran (1950) to study stochastic phenomena which are distributed in space in two or more dimensions. Moran's I, also known as Moran's I Index, simultaneously measures spatial autocorrelation based on both feature locations and feature values. Given a set of features and an associated attribute, it evaluates whether the pattern expressed is clustered, dispersed, or random. Moran's I statistic is based on cross-products of the deviations from the mean and is calculated as follows for n observations for variable x at locations i and j:

$$I = \frac{n \sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{W \sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{5.1}$$

where w_{ij} is a spatial weight between location i and j, $W = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}$ and $\bar{x}$ is the mean of variable x (Cliff & Ord, 1972). The weight reflects the spatial influence of i on j, where options for this weight includes:

- Binary:

$$w_{ij} = \begin{cases} 1 & \text{if sites } i \text{ \& } j \text{ pertains to distance class } h, \\ 0 & \text{if } i \text{ \& } j \text{ are in different distance class } h, \\ 0 & \text{if } i = j \end{cases}$$

- Power distance weight:

$$w_{ij} = 1/d_{ij}^{\alpha} \text{ typically } \alpha = 1 \text{ or } 2$$

  where d_{ij} is the distance between spatial locations i and j and α is a pre-specified parameter that determines the centre of the spatial location and defines the degree to which the two locations are deemed close.

- Exponential distance weight:

$$w_{ij} = \exp\left(-\alpha d_{ij}\right)$$

- General for non-binary:

$$w_{ij} = scale/(1 + h^{power}), \;\; scale, \, power > 0$$

- Measure on lattices:

$$w_{ij} = \begin{cases} 1 & \text{if sites } i \text{ \& } j \text{ are connected,} \\ 0 & \text{if sites } i \text{ \& } j \text{ are not connected.} \end{cases}$$

- Distance Weight:

  If the spatial weights are set to zero beyond a radius *d* and decrease monotonically to zero with increasing distance, d_{ij}, then the distance weighting function is given by

$$w_i(j) = \begin{cases} \left(1 - \dfrac{d_{ij}^2}{d^2}\right)^2 & \text{if} \quad d_{ij} \leq d, \\ 0 & \text{if} \quad d_{ij} > d \end{cases}$$

The null hypothesis states that feature values are randomly distributed across the study area. A Z-score and p-value are calculated to indicate whether to reject the null hypothesis or not. A Moran's I value near +1.0 indicates clustering while a value near -1.0 indicates a dispersed pattern (Cliff & Ord, 1972).

Geary's C

Geary's C statistic is another measure of spatial autocorrelation, originally proposed by Geary (1954). It is based on the deviations in responses of each observation with one another:

$$C = \frac{n-1}{2W}\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - x_j)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \tag{5.2}$$

This statistics ranges from 0 (maximal positive autocorrelation) to a positive value for high negative autocorrelation. Its expectation is 1 in the absence of autocorrelation, regardless of the specified weight matrix (Cliff & Ord, 1972).

Moran's I is a more global measurement it has been shown to be more consistent and powerful than Geary's C. However, Geary's C is more sensitive to differences in smaller neighbourhoods (Leung et al., 2000). These statistics can be used to assess spatial autocorrelation in a fitted GLMM's residuals. This then gives an indication that there is spatial variability in the data that needs to be taken into consideration in the statistical model. If significant spatial autocorrelation is detected, then the GLMM can be extended to the spatial generalised linear mixed model (spatial GLMM) in order to account for such spatial autocorrelation. This model is discussed in the next section.

5.3 Spatial Generalised Linear Mixed Models

The spatial GLMM takes on the same form as the non-spatial GLMM presented in Chapter 4:

$$g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} \tag{5.3}$$

As similarly defined in Equation 4.3 for the GLMM, $g(.)$ is the known link function that links the conditional mean of $\boldsymbol{y}$ and the linear form of the predictors. However, the response vector is now spatially indexed such that $\boldsymbol{y} = [y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n)]$, for spatial locations $\boldsymbol{s}_i$, $i = 1, \ldots, n$, and it is further assumed that $\boldsymbol{s} \in D$, if D is some d-dimensional Euclidean space. $\boldsymbol{X}$ is the $n \times (p+1)$ design matrix for fixed effects, $\boldsymbol{\beta}$ is a $(p+1) \times 1$ vector of regression coefficients for the fixed effects, $\boldsymbol{Z}$ is the $n \times q$ design matrix for the random effects and $\boldsymbol{\gamma}$ is a $q \times 1$ vector of random effect coefficients. However, now for the spatial GLMM, $\boldsymbol{\gamma} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{G})$ where $\boldsymbol{G}$ is a $q \times q$ spatial

covariance matrix, with q being the number of spatial locations.

Finding an appropriate covariance function to account for spatial autocorrelation is an important procedure in fitting a spatial GLMM. In general, it is assumed that spatial autocorrelation does not depend on the location of responses $y(\boldsymbol{s}_i)$ and $y(\boldsymbol{s}_j)$, but rather the distance between them (Cressie, 1991). Thus, the covariance between two responses at different sites is given by $cov(y(\boldsymbol{s}_i), y(\boldsymbol{s}_j)) = c(.)$, where $c(.)$ is a function of the distance, h, between the locations. In addition, there are two underlying assumptions regarding this covariance: (i) there is a constant mean: $E[y(\boldsymbol{s}_i)] = E[y(\boldsymbol{s}_j)]$, and (ii) it is second order stationary (Cressie, 1991). There are various spatial covariance structures that exist. The most common structures include the exponential, Gaussian, Matérn, power and spherical structures.

Spatial dependency may be described by a range of functions. One such function is the semivariogram, also simply referred to as the variogram. In geostatistics, the variogram is a useful function used to fit a model for the spatial correlation in the data. It is considered essential to first examine the empirical semivariogram before fitting a spatial covariance structure in a spatial GLMM (Reade et al., 2016). The semivariogram measures spatial variability as a function of distance between two locations. Specifically, it is defined as one-half the variance of the difference between two observations, made at different locations (Littell et al., 2006).

The empirical semivariogram gives us a visual depiction of the actual spatial variability of the data. The basic elements of a semivariogram are the nugget, sill and range, as illustrated in Figure 5.1. The nugget can be defined as the intercept of the semivariogram where $d = 0$. In addition, the nugget effect illustrates errors that are spatially independent or, the variance occurring at a particular location. The sill is the value that the semivariogram tends towards for large values of d. At large distances, variables cease to be correlated (Littell et al., 2006). Therefore, it is inferred

that when d is very large, the sill corresponds to the variance of an observation (Littell et al., 2006). From Figure 5.1, one can also see that the range can be defined as the value of d at which the semivariogram reaches the sill. This then informs us that for all distances less than this value of d, observations are spatially correlated. While, at distances greater than or equal to this value of d, observations are no longer spatially correlated (Littell et al., 2006).

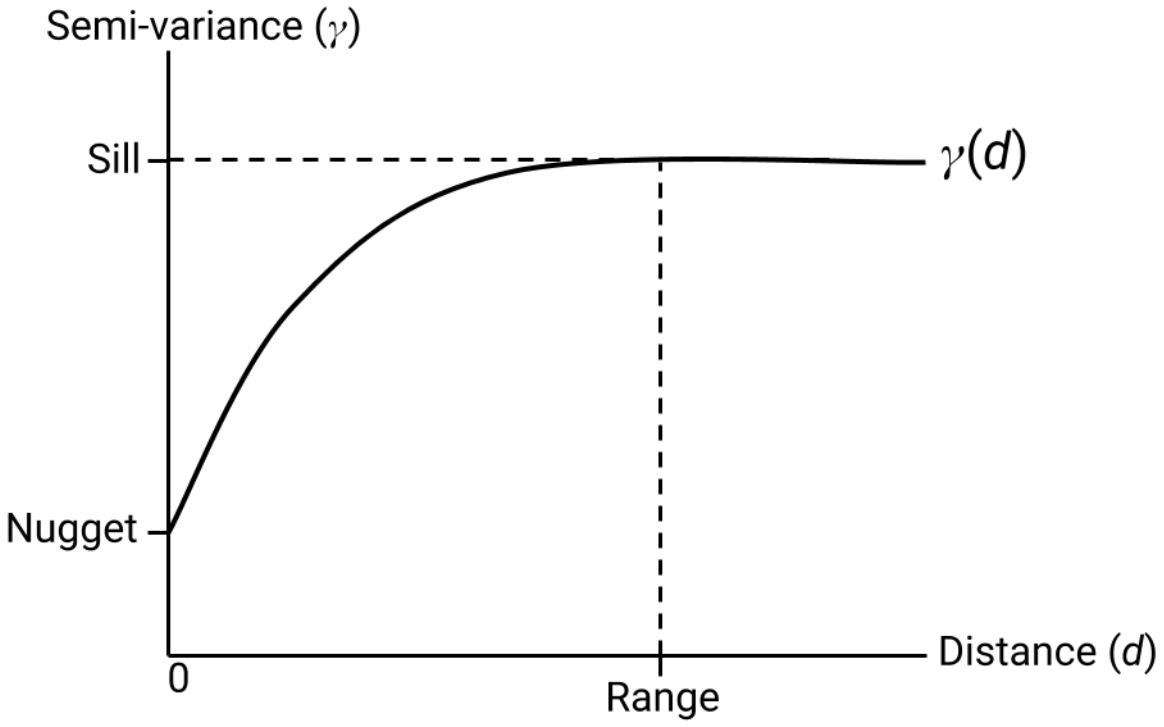

Figure 5.1: The nugget, sill and range parameters illustrated on a idealized variogram function (Google Earth Engine, 2020).

5.4 Examining Residual Autocorrelation in the MDHS data

Prior to fitting the spatial GLMM, it must first be established whether there is autocorrelation present in the MDHS data by detecting the presence of spatial auto-

correlation in the residuals. In addition to the Moran's I and Geary's C statistics, the variogram procedure also aids in establishing whether there is autocorrelation found in the data by detecting the presence of spatial autocorrelation in the residuals. For the purpose of this, the residuals were extracted from the fitted GLMM presented in Section 4.5. The average residual was then calculated for each cluster, as the geographical coordinates (latitude and longitude) were provided at cluster level for the MDHS data. This process of constructing and examining the structure of the empirical semi-variogram is then followed based on these average residuals.

In doing this process, one needs to first group the spatial locations into intervals in-line with a common distance between them. In SAS, the **PROC VARIOGRAM** procedure is used, which produces the empirical semivariogram and determines the distance between each spatial location using the uniqueness of the geographical coordinates for a particular cluster. The procedure also requires that one specifies the size of the lag class and maximum number lags via **LAGDISTANCE (lagd)** and **Maxlag** commands, respectively in SAS.

In order to make these specifications, the pairwise distribution of the data using a variety of class numbers was examined. Specifying the appropriate number of intervals is based on the researcher's discretion. However, Journal & Huijbregts (1978) have recommended that the lag classes be specified such that each class contains a minimum of 30 location pairs and only lags up to approximately half of the extreme distance between points be considered. To start off with, we chose to group our locations across 50 classes. Based on the resulting pairwise distribution, the following information was obtained and presented in Table 5.1.

Table 5.1: Pairwise Information for 50 classes

Number of lags	51
Lag Distance	0.16
Maximum Data in Latitude	7.63
Maximum Data in Longtitude	3.00
Maximum Data Distance	8.20

This information in Table 5.1 was used to construct the empirical semivariogram. Using the common lag distance of 0.16 and the maximum data distance of 8.20, it was determined that the maximum number of classes should be 26 ($max\ class = (8.20 \div 2) \div 0.16 \approx 26$). Figure 5.2 displays the resulting empirical semivariogram.

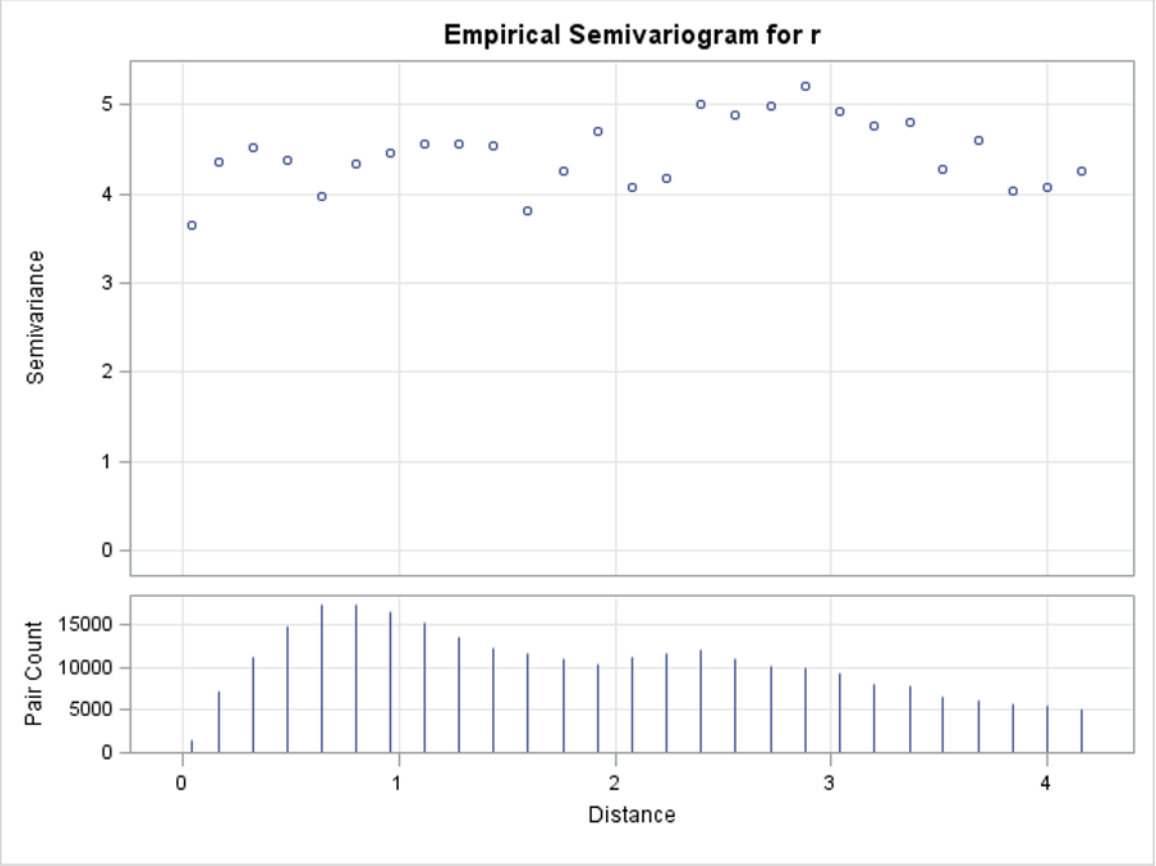

Figure 5.2: Empirical semivariogram for the MDHS data

Figure 5.2 indicates that as the distance between clusters increases, the semivariogram fluctuates. This may be an indication that spatial autocorrelation is present.

However, to confirm this, we make use of the Moran' I and Geary's C statistics. Table 5.2 below presents the results of these two autocorrelation tests. While Moran's I indicates that there is no significant spatial autocorrelation (P-value = 0.7782), Geary's C indicates otherwise (P-value = 0.0028). This suggests that there is spatial autocorrelation on a more local scale compared to a global scale. In addition, the positive statistic for Geary's C indicates that there is possibly a clustering of high residual values with reference to their average.

Table 5.2: Autocorrelation test results

Assumption	Coefficient	Observed	Expected	Std. Dev.	Z	P-value
Normality	Moran's I	-0.00539	-0.00127	0.0146	-0.282	0.7782
Normality	Geary's C	0.89276	1	0.0359	-2.987	0.0028

Now that we have observed the presence of residual spatial autocorrelation, it is necessary to follow a modelling procedure that will allow one to account for this spatial autocorrelation. This will be done by fitting a spatial GLMM to the MDHS data. However, prior to fitting this model, we need to determine a spatial covariance structure that best suits the data. We do this by specifying a range of variogram models in the **PROC VARIOGRAM** procedure and examine the AIC, making use of the *small-is-better* criterion. Table 5.3 below represents the results of this process along with the corresponding AICs. Based on this, the spherical variogram is the most appropriate structure and thus will be selected as the spatial covariance structure for the cluster random effect in the spatial GLMM, the results of which are discussed in the next section.

Table 5.3: Fit of the spatial covariance structure for the variogram

Spatial Model	**Spherical**	Exponential	Power	Matérn	Gaussian
AIC	**32.73140**	32.73143	32.73485	34.73376	32.73289

5.5 Spatial Generalised Linear Mixed Model Applied to MDHS Data

To fit the spatial GLMM, a similar procedure that was used to fit the GLMM in Section 4.5 was followed, where the **PROC GLIMMIX** procedure in SAS was used. Once again, a cluster-level random effect was incorporated into the model as a random intercept. However, a spherical spatial covariance structure was specified for $\boldsymbol{G}$. This accounted for not only possible correlations between the observations within a cluster, but also spatial autocorrelation between the clusters. Based on the **COVTEST** procedure, the null hypothesis of the covariance parameters equal to zero was rejected (Table 5.4). This indicates that the cluster random effect was significant in the model. The ratio of the Pearson Chi-square statistic to its degrees of freedom was 0.86, this illustrates that the variability in the data is well modelled and that there were no consequences of residuals overdispersion.

Table 5.4: Test of covariance parameters based on the likelihood.

Label	DF	-2Log Likelihood	χ^2	P-value
No G - side effects	2	2059.61	29.69	<0.0001

The estimated variance component for the cluster effect was 0.6661, which represents the partial *sill*. The estimated range, based on the spherical covariance structure, was 22.0000. This implies that observations than 22 units apart are not spatially correlated.

The final spatial GLMM is presented in Table 5.5, where the fixed effects including the two-way interactions explored as per the SLR model and GLMM were examined. These results largely concurred with that of the non-spatial GLMM, where the variables age, hearing of family planning on the radio, union type, socio-economic status, contraceptive use, the gender of the head of household, education level and tested for HIV were significantly associated with the likelihood of teenage preg-

nancy at a 5% level of significance. The effect of accounting for spatial autocorrelation resulted in slightly lower P-values for some of the effects, however this did not have any consequences in the significance of the effects when compared to the results of the non-spatial GLMM.

Table 5.5: Analysis of effects for the final spatial GLMM

Effect	**F-Value**	**P-Value**
Region	0.15	0.8584
Family Planning via Radio	5.84	0.0157
Age	37.18	< 0.001
Type of Residence	0.31	0.5796
Age at First Sex	1.57	0.2075
Total number of Partners	0.00	0.9537
Union Type	60.79	<0.001
Socio-Economic Status	3.95	0.0195
Contraceptive Use	28.66	<0.001
Head of Household gender	4.68	0.0307
Religion	2.87	0.0354
Education Level	7.17	0.0008
Tested for HIV	229.48	<0.001
Region$*$Education Level	1.16	0.3249
Age at First Sex$*$Union Type	2.21	0.0659

Table 5.6 presents the estimated odds ratios (OR) and their 95% confidence intervals (CI) for the variables that were not included in the interaction effects in the spatial GLMM. As expected after considering the results of the analysis of fixed effects in Table 5.5 above, the results of the odds ratios and confidence intervals almost mimic that of the non-spatial GLMM. In addition, the estimated log-odds of teenage pregnancy based on the two interaction effects given in Figures 5.3 and 5.4 demonstrate the same findings as those from the non-spatial GLMM. Therefore, it is clear that accounting for spatial autocorrelation in the GLMM did not greatly alter the main findings of the non-spatial GLMM. It is necessary to consider the principle of par-

simony, this advocates choosing the simplest scientific explanation that fit the evidence. For this reason, the results of the fixed effects for the spatial GLMM will not be further discussed, as the same conclusions can be made to that of the non-spatial GLMM.

Table 5.6: Estimated odds ratios (OR) and corresponding 95% confidence intervals (CI) for the variables not included in interactions for the spatial GLMM

Variables	Odds Ratio (95% CI)
Age	1.436 (1.278; 1.613)*
Type of Residence (Ref=Urban)	
Rural	0.888 (0.584; 1.352)
Family Planning via Radio (Ref=Yes)	
No	1.392 (1.064; 1.820)*
Total number of partners (Ref=4+)	
1 to 3	0.981 (0.517; 1.863)
Socio-economic status (Ref=Rich)	
Middle	1.690 (1.171; 2.439)*
Poor	1.260 (0.901; 1.763)
Contraceptives (Ref=Yes)	
No	0.447 (0.333; 0.600)*
Head of household gender (Ref=Male)	
Female	1.352 (1.028; 1.778)*
Religion (Ref=Christian)	
Muslim	1.795 (1.171; 2.753)*
Catholic	1.350 (0.945; 1.928)
Other	1.080 (0.789; 1.479)
Tested for HIV (Ref=Yes)	
No	0.082 (0.059; 0.113) *

*significant at 5% level of significance

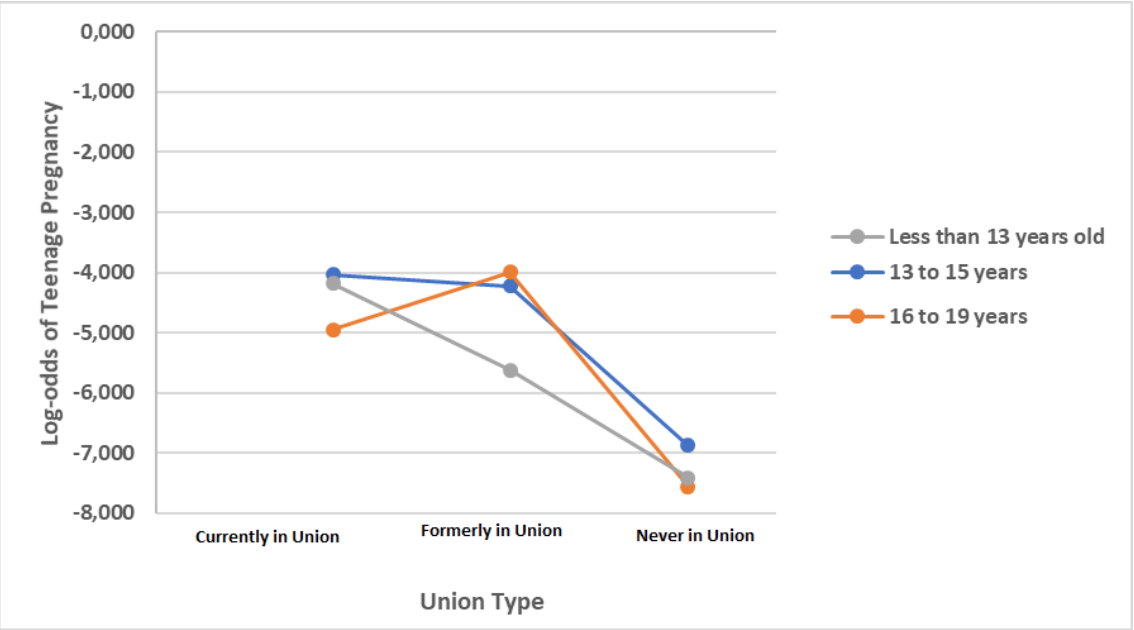

Figure 5.3: The estimated log-odds of teenage pregnancy associated with Age at first sex and Union type for the spatial GLMM model

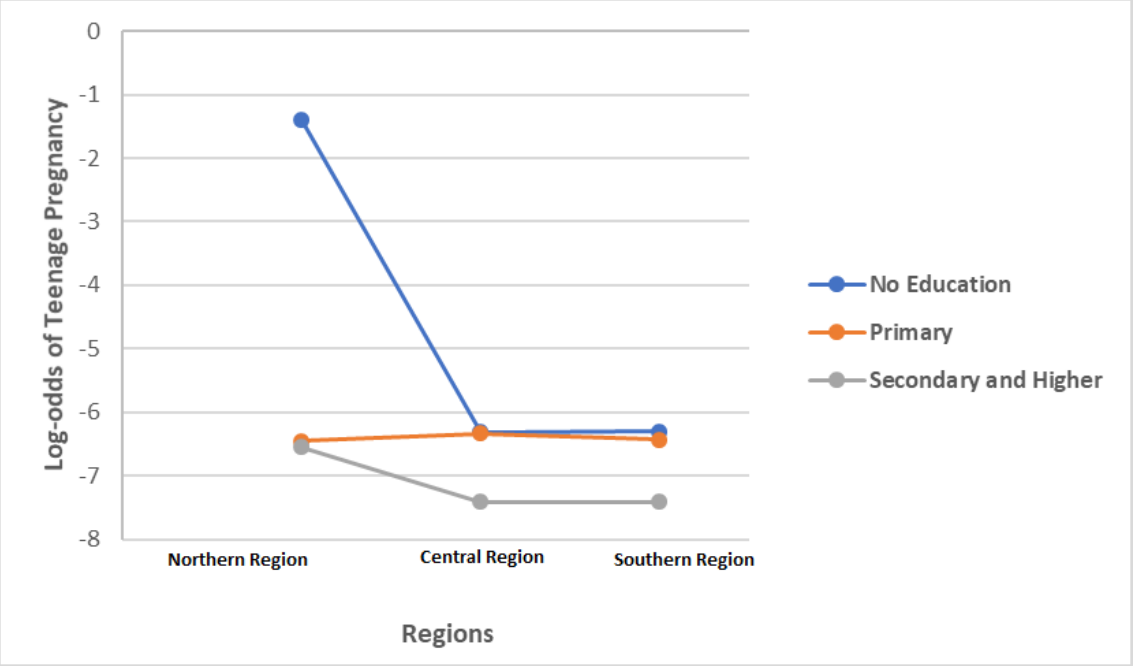

Figure 5.4: The estimated log-odds of teenage pregnancy associated with Regions and education level for the spatial GLMM model

5.6 Summary

The main objective for the application of spatial analysis of the MDHS data was to account for spatial autocorrelation in the observations between the clusters. Spatial autocorrelation was assessed in the residuals of the non-spatial GLMM presented in Section 4.5 by making use of the semivariogram as well as the Moran's I and Geary's C statistics. Geary's C statistic, which was positive, indicated significant spatial autocorrelation in the residuals at a local level, which means there may be hot spots or clusterings of high residual values in close proximity to each other. A spatial GLMM was then fitted to the MDHS data, where a spherical spatial covariance structure was specified for the cluster random effect. This spatial covariance structure made use of the cluster's geographical coordinates provided with the MDHS data. The results of this model were almost identical to that of the non-spatial GLMM.

Chapter 6

Discussion and Conclusion

The main objective of this thesis was to investigate the factors associated with teenage pregnancy among 15 to 19 year old females in Malawi. This was done by considering a design-based approach, where a survey logistic regression model that uses the survey weights during parameter estimation was employed, as well as a model-based approach that accounts for the effect of clustering. The model-based approach involved fitting a non-spatial and spatial generalised linear mixed model that incorporated a random effect at cluster level. The survey logistic regression model assumed independence in the observations, however the generalised linear mixed model allowed for possible correlations among the observations within the same cluster. Further extending the generalised linear mixed model to the spatial case aided in accounting for spatial autocorrelation between the clusters. Similarities in the observations between neighbouring clusters may be due shared resources. For example, individuals residing in neighbouring clusters may attend the same educational institution, or have the same access to health care services, or the same exposure and knowledge on how to prevent teenage pregnancies. In addition, individuals residing in neighbouring clusters may have similar cultural practices and beliefs. These shared characteristics can contribute to spatial variation and autocorrelation in the observations. While the result of Geary's C statistic indicated significant local spatial autocorrelation in the residuals of the non-spatial GLMM, the results of the

spatial GLMM were almost identical to that of the non-spatial GLMM. Therefore, it did not warrant the additional complexity of the spatial covariance structure into the GLMM.

This thesis used data from 2648 sexually active females between the ages of 15 and 19 years old obtained during the 2015-16 Malawi Demographic and Health Survey. The overall observed prevalence of teenage pregnancy among the sample was 57.7%. However, this prevalence varied between 56.7% and 62.3% based on the region of residence in Malawi, where the Northern region experienced the greatest burden. All three of the statistical models revealed that a female's age, the event of hearing of family planning on the radio, their union type, socio-economic status, contraceptive use, education level and the event that they have tested for HIV were significantly associated with the likelihood of teenage pregnancy. In addition, the GLMM revealed that the gender of the head of household was significantly associated with teenage pregnancy and the SLR revealed that the age at first sex and the region of residence had a significance association with teenage pregnancy.

Some of the important findings of this study include an increased likelihood of teenage pregnancy with an increase in age, however it is noted that we do not know the age at which the individual experienced pregnancy. The likelihood of teenage pregnancy was significantly higher for those who had not recently heard about family planning on the radio. In addition, the likelihood of teenage pregnancy was highest among those with no formal education, especially in the Northern region of Malawi. These two findings are in agreement with other studies which have highlighted the need for education and awareness of family planning, which can go hand in hand with sex education within schools (Mjwara, 2014). In general, individuals who complete school are better equipped to make rational decisions about their sexual behaviour, possibly avoiding pregnancy. While this study revealed a lower likelihood of teenage pregnancy among those who do not use contraceptives,

this may be an indication of a low level of knowledge of the correct use of contraceptives to avoid pregnancies, as also found in other studies (Maharaj, 2006; Makinwa-Adebusoye, 1992). Those with a middle socio-economic background were associated with an increased likelihood of teenage pregnancy compared to those with a rich socio-economic background. This result is consistent with that of Kaphagawani (2006) and Mutara (2015), where a lower socio-economic status has been shown to contribute to teenage pregnancy in multiple ways. Financial inadequacies can influence a young girl to leave school early and enter into a sexual relationship, therefore making them more at risk for pregnancy. In addition, this study revealed that females in the Muslim religious group were significantly more likely to experience teenage pregnancy compared to those in the Christian religious group. However, this could be as a result of early marriage practises in the Muslim religion. The interaction between type of union and age at first sex revealed a higher likelihood of teenage pregnancy among those who were currently or formerly in union regardless of the age at first sex. This suggests that a fair number of the teenage pregnancies may have been planned due to marriage.

There are numerous limitations associated with this study. Firstly, the cross-sectional nature of the data means we cannot establish a causal effect between teenage pregnancy and the factors considered or effect over time. In addition, the results of the survey were based on self-reporting, which can lead to under-reporting of certain important but sensitive information. This study was also not able to consider the effect of other factors that have been shown to be associated with teenage pregnancy, such as abuse, parents' education level, and whether or not the pregnancy was planned, among others, as this information was not available in the data. Data on substance abuse among the participants was collected, however substance use was extremely low among the sample and thus was not considered in this study. Data on religion had Christianity and Catholicism classified as separate religions groups, although this is widely considered as one religion, hence the likelihood that

respondents were confused was high.

In light of the findings in this study, it is recommended that appropriate educational programmes regarding family planning be prioritised. Such programmes should educate young girls on the benefits, and correct and safe use of contraceptives as well as encourage them to complete their schooling. In addition, programmes that empower young girls against the pressures of early marriage or early sexual activities would also support the endeavour of lowering the burden of teenage pregnancy. While this study considered the spatial autocorrelation in the observations based on the geographical coordinates of the clusters, the method of kriging to predict the likelihood of teenage pregnancy at unmeasured locations was not considered. This presents an extension to the spatial analysis already performed in this study and a possible future direction. Spatial modelling can also aid in identifying the geographical areas of Malawi that experience the highest burden of teenage pregnancy. This will assist in developing a more targeted approach to interventions, which will enable a more effective delivery system of limited resources in the country.

References

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

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.

Avert (2019). HIV and AIDS in Malawi. https://www.avert.org/professionals/hiv-around-world/sub-saharan-africa/malawi. [Online; accessed June 2020].

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

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

Blanc, A. K., & Way, A. A. (1998). Sexual behavior and contraceptive knowledge and use among adolescents in developing countries. *Studies in family planning*, (pp. 106–116).

Cavazos-Rehg, P. A., Krauss, M. J., Spitznagel, E. L., Schootman, M., Cottler, L. B., & Bierut, L. J. (2011). Substance use and the risk for sexual intercourse with and

without a history of teenage pregnancy among adolescent females. *Journal of studies on alcohol and drugs*, *72*(2), 194–198.

Chalasani, S., Kelly, C. A., Mensch, B. S., & Soler-Hampejsek, E. (2012). Adolescent pregnancy and education trajectories in malawi. *Pregnancy and Education, Extended A*, *21*.

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].

Chipeta, E. K., Chimwaza, W., & Kalilani-Phiri, L. (2010). Contraceptive knowledge, beliefs and attitudes in rural malawi: misinformation, misbeliefs and misperceptions. *Malawi Medical Journal*, *22*(2).

Cliff, A., & Ord, K. (1972). Testing for spatial autocorrelation among regression residuals. *Geographical analysis*, *4*(3), 267–284.

Cressie, N. (1991). *Statistics for Spatial Data*. John Wiley & Sons, Inc.

Geary, R. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician*, *5*, 115–145.

Gita D., R. C. S. e. T. D. K., Mishra (2009). Early life circumstances and their impact on menarche and menopause. https://www.medscape.com/viewarticle/589004_3.

Google Earth Engine (2020). Vector to raster interpolation. https://developers.google.com/earth-engine/guides/interpolation. [Online; accessed December 2020].

Habitu, Y. A., Yalew, A., & Bisetegn, T. A. (2018). Prevalence and Factors Associated with Teenage Pregnancy, Northeast Ethiopia, 2017: A Cross-Sectional Study. *Journal of Pregnancy*, *vol. 2018*(Article ID 1714527), 1–7.

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.

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.

International Bank for Reconstruction and Development / The World Bank (2016). Adolescent Girls in Malawi: Executive Summary. [Policy Brief: Malawi].

Jewkes, R., Vundule, C., Maforah, F., & Jordaan, E. (2001). Relationship dynamics and teenage pregnancy in south africa. *Social science & medicine*, *52*(5), 733–744.

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

Journal, A., & Huijbregts, C. (1978). *Mining Geostatistics*. New York: Academic Press.

Kaphagawani, N. C. C. (2006). *Risk factors for unwanted/unplanned teenage pregnancy in Zomba District, Malawi*. Ph.D. thesis.

Karen Grace-Martin (2011). Covariance Matrices,Covariance structures and Bears. https://www.theanalysisfactor.com/covariance-matrices/. [Online; accessed August 2021].

Kauye, F., & Mafuta, C. (2007). Malawi. *International Psychiatry*, *4*(1), 9–11.

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

Leung, Y., Mei, C. L., & Zhang, W. X. (2000). Testing for spatial autocorrelation among the residuals of the geographically weighted regression. *Environment and Planning A*, *32*(5), 871–890.

Levandowski, B. A., Kalilani-Phiri, L., Kachale, F., Awah, P., Kangaude, G., & Mhango, C. (2012). Investigating social consequences of unwanted pregnancy and unsafe abortion in malawi: the role of stigma. *International Journal of Gynecology & Obstetrics*, *118*(S2).

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.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for Mixed Models*. SAS Institute Inc, 2 ed.

Lloyd, C. D. (2010). *Local models for spatial analysis*. CRC press.

Machira, K., & Palamuleni, M. E. (2017). Health care factors influencing teen mothers' use of contraceptives in malawi. *Ghana Medical Journal*, *51*(2), 88–93.

Maharaj, P. (2006). Reasons for condom use among young people in kwazulu-natal: prevention of hiv, pregnancy or both? *International family planning perspectives*, (pp. 28–34).

Makinwa-Adebusoye, P. (1992). Sexual behavior, reproductive knowledge and contraceptive use among young urban nigerians. *International Family Planning Perspectives*, (pp. 66–70).

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.

Mjwara, N. P. (2014). *A qualitative study of early child bearing: experiences of Black woman in a South African township.*. Ph.D. thesis.

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

Moyo, S., & Sill, M. (2014). *The southern African environment: Profiles of the SADC countries*. Routledge.

Mutara, N. M. G. (2015). Factors contributing to teenage pregnancies in a rural community of zimbabwe. *Journal of Biology, Agriculture and Healthcare*, *5*(14).

National Statistical Office (Malawi; NSO), I. (2017). 2015–16 malawi demographic and health survey key findings.

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

Office/Malawi, N. S., & ICF (2015-16). *Malawi Demographic and Health Survey 2015-16*. Zomba, Malawi: National Statistical Office.

Organization, W. H., et al. (2004). *Adolescent pregnancy*. World Health Organization.

PennState (2020). Analysis of variance and designs of Exp. https://online.stat.psu.edu/stat502/lesson/10/10.3. [Online; accessed August 2021].

Rabiu, A., et al. (2018). The role of traditional contraceptive methods in family planning among women attending primary health care centers in kano. *Annals of African medicine*, *17*(4), 189.

Reade, S., Zewotir, T., & North, D. (2016). Modelling household electricity consumption in ethekwini municipality. *Journal of Energy in Southern Africa*, *27*(2), 38–49.

SAS, S., & Guide, S. U. (1999). Cary, nc: Sas inst.

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

Singh, S., Bankole, A., & Woog, V. (2005). Evaluating the need for sex education in developing countries: sexual behaviour, knowledge of preventing sexually transmitted infections/hiv and unplanned pregnancy. *Sex education*, *5*(4), 307–331.

Stephenson, R., Simon, C., & Finneran, C. (2014). Community factors shaping early age at first sex among adolescents in burkina faso, ghana, malawi, and uganda. *Journal of health, population, and nutrition*, *32*(2), 161.

Stone, N., & Ingham, R. (2002). Factors affecting british teenagers' contraceptive use at first intercourse: The importance of partner communication. *Perspectives on sexual and reproductive health*, (pp. 191–197).

UNICEF (2019). Early Childbearing. https://data.unicef.org/topic/child-health/adolescent-health/#:~:text=In%202018%2C%20the%20estimated%20adolescent,regional%20rate%20in%20the%20world. [Online; accessed June 2020].

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

Wellings, K., Collumbien, M., Slaymaker, E., Singh, S., Hodges, Z., Patel, D., & Bajos, N. (2006). Sexual behaviour in context: a global perspective. *The Lancet*, *368*(9548), 1706–1728.

WHO (2020). Adolescent pregnancy. https://www.who.int/news-room/fact-sheets/detail/adolescent-pregnancy. [Online; accessed June 2020].

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

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

SAS Codes

The following presents the key for each variable used in the models:

V781	Testing for HIV
SES	Socio-Economic Status
ReligionNew	Religion
FamilyPlanningRadio	Family Planning Radio
TypeOfRes	Type Of Residence
AgeFirstSex	Age at First Sex
TotalPartners	Total number of Partners
UnionType	Union Type
SexHeadHouse	Head of Household gender
EduLevel	Education Level

Survey Logistic Regression Codes

Final SLR MODEL Applied to MDHS Data

```
proc surveylogistic data=sandile;
class FamilyPlanningRadio(ref='Yes') Region(ref='Southern region') TypeOfRes (ref='Urban')
```

```
AgeFirstSex (ref='Less than 13 years old') V781 (ref='Yes') TotalPartners (ref='4+')
UnionType (ref='Never in union') SES (ref='Rich') ContraUse (ref='Yes')
SexHeadHouse (ref='Male') EduLevel (ref='Secondary and Higher') ReligionNew (ref='Christian')/
param=reference;
model TeenagePregnancy (descending)= Age TypeOfRes FamilyPlanningRadio TotalPart-
ners SES
ContraUse SexHeadHouse ReligionNew V781 EduLevel|Region AgeFirstSex|UnionType ;
strata Strata;
cluster Cluster;
weight SamplingWeight;
run;
```

Generalised Linear Mixed Model codes

Final Generalised Linear Mixed Model

```
proc glimmix data =sandile method=laplace;
class FamilyPlanningRadio(ref='Yes') Region(ref='Southern region') TypeOfRes(ref='Urban')
AgeFirstSex(ref='Less than 13 years old')
V781(ref='Yes') UnionType(ref='Never in union') TotalPartners(ref='4+') SES(ref='Rich')
ContraUse (ref='Yes')
SexHeadHouse (ref='Male') EduLevel (ref='Secondary and Higher') ReligionNew
(ref='Christian') ;
model TeenagePregnancy (descending) = Age FamilyPlanningRadio TypeOfRes To-
talPartners SES
ContraUse SexHeadHouse ReligionNew V781 EduLevel|Region AgeFirstSex|UnionType
/ link=logit dist=binary oddsratio solution ;
random int / subject=Cluster type=VC ;
covtest zerog ;
run;
```

Spatial Generalised Linear Mixed Model codes

Variogram code

```
proc variogram data=sandile plots(only)=semivar;
compute autocorrelation lagd=0.16 maxlag=26;
coordinates xc=longitude yc=latitude;
model form=(sph);
var Residuals;
run;
```

Final Spatial Generalised Linear Mixed Model

```
proc glimmix data=sandile method=laplace;
class FamilyPlanningRadio(ref='Yes') Region(ref='Southern region') TypeOfRes(ref='Urban')
AgeFirstSex(ref='Less than 13 years old')
V781(ref='Yes') UnionType(ref='Never in union') TotalPartners(ref='4+') SES(ref='Rich')
ContraUse (ref='Use Contraceptives')
SexHeadHouse (ref='Male') EduLevel (ref='Secondary and Higher') Religion (ref='Christian');
model TeenagePregnancy (descending) = Age FamilyPlanningRadio TypeOfRes To-
talPartners SES
ContraUse SexHeadHouse ReligionNew V781 EduLevel|Region AgeFirstSex|UnionType
/ link=logit dist=binary oddsratio solution ;
random int / subject=Cluster type=sp(sph)(longitude latitude);
covtest zerog ;
run;
```