A Modelling Approach to the Analysis of Complex Survey Data

A thesis submitted in fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Faculty of Science

in the

DEPARTMENT OF STATISTICS

RHODES UNIVERSITY

by

Olwethu Dlangamandla

Supervisor: Dr Amos Chinomona

Co-supervisor: Mr Jeremy Baxter

March 2021

Declaration

I hereby declare that this thesis represents my own work which has been done after registration for the degree of Master of Science at Rhodes University, and has not been previously submitted to any other institution for obtaining any other qualification.

Olwethu Dlangamandla (Student)

Dr Amos Chinomona (Supervisor)

Mr Jeremy Baxter (Co-supervisor)

March 2021

Abstract

Surveys are an essential tool for collecting data and most surveys use complex sampling designs to collect the data. Complex sampling designs are used mainly to enhance representativeness in the sample by accounting for the underlying structure of the population. This often results in data that are non-independent and clustered. Ignoring complex design features such as clustering, stratification, multistage and unequal probability sampling may result in inaccurate and incorrect inference. An overview of, and difference between, design-based and model-based approaches to inference for complex survey data has been discussed. This study adopts a model-based approach. The objective of this study is to discuss and describe the modelling approach in analysing complex survey data. This is specifically done by introducing the principle inference methods under which data from complex surveys may be analysed. In particular, discussions on the theory and methods of model fitting for the analysis of complex survey data are presented. We begin by discussing unique features of complex survey data and explore appropriate methods of analysis that account for the complexity inherent in the survey data. We also explore the widely applied logistic regression modelling of binary data in a complex sample survey context. In particular, four forms of logistic regression models are fitted. These models are generalized linear models, multilevel models, mixed effects models and generalized linear mixed models. Simulated complex survey data are used to illustrate the methods and models. Various **R** packages are used for the analysis. The results presented and discussed in this thesis indicate that a logistic mixed model with first and second level predictors has a better fit compared to a logistic mixed model with first level predictors. In addition, a logistic multilevel model with first and second level predictors and nested random effects

provides a better fit to the data compared to other logistic multilevel fitted models. Similar results were obtained from fitting a generalized logistic mixed model with first and second level predictor variables and a generalized linear mixed model with first and second level predictors and nested random effects.

Keywords: Complex survey data, simple random sampling, complex sampling, stratification sampling, clustering sampling, multi-stage sampling, systematic sampling, model-based approach, logistic regression models, generalized linear models, multilevel models, mixed effects models and generalized linear mixed models.

Contents

List of Tables

List of Abbreviations

AIC	Akaike Information Criterion
AIDS	Acquired Immune Deficiency Syndrome
BIC	Bayesian Information Criterion
deff	Design Effect
GLMs	Generalized Linear Models
GLMMs	Generalized Linear Mixed Model
GLS	Generalized Least Squares
HIV	Human Immunodeficiency Virus
ICC	Inter-Cluster Correlation
iid	Independently and Identically Distributed
IGLS	Iteratively Generalized Least Squares
IRLS	Iteratively Reweighted Least Squares
LM	Linear Models
LMMs	Logistic Mixed Models
MAR	Missing At Random

MCAR	Missing Completely At Random
meff	Misspecification Effect
MNAR	Missing Not At Random
MLM	Multilevel Logistic Modelling
ML	Maximum Likelihood
MPML	Pseudo-Maximum Likelihood Estimator
MSc	Master of Science
NRF	National Research Foundation
OLS	Ordinary Least Squares
OR	Odds Ratio
PSUs	Primary Sampling Units
REML	Restricted Maximum Likelihood
SADHS	South Africa Demographic and Health Survey
SSU	Secondary Sample Units
SRS	Simple Random Sampling without Replacement
SRSWR	Simple Random Sampling with Replacement
WLS	Weighted Least Squares

Acknowledgments

This study would not have been possible without the support of many people. Firstly, I would like to express my deepest gratitude to my primary supervisor, Dr Amos Chinomona, for his vast knowledge on the topic, patience, motivation and excellent guidance throughout. Truly, I can not imagine having a better supervisor for this research study. Dr Amos Chinomona together with Mr Jeremy Baxter believed in me from the beginning to the end of this dissertation that support helped me through this research study. I would also like to extend my gratitude to the staff members of the Statistics Department at Rhodes University for their helpful encouragement, comments and suggestions. I want to also acknowledge the financial support from National Research Foundation (NRF) block grant, thank you for funding this project in 2018. A special thank you to my family, partner and friends, I acknowledge the contributions you have made towards my studies and I will forever remain grateful. Your love, friendship and best wishes helped me to stay focused on this project and gave me strength to continue through hard times.

Chapter 1

Introduction

Large sample surveys often use complex sampling designs instead of simple random sampling (SRS) to collect data. These complex sampling methods are used mainly to improve the quality of data and enhance population coverage while balancing issues of cost and feasibility. The complex sampling designs often include stratification, clustering, multistage sampling and unequal probability sampling as described in Lohr (1999) and Lee and Forthofer (2006). Statistical approaches to analysing such data from complex samples ought to accommodate the distinct sampling features embedded in the survey data to ensure that statistics obtained from and models fitted to such data provide appropriate estimates and inferences. Such survey data are often non-independent, are clustered and hierarchical in structure. Most statistical theory usually explore statistical analysis methodology that is applied to data obtained from simple random samples (SRS), whereas most survey data analyzed in practice originate from non-SRS designs.

According to Lumley (2004) the features of complex sampling designs often introduce a number of complexities, although these often bring gains in precision. Therefore, taking into account the sampling design is important in complex survey data analysis. If these features are not included in the analysis, inference from such data may be incorrect (Chambers and Skinner (2003) and Lohr (2010)).

Many approaches have been proposed for modelling and analysing complex survey data. These

approaches often differ in the conditions underlying their use, the data requirements, the target of inference, statistical efficiency, computational demands, and the skills required for their implementation Pfeffermann (2011).

In this study a logistic regression model is fitted in a generalized linear modelling (GLM) framework taking into account the complex sampling design features explored. Adjustments to the conventional methods to take into account the survey design elements are necessary in order to make correct inferences about the population. As a result survey adjusted logistic regression models were fitted. Specifically linear mixed effects models (LMMs) and generalized linear mixed effects models (GLMMs), that have increased in popularity as expressed by Douglas et al. (2014), that extend the traditional linear models to include a combination of fixed and random effects as predictor variables are explored. The inclusion of the random effects allows one to explicitly model the non-independent observational units that characterize most practical hierarchical structured and clustered data. Random effects allow the estimation of variance in the response variable within and among these groups. Multilevel models with parameters that vary at more than one level are often used for data from complex surveys involving multistage sampling, stratification, and unequal sampling probabilities are presented. We also consider multilevel regression modelling that allows the use of both primary and secondary units in the same model as defined in Chapter 4. These models include both fixed and random effects developed from hierarchical approaches for analysis Wong and Mason (1985) and Goldstein (2011).

The main objective of this research study is to explore the modelling approach of complex survey data. We explore the widely used logistic regression analysis statistical modelling methods for binary data. The goal is to engage with the procedures and fit different models to simulated complex survey data analysis.

Chapter 2 of this research study begins with a discussion of the features of basic sampling designs and their use in complex samples, and the impact of sample characteristics on variance estimation and their relationship to sample design effects. The fundamental concepts underlying analysis of survey data, including the design effect, misspecification effect and the use of weights

are also discussed. The theory behind the differences between model-based and design-based estimation is presented.

In Chapters 3, 4 and 5, the theory underlying the models used, model estimation methods and model fitting approaches are discussed. Data were simulated to resemble complex survey data with an assurance that the assumptions underlying the linear modelling formulations are met. The models presented in the earlier chapters are fitted to the simulated data. The results are presented in Chapter 7. The necessary sampling design features are incorporated into the simulation process.

Various **R** packages are used for analysing the complex survey data, see for example Cochran (1977) and Hosmer and Lemeshow (2000). Chapter 7 presents various illustrations of survey data analysis. The emphasis is on the demonstration of the effects of incorporating weights and the data structure on the analysis. Complex survey data which are simulated in Chapter 6 are used to illustrate various analyses, including descriptive analysis, linear regression analysis and logistic regression analyses using the model-based perspective. The variables used in each analysis are selected to illustrate the methods rather than to present substantive findings. Finally Chapter 8 finishes with conclusion and recommendations.

Chapter 2

General Fundamental Concepts

2.1 Introduction

This chapter outlines the fundamental theory underpinning the basic sampling designs that are encountered in practice and the methods of analyzing data obtained from such designs. In particular, this chapter reviews the fundamental theory surrounding the analysis of complex survey data including the use of sampling weights to ensure better statistical inference. Missing data are often inevitable in complex survey data, usually in the form of nonresponse and has a potential bias in research findings, Lee and Forthofer (2006). Sources of, and methods for, handling missing data are also discussed. Alternative approaches to making inferences for complex survey data, the design-based approach and the model-based approach are discussed. Much of the discussion from this chapter on survey sampling theory, analyzing complex survey data and how to overcome problems that often arise, are explained in detail in Kish (1965), Lee and Forthofer (2006), Skinner et al. (1989) and Lohr (1999).

2.2 Simple Random Sampling

A simple random sampling design is the most well-known and widely applied probability sampling method in which the sampling can be conducted in two ways, namely simple random sampling with

replacement (SRSWR) and simple random sampling without replacement (SRS). Under SRSWR, a sample of size n is selected from a population of size N by selecting units one by one replacing the units at each draw, Cochran (1977). This method can be thought of as taking n independent samples of size 1 with replacement. The first sample element is selected with probability $\frac{1}{N}$, then it is placed back into the sample, the second and subsequent element is also selected with probability $\frac{1}{N}$ since the population size does not change. This procedure is repeated until the desired sample size n is achieved and it may include duplicate items from the population. Under simple random sampling without replacement, units are not returned after each draw. The aim of simple random sampling without replacement is to ensure that each distinct sample of n elements has the same probability of being selected. There are $\binom{N}{n}$ possible subsets of size n that can be selected from a population of s ize N each with probability $\frac{1}{\binom{N}{n}}$ of being the selected sample. Lohr (1999) describes why the simple random sampling without replacement is advantageous, primarily the capability that the sample contains no duplicates.

2.3 Complex Sampling

Most practical data are rarely collected using a simple random sampling design, rather complex sampling designs are used. As mentioned in Chapter 1, these designs often include stratification, clustering, multistage sampling and unequal probability sampling, see Lohr (1999) and Lee and Forthofer (2006). These are mainly aimed at reflecting the underlying structure of the target population hence ensuring representativeness. Appropriate analysis of such data needs to take account of the complex sampling designs. The complex designs often induce dependence among sampling units thereby rendering most conventional statistical methods inappropriate. The fundamental theory that underlies linear models and generalized linear models (GLMs) assumes independence of sampling units, thus these methods become inappropriate for analyzing such data. In particular GLMs, as extensions of linear models, are mainly designed to accommodate both normal and non-normal data, however both assume independence of observation units. Nonindependence of units

due to complex designs often result in homogeneity that may produce small standard errors as compared to simple random samples, as expressed by Kish (1965) and Skinner et al. (1989). This implies that methods that accommodate complex designs are required in analyzing such data. We present the details of these complex designs in the proceeding subsections.

2.3.1 Stratification Sampling

Under stratified random sampling, the population is first divided into separate nonoverlapping groups, called strata. Subsequently separate independent simple random samples (SRS) are selected from each stratum. A combination of the sum total of these SRSs forms the full sample. Stratification is recommended for several reasons including accounting for natural groupings within a population, decreasing the size of standard errors relative to an SRS of the same size, enabling the deliberate oversampling of specific subpopulations, facilitating the use of different survey methods within strata, and permitting analysis within strata Heeringa et al. (2010). The main limitations to its use are the availability of strata information on all members of the population and the extra complexity in estimation and inference.

2.3.2 Cluster Sampling

In cluster sampling, observation units in the population are aggregated into larger sampling units called clusters. The population is divided into groups each of which is similar to stratified sampling. However, particularly in a one-stage sampling, clusters are selected at random, hence they (clusters) are treated as the observation units. Clusters are often of unequal sizes and Kish (1965) lists some problems that arise due to unequally sized clusters when all elements in the selected clusters are sampled. Under cluster sampling, the sample size is regarded as a random variable since it depends on the size of the particular cluster selected. Another problem is that the variance estimates are often not unbiased and the formulas are usually complex. The advantages of clustering include a reduction of costs, time and convenience for a given accuracy.

2.3.3 Multi-Stage Sampling

If the population contains a very large number of units of varying characteristics, then multistage sampling is often used Samphath (2005). A first-stage sample is selected based on clusters, then a second-stage sample is created by drawing subsamples from the selected clusters. If the second-stage sample is based on subclusters, then a third stage to the sample could be added. Multistage sampling can be considered a complex form of cluster sampling because it involves dividing the population into groups or clusters. One or more clusters are chosen at random and every element within the chosen cluster(s) is sampled. Multistage sampling is the most common complex sampling design used in practice and often includes other basic designs such as SRS, stratification and cluster sampling. For example in the first stage the population is stratified, then for the second stage from each stratum a SRS of clusters is selected and so forth.

2.3.4 Systematic Sampling

Systematic sampling is a sampling method used when the population elements are arranged in a specific order from which a sample can be drawn in a systematic way rather than generating a simple random sample. To select a systematic sample of n elements from a population of N elements, the N elements in the population are divided into n groups of k elements, where $k = \frac{N}{n}$ and then the first element is randomly selected out of the first k elements in the population. Thereafter, every k^{th} unit is selected until a sample of n elements is obtained. This sampling design is called an every k^{th} systematic sample Cochran (1977) and is considered as a convenient substitute for simple random sampling. Systematic sampling has some advantages over other methods in general and over simple random sampling in particular. Some of these advantages include that the population size does not need to be known for systematic sampling and any bias in the population will tend to be more represented in a systematic rather than in simple random sampling Elsayir (2014). Hence a systematic sample can be expected to give more precise estimates. The main disadvantages are that conventional methods for estimating the sampling variance do not apply, and that systematic sampling may be poor when the ordering of the population is based on inaccurate knowledge,

Zhang (2008).

2.4 Basic Concepts for the Analysis of Survey Data

2.4.1 The Weighting

As described by Lumley (2010), if an individual element is sampled with sampling probability π_i it represents $\frac{1}{\pi_i}$ elements in the population and the the value $\frac{1}{\pi_i}$ is called the sampling weight. Sampling weights are used to correct discrepancies in the sample that might lead to differences in representativeness between the sample and the target population. Such discrepancies include the selection of units with unequal probabilities, non-coverage of the population and non-response. There are two approaches to analyse complex survey data, design-based and model-based as explained in detail in section 2.5. These approaches differ in how they incorporate complexities of sampling designs, weights are often used for this. Design-based analysis uses the sampling design as the sole source of variability and thus the sampling weights are used to account for the sampling design as a source of variability. Model-based analysis include a model that assumes to have generated the data in addition to sampling designs and this renders the role of sampling weights to be considered unnecessary. The debate of whether weights should be used or not in model-based analysis continues. Hoem (1989) and Fienberg (1989) suggest that weights should not be used in model-based analysis with the exception of when there is informative sampling. Kalton (1989) on the other hand strongly suggests that the weights should always be used as they help against model misspecification. Lohr and Liu (1994) suggest that it is best to run two analyses, one with weights and one without weights. Then if it happens that the two analyses match, the use of weights does not matter, otherwise adjust the model until the weighted and unweighted match. For this study weights will not be used.

2.4.2 Design Effect and Misspecification Effect

Complex sampling designs tend to result in possible unequal probability of selection into the sample for individual units of analysis due to the varying sizes of clusters and strata, the lack of independence of individual units within randomly sampled clusters unlike simple random sampling. It is argued that the use of complex designs results in increased precision of estimates due to stratification, but decreased precision of estimates due to the use of clustering. The general properties of a complex sampling designs which were reviewed by West (2008) led to the use of the design effect (deff). The design effect, as defined by Kish (1965), is a ratio of the variance of an estimate obtained under a complex design to the variance of the same estimate under simple random sampling. The effect of the complex design on the variance of $\hat{\theta}$ is given by the design effect. As described by Kish (1965), let $\hat{\theta}$ denote the estimate of a finite population parameter θ obtained from a sample of size n from a complex survey design with $Var_{complex}\left(\hat{\theta}\right)$ denoting the design variance of $\hat{\theta}$. Let $Var_{SRS}\left(\hat{\theta}\right)$ denote the variance of $\hat{\theta}$ calculated under a simple random sampling design of the same sample size n, then

$$\text{deff}\left(\hat{\theta}\right) = \frac{Var_{complex}\left(\hat{\theta}\right)}{Var_{SRS}\left(\hat{\theta}\right)}.$$

If $\text{deff}\left(\hat{\theta}\right) < 1$, the complex design is better, as this sampling design has produced an estimator with lower variance, than a corresponding SRS with respect to $\hat{\theta}$. $\text{deff}\left(\hat{\theta}\right)$ is completely a design-based measure. The design effect provides a measure of the precision gained or lost by the use of complex designs instead of an SRS Lohr (1999). As much as deff is important, it is not a method to avoid the calculation of variances because an estimation of the variance from the complex design is needed to find the design effect.

As described by Skinner et al. (1989) the misspecification effect (meff) measures the effect on the sampling variance of an estimator under incorrect specification of both the sampling scheme and the model considered. Misspecification effect is similar to Kish's design effect, except that meff is used at the analysis stage, where $Var_{SRS}\left(\hat{\theta}\right)$ cannot be estimated because there is no additional

simple random sample available. $E_{complex}(v_0)$ denotes the expected variance of a complex survey design. meff is given by

$$\text{meff}\left(\hat{\theta}, v_0\right) = \frac{Var_{complex}\left(\hat{\theta}\right)}{E_{complex}(v_0)},$$

where v_0 may be a design-based estimator or a model-based estimator. The meff may be defined either as a design-based measure or as a model-based measure. When taken as a model-based measure, the quantities $E_{complex}$ and $Var_{complex}$ are based on the true model distribution. Therefore under a model-based approach, $\text{meff}\left(\hat{\theta}, v_0\right)$ depends only on the model relationship between the units in the actual sample selected and not on how the sample was selected. When $\text{meff} < 1$, v_0 is overestimating $Var_{complex}\left(\hat{\theta}\right)$ and when $\text{meff} > 1$, v_0 is underestimating $Var_{complex}\left(\hat{\theta}\right)$.

2.4.3 Handling of Missing Data in Complex Surveys

Missing data in surveys are a common occurrence and can have a significant effect on the inference that can be drawn from the collected data. Missing data in survey research have to be taken into account in analysis because they are capable of reducing the representativeness of the sample and while the analysis might run, the results may not be statistically significant because of the lost data. This may result in misleading and inaccurate conclusions. Hence it is important to properly account for missing data or non-response when conducting an analysis of survey data.

There are a variety of reasons for missing data in surveys. Brick and Kalton (1996) suggest that missing survey data results mainly from four sources namely total or unit non-response, non-coverage, item non-response and partial non-response. Brick and Kalton (1996) describes these as follows:

Total or unit non-response occurs when, for instance, an eligible sample member fails to respond possibly due to non-contact, language barrier, refusal to answer or does not provide enough information for the response to be considered usable. Unit non-response occurs when no survey data are collected from an element (unit of analysis).

Non-coverage occurs when, for instance, persons (or elements) in the target population are not

included in the sampling frame from which the sample is selected.

Item non-response occurs when, for example, participants do not provide a valid answer to a question. This may happen if, for instance, respondents feel a question is sensitive or do not know the response.

Partial non-response results when, for instance, participants in a multi phase survey provide data for some, but not all phases of data collection. For example, when the participant cuts off the interview in the middle and end up not answering some questions.

Understanding the reasons why data are missing is important as it informs appropriate methods of analysing data with missing values. There are three types of missing data, classified according to the mechanisms that cause the missing data. These are missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) Rubin (1976). Determining the mechanisms help to identify the most appropriate analysis method. These methods for handling missing data in surveys range from traditional methods, that include list-wise deletion or complete-case analysis, pairwise deletion, simple imputation and advanced methods, that include multiple imputation and maximum likelihood Chinomona and Mwambi (2015).

2.5 Approaches to Analysing Survey data

The are two fundamental approaches used for analyzing complex survey data, the design-based and the model-based approach. They differ in how they incorporate the complexities of the sampling design, such as stratification, clustering and/or unequal probabilities of selection, into the survey data analysis. They also differ in the way the stochastic process that is responsible for generating the randomness in the data is handled. The design-based approach is one which has been traditionally used in sample surveys and it is called the design-based approach because the sampling design is considered important Cassel et al. (1977). In contrast to a design-based approach is a model-based approach assumes a superpopulation approach. The superpopulation approach assumes an infinite population and the sampling designs are considered to be unimportant because the sample

is assumed to be given and inferences are conditional on a given sample.

2.5.1 The Design-based Approach

Hargovan (2007) outlines that under the design-based inference population parameters are considered to be fixed. Suppose that U represent the finite population of N elements, and θ_U denotes a target parameter which is to be estimated by $\hat{\theta}$. This target parameter is a function of the finite population quantities, however as a census is usually not taken, the parameter must be estimated from a sample of n elements. For a given sampling design, $p_U(.)$, suppose that there are t_n possible samples of size n from the population. Let $p_U(i)$ be the probability of selecting the i^{th} sample and define $\hat{\theta}_U^{(i)}$ to be the estimate of θ_U from the i^{th} sample. $\{\hat{\theta}^{(i)}_{p_U,p_U(i)}\}_{i=1}^{t_n}$ defines the randomization distribution for sampling design $p_U(.)$. The randomization distribution is also referred to as the sampling distribution, Lohr (1999). The estimates of the parameters from all the possible samples from the population are needed to know the randomization distribution. However in practice only one sample is typically selected from the population. To account for this, sampling weights can be used in the analysis. The sampling weight, is defined as in section 2.4.1, is denoted by $w_{ki} = \frac{1}{\pi_{ki}}$, where π_{ki} is the probability that the ki^{th} element is included in the sample Korn and Graubard (1999). Note that the ki subscripts will vary according to the sampling design, for example k may refer to clusters instead of strata. The sampling weight, w_{ki}, corresponds to the number of items which element ki represents in the population. These weights are important for scaling down the bias as design-based analyses without the weights tend to have very large bias for gaining information about the non-realized samples and accounting for the sampling structure. Therefore the key distinct points for design-based analysis are that the population of interest is the specific finite population which was sampled, that variability is induced by the sampling design and the fact that sampling weights are crucial for the analysis.

2.5.2 The Model-based Approach

In a model-based approach to the analysis of survey data the targets of inference are the model parameters, for example regression coefficients. In a model-based approach, the interest is in making inferences about $\boldsymbol{\beta}$, as it is assumed that y-values are from an infinite superpopulation and are independent Chambers and Skinner (2003). Suppose that under a model it is assumed that the sample observations, $y_1,...,y_n$, are random variables which, given $x_1,...,x_n$, satisfy

$$y_t = x_t^T \boldsymbol{\beta} + \varepsilon_t \tag{2.1}$$

for $t = 1,...,n$, where ε_t has mean 0 variance σ^2 and is uncorrelated with ε_t^T for $t \neq t^T$, where T is the transpose. Standard statistical theory would imply that the ordinary least squares estimator for β is the best linear unbiased estimator, see section 3.1, given by

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

with mean

$$E\left(\hat{\beta}\right) = \beta$$

and variance

$$\text{Var}\left(\hat{\beta}\right) = \sigma^2 \left(X^T X\right)^{-1}$$

where $\boldsymbol{X}$ and $\boldsymbol{y}$ are based on the sample. The model-based parameter estimates for instance $\hat{\boldsymbol{\beta}}$, are calculated by ignoring the sample survey design and treating the data as a simple random sample from an infinite population. In this approach the sample design is assumed to be ignorable Rubin (1976).

A superpopulation model is a way of formalizing the relationship between a target variable and auxiliary data. The superpopulation can be given several interpretations, the basic one being to view the finite population as a random realization of the superpopulation Cassel et al. (1977). In

classical sampling theory a model-based approach is based on a superpopulation model, say ξ. This model, ξ, is used to predict the nonsampled values of the population Rueda and Sanchez-Borrego (2009), Graubard and Korn (2002).

To implement this approach, one needs to formulate a statistical model that describes how the target response variable is generated. A hypothetical (infinite) population is then defined by all possible values of the target response variable generated by the model. Subsequently an assumption about the source of the underlying variability in the population data is made. Such an assumption allows one to treat the errors in a regression model as independently and identically distributed (iid) realisations of a random variable with mean zero and a constant variance. As described by Vallient (2009), in the model-based approach a prediction is made for each nonsample unit and the total of the predictions is added to the observed sample total to estimate the parameters such as T, where the population total of Y is $T = \sum_U Y_i$, where U denotes the set of N units in the population.

Hargovan (2007) further explains that under model-based inference one assumes that for element ki there is an associated value Y_{ki} which was randomly generated by the model ξ. This model is usually a population model, that is the model for the stochastic process which is assumed to have generated the finite population values $y_1, ..., y_N$, as explained by Chambers and Skinner (2003). This is considered as a two-stage process where the first stage generates the finite population of size N from the model ξ, and the second stage samples the finite population. In this scenario, the estimation process can be for a finite population parameter or a superpopulation parameter. The finite population parameter estimation mimics the design-based analysis, while still assuming that the Y_{ki} value associated with element ki is random. Model-based statistical inference has gained more attention in complex survey data analysis because of its several advantages over design-based approaches. These advantages include inference that can be obtained from standard software and estimation routines, so that special techniques for weighted estimation are not needed. Most software packages permit examination of diagnostics for model-based approaches, but they do not have routines for diagnostics for design-based approaches. In addition to these advantages, model-based estimators of analytic parameters typically have smaller standard errors than design-based

estimators Pfeffermann (1993).

The proceeding chapters present the underlying theory of the models that are considered in this thesis. In particular, details of how each model can be fitted to survey data are discussed. Since we will consider a binary response variable, the theory of logistic models is presented alongside.

Chapter 3

Generalized Linear Modelling for Survey Data

3.1 Generalized Linear Models

Generalized linear models (GLMs) were proposed by Nelder and Wedderburn (1972) and further expanded by McCullagh and Nelder (1989), GLMs are an extension of linear models and are designed to accommodate both normal and non-normal data. There are three components that make up a GLM, namely the random component, the system component and the link function. The random component refers to the probability distribution of the response variable, denoted by $\mathbf{Y}$. The systematic component specifies the explanatory variables denoted by $\mathbf{X}=(x_1, x_2, ..., x_p)$ in the model given as a combination of linear predictors. The link function, denoted by $\boldsymbol{\eta}$ or $g(\boldsymbol{\mu})$ specifies the link between the random and the systematic components. It gives how the expected value $\boldsymbol{\mu} = E(\mathbf{Y})$ of the response relates to the linear predictor of the explanatory variables.

GLMs extend the linear model by modelling $g(\mu_i) = \mathbf{X}_i\beta$ where $y_i \sim EF(\mu_i, \phi)$, where g is any smooth monotonic link function and $EF(\mu_i, \phi)$ is any exponential family of distributions. A distribution belongs to the exponential family if its distribution can be expressed as

$$f_y = (y, \boldsymbol{\theta}, \phi) = \exp\left\{\frac{y\theta - b(\boldsymbol{\theta})}{a(\phi)} - c(y, \phi)\right\} \tag{3.1}$$

where $\boldsymbol{\theta}$ is the natural parameter, ϕ is the dispersion parameter, $b(\boldsymbol{\theta})$ and $c(y, \phi)$ are known functions, Dobson (2002).Under the GLM approach the normality assumption is relaxed to include distributions such as the Gamma, Poisson, Binomial. Instead of modelling $\boldsymbol{\mu} = E(\boldsymbol{y})$, directly a link function $g(\boldsymbol{\mu})$ is given as

$$g(\boldsymbol{\mu}) = \eta = \mathbf{X}^T \boldsymbol{\beta}.$$

The mean and the variance of $\mathbf{Y}$ under GLM can be derived from equation 3.1.

The log-likelihood is given by

$$l(y, \boldsymbol{\theta}, \phi) = \log f_y(y, \boldsymbol{\theta}, \phi) = \frac{y\boldsymbol{\theta} - b(\boldsymbol{\theta})}{a(\phi)} - c(y, \phi)$$

through the relations

$$E\left(\frac{\partial l}{\partial \boldsymbol{\theta}}\right) = 0$$

and

$$E\left(\frac{\partial^2 l}{\partial \boldsymbol{\theta}^2}\right) + E\left(\frac{\partial l}{\partial \boldsymbol{\theta}}\right)^2 = 0 \tag{3.2}$$

where

$$\frac{\partial l}{\partial \boldsymbol{\theta}} = \frac{y - b'(\boldsymbol{\theta})}{a(\phi)} \tag{3.3}$$

and

$$\frac{\partial^2 l}{\partial \boldsymbol{\theta}^2} = \frac{b''(\boldsymbol{\theta})}{a(\phi)}. \tag{3.4}$$

From equations 3.3 the mean is obtained as

$$E\left(\frac{\partial l}{\partial \boldsymbol{\theta}}\right) = \frac{\mu - b'(\boldsymbol{\theta})}{a(\phi)} = 0$$

so that

$$E(Y) = \mu = b'(\boldsymbol{\theta}).$$

The variance function of Y, under a GLM, describes how the variance $\text{Var}(Y)$, depends on both the dispersion parameter ϕ and on a function of the mean μ. Using equations 3.2, 3.3 and 3.4, $\text{Var}(Y)$ is given by

$$-\frac{b''(\boldsymbol{\theta})}{a(\phi)} + \frac{\text{Var}(Y)}{a^2(\phi)} = 0$$

implying that

$$\text{Var}(Y) = b''(\boldsymbol{\theta})\, a(\phi)$$

as shown by McCullagh and Nelder (1989).

The Maximum likelihood (ML) estimation method is used to estimate the regression parameters making use of iterative numerical procedures such as the Newton Raphson or Fishers scoring method. Maximum likelihood estimation is aimed at estimating the value of the parameter, say $\boldsymbol{\theta}$, that maximizes the log-likelihood function. Let $x_1, x_2, ..., x_n$ be the observed values that have joint density $f(x_1, x_2, ..., x_n|\boldsymbol{\theta})$, the likelihood of $\boldsymbol{\theta}$ is given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} f(x_i|\boldsymbol{\theta}).$$

The log-likelihood function is

$$l(\theta|x_1, x_2, ..., x_n) = \sum_{n=1}^{N} \ln f(x_N|\boldsymbol{\theta}).$$

The process of finding the estimate of $\boldsymbol{\theta}$ involves computing the first derivatives, or the score,

and second derivatives, termed the Hessian matrix, of the log-likelihood function with respect to the parameter vector $\boldsymbol{\theta}$, as described in Jennrich and Sampson (1976) and Jorgensen (1983). In particular, the first derivative of the log-likelihood function with respect to the parameter vector $\boldsymbol{\theta}$ is given by

$$G(\theta) = \frac{\partial \ln L(\theta)}{\partial \theta}.$$

The maximum likelihood estimate of $\boldsymbol{\theta}$, denoted $\hat{\boldsymbol{\theta}}$, is obtained by solving the set of first order partial derivatives

$$G(\hat{\theta}) = \frac{\partial \ln L(\theta)}{\partial \theta} \mid_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0.$$

The second derivative of the log-likelihood function with respect to the parameter vector $\boldsymbol{\theta}$, known as the Hessian, is given by

$$H(\theta) = \frac{\partial^2 \ln L(\theta)}{\partial \theta \partial \theta'}.$$

The Hessian is used to establish a maximum for the log-likelihood function that has been attained. The Hessian also plays a role in determining the precision of the maximum likelihood estimator.

The Newton-Raphson algorithm is an iterative numerical integration techniques that employs an iterative procedure to solve nonlinear equations. This algorithm uses first order derivative vectors and second order matrices of maximized functions. In particular, the Newton-Raphson algorithm is based on the estimation of a function that has to be optimized. Millar (2011) shows how the Newton Raphson algorithm is derived from the quadratic estimate of the log-likelihood function through derivatives of linear estimation. Given $\boldsymbol{\theta}$, the first derivative of the log-likelihood is

$$l'(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta}; y)}{\partial \boldsymbol{\theta}} = G(\boldsymbol{\theta}).$$

The Newton Raphson algorithm estimates l' using an extension of the Taylor series using the pa-

rameter value $\boldsymbol{\theta}^{(k)}$. The Newton Raphson algorithm's iteration formula is given by

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - H\left(\boldsymbol{\theta}^{(k)}\right)^{-1} l'\left(\boldsymbol{\theta}^{(k)}\right)$$

or

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - H\left(\boldsymbol{\theta}^{(k)}\right)^{-1} G\left(\boldsymbol{\theta}^{(k)}\right).$$

In this case $H\left(\boldsymbol{\theta}^{(k)}\right)$ is the Hessian matrix, the second derivative of $l(\boldsymbol{\theta})$ evaluated on the k^{th} iteration. To implement the Newton Raphson algorithm both the first and second derivatives of the log-likelihood function, $G(.)$ and $H(.)$ respectively, are needed at each iteration. A theoretical advantage of the Newton-Raphson procedure is its quadratic convergence rate Jennrich and Sampson (1976).

An alternative but similar method is the Fisher scoring algorithm. The main difference is that the Fisher scoring uses the information matrix. The information matrix is equal to the negative expected Hessian of the log-likelihood while the Newton Raphson algorithm uses the second derivative matrix of the observed value Jennrich and Sampson (1976). Fisher scoring's iteration formula is given by

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + I\left(\boldsymbol{\theta}^{(k)}\right)^{-1} I'\left(\boldsymbol{\theta}^{(k)}\right),$$

where $I\left(\boldsymbol{\theta}^{(k)}\right)$ is the k^{th} estimate of the observed information matrix given by

$$I(\boldsymbol{\theta}) = -E\left(\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j \partial \boldsymbol{\theta}_k}\right).$$

As discussesd by Dobson (2002) and McCullagh and Nelder (1989), hypothesis tests applied to GLMs do not require normality of the response variable, nor do they require homogeneity of variances. Hypothesis tests under the GLM framework are performed by comparing how well two nested models fit the data. As per Dobson (2002), let M_0 denote a reduced model or a special case of a full or more general model denoted by M_1. If M_0 corresponds to H_0 and if it fits the data as

well as M_1 then M_0 is preferred hence H_0 is retained otherwise H_0 is rejected in favour of H_1.

For instance, consider a more general hypothesis

$$H_1 : \beta = \beta_1 = \left(\beta_1, \quad . \quad . \quad . \quad , \beta_p \right)',$$

corresponding to M_1, with $p < N$. Consider the null hypothesis

$$H_0 : \beta = \beta_0 = \left(\beta_1, \quad . \quad . \quad . \quad , \beta_q \right)',$$

corresponding to model M_0 with $q < p < N$.

H_0 can be tested against H_1 using the difference of the deviance statistics, namely

$$D = D_0 - D_1$$

$$= 2\left[l\left(b_{max}; y\right) - l\left(b_0; y\right)\right] - 2\left[l\left(b_{max}; y\right) - l\left(b_1; y\right)\right],$$

$$= 2\left[l\left(b_1; y\right) - l\left(b_0; y\right)\right],$$

then

$$\chi^2_{(N-q)}, \chi^2_{(N-p)} \text{ and } \chi^2_{(p-q)}.$$

If both models describe the data well then $D_0 \sim \chi^2_{(N-q)}$ and $D_1 \sim \chi^2_{(N-p)}$ so that $D \sim \chi^2_{(p-q)}$, provided that certain independence conditions hold. If the value of D is consistent with the $\chi^2_{(p-q)}$ distribution, then generally the model M_0 corresponding to H_0 is preferred because it is reduced, it is a parsimonious model, that is, it has less parameters. If the value of D is in the critical region then H_0 is rejected in favour of H_1 on the grounds that model M_1 provides a significantly better description of the data.

By assuming that the distribution of Y belongs to the exponential family it is possible to derive maximum likelihood estimates for the coefficients β of a GLM Dobson (2002). The maximum likelihood estimates of β can be found through the log likelihood function

$$l(\boldsymbol{\mu}; y) = \sum_{i=1}^{n} \log f(y_i, \beta).$$

The likelihood is minimized by an iteratively reweighted least squares (IRLS) algorithm.

Stirling (1984) and Green (1984) provide a trial estimate of the parameters $\hat{\beta}$. The estimated linear predictor $\hat{\eta}_i = \boldsymbol{x}_i' \hat{\beta}$ is calculated and used to obtain the fitted values $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$. Using these quantities, a set of predictors corresponding to these effects are computed, called the working dependent variable and denoted as z_i, which is a linearized form of the link function applied to Y and is calculated as

$$z_i = \hat{\eta}_i + (y_i - \hat{\mu}_i) \frac{\partial \eta_i}{\partial \mu_i}. \tag{3.5}$$

The iterative weights are then calculated, these are functions of the fitted values of $\hat{\mu}$, that is

$$w_i = \frac{p_i}{\left[b''(\theta_i) \left(\frac{\partial \eta_i}{\partial \mu_i} \right)^2 \right]} \tag{3.6}$$

where $b''(\theta_i)$ is the second derivative of $b(\theta_i)$ evaluated at the trial estimate. It is assumed that $a_i(\phi)$ has the form $\frac{\phi}{p_i}$. The weight is inversely proportional to the variance of the working dependent variable z_i given the current estimates of the parameters, with proportionality factor ϕ.

Then an improved estimate of β, namely the weighted least-squares estimates, degenerate the working dependent variable z_i on the predictors x_i using the weights w_i, obtained via

$$\hat{\beta} = \left(\mathbf{X}' \mathbf{W} \mathbf{X} \right)^{-1} \mathbf{X}' \mathbf{W} \mathbf{z}, \tag{3.7}$$

where $\mathbf{X}$ is the model matrix, $\mathbf{W}$ is a diagonal matrix of weights with entries w_i given by equation 3.6 and $\mathbf{z}$ is a response vector with entries z_i given by equation 3.5. The procedure is

repeated until convergence.

Similar to the residual sum of squares in linear regression, the goodness-of-fit of a generalized linear model can be assessed by the scaled deviance

$$D(y,\hat{\mu}) = 2\left[l(y,y) - l(\hat{\mu},y)\right],$$

where $l(y;y)$ is the log-likelihood achievable for an exact fit in which the fitted values are equal to the observed values, and $l(\hat{\mu};y)$ is the log-likelihood function calculated at the estimated parameters β. The deviance function is useful for comparing two models when one model is nested within the other. The deviance is additive for such nested models if the maximum likelihood estimates are used McCullagh and Nelder (1989).

3.2 Logistic Regression Models

Logistic regression modelling is a special type of generalized linear modelling used to fit a binary response variable as a function of a set of explanatory variables $X = (x_1, x_2, ..., x_p)'$ McCullagh and Nelder (1989). The set of explanatory variables can be discrete and/or continuous. If a particular observed outcome for the response variable Y is the desirable possible outcome, referred to as a "success", it is usually coded as $Y = 1$ and the contrary outcome, referred to as a "failure", coded as $Y = 0$, that is

$$y_i = \begin{cases} 1 & \text{if success,} \\ 0 & \text{otherwise} \end{cases}$$

as discussed by Hosmer and Lemeshow (2000). The logistic regression model is used to predict the odds of being successful based on the values of the independent variables (predictors). Logistic regression estimates the probability that a characteristic outcome is a success given the values of explanatory variables, expressed as

$$\pi = P(Y = 1|X = x).$$

The model is expressed as

$$\pi_i = P(Y_i = 1|X = x_i),\ i = 1,2,\ldots,p,$$

which implies that

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{i=1}^{p}\beta_i x_i,$$

therefore

$$\pi_i = \frac{\exp\left\{\beta_0 + \sum_{i=1}^{p}\beta_i x_i\right\}}{1+\exp\left\{\beta_0 + \sum_{i=1}^{p}\beta_i x_i\right\}} \text{ for } i = 1,2,...,p. \tag{3.8}$$

As demonstrated by Hosmer and Lemeshow (2000), the parameters for the model given in equation 3.8 are obtained via MLE as described in section 3.1 above. As per section 3.1, the likelihood function expresses the values of β in terms of known, fixed values for Y as

$$L(\beta) = \prod_{i=1}^{N}\frac{n_i!}{Y_i!\,(n_i - Y_i)!}\pi_i^{Y_i}(1-\pi_i)^{n_i - Y_i} \tag{3.9}$$

$$\propto \prod_{i=1}^{N}\left(\frac{\pi_i}{1-\pi_i}\right)^{Y_i}(1-\pi_i)^{n_i}$$

$$\propto \prod_{i=1}^{N}\exp(x_i\beta y_i)\left(1+\exp\left(x_i^T\beta\right)\right)^{-n_i}$$

For a binary response, there are $\binom{n_i}{Y_i}$ different ways to obtain Y_i successes from n_i trials. The probability of a success for any one of the n_i trials is π_i. As such the probability of observing $n_i - Y_i$ failures is $(1-\pi_i)^{n_i - Y_i}$ and the probability of observing Y_i successes is π_i^{yi}. Since the logarithm is a monotonic function, any maximum of the likelihood function will also be a maximum of the log

likelihood function. Taking the natural log of equation 3.9 yields the log likelihood function

$$l(\beta) = \ln[L(\beta)] = \sum_{i=1}^{N} x_i^T \beta y_i - \sum_{i=1}^{N} n_i \log\left(1 + \exp\left(x_i^T \beta\right)\right). \tag{3.10}$$

The first derivative of $x_i\beta$ with respect to β_j is x_{ij}, thus

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{N} y_i x_{ij} - \sum_{i=1}^{N} n_i \left(\frac{1}{1 + \exp\left(\boldsymbol{x}_i^T \beta\right)}\right) \exp\left(\boldsymbol{x}_i^T \beta\right) x_{ij} = \sum_{i=1}^{N} (y_i - \mu_i) x_{ij}, \tag{3.11}$$

where $E(y_i) = n_i \pi_i$. The maximum likelihood estimates for β can be found by setting each of the $j+1$ equations in equation 3.11 equal to zero and solving for each β_j.

The second derivatives are used in computing the standard errors of the parameter estimates, $\hat{\beta}$ are given by

$$\frac{\partial^2 l(\beta)}{\partial \beta_j \partial \beta_k} = -\sum_{i=1}^{N} n_i x_{ij} \frac{\partial}{\partial \beta_k} \left(\frac{\exp\left(x_i^T \beta\right)}{1 + \exp\left(x_i^T \beta\right)}\right) = -\sum_{i=1}^{N} n_i \pi_i (1 - \pi_i) x_{ij} x_{ik}$$

in McCullagh and Nelder (1989). The logistic regression in a GLM framework is discussed in section 3.3.

3.3 Logistic Regression in the GLM framework

As detailed in section 3.1 and expressed in equation 3.1, a density $f(y_i)$ belongs to the exponential family if it can be expressed as

$$f_y(y, \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} - c(y, \phi)\right\},$$

which can be expressed as

$$f_y(y_i|\omega_i) = \exp\left\{y_i \omega_i - b(\omega_i) + e(y_i)\right\},$$

where ω_i is called the natural parameter, $b(\omega_i)$ is a function whose form depends on the specific distribution and $e(y_i)$ is a function of the data that does not depend on the natural parameters of the model.

With reference to sections 3.1 and 3.2 it can be shown that a Bernoulli distribution is given by

$$\pi_i^{y_i}(1-\pi_i)^{1-y_i} = \exp\left\{y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \exp(1-\pi_i)\right\}$$

$$= \exp\{y_i\omega_i - b(\omega_i) + 0\},$$

where $\omega_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$, $b(\omega_i) = \exp(1+\exp(\omega_i))$ and $e(y_i) = 0$. It can be shown McCullagh and Nelder (1989), that the mean of an exponential family distribution is

$$E(y_i) = \frac{\partial b(\omega_i)}{\partial \omega_i} = b'(\omega_i).$$

The variance of an observation y_i, denoted as $\mathrm{Var}(\mu_i)$ indicating that it depends on the mean, is given by

$$\mathrm{Var}(y_i) = \frac{\partial^2 b(\omega_i)}{\partial \omega_i^2} = \mathrm{Var}(\mu_i).$$

In the GLM framework of McCullagh and Nelder (1989), the natural parameter ω_i is given by

$$\omega_i = b'^{-1}(\mu_i) = g(\mu_i) = x_i^T\beta,$$

where β is a vector of regression coefficients and x_i is the vector for the i^{th} observation. The function $g(.)$ is called the canonical link function because it relates the mean of the distribution through a transformation to the predictors. The inverse of $g(.)$, denoted as $b'(.)$, is called the response function and maps the linear predictor $\eta_i = x_i^T\beta$ onto the mean of the observations. Discussion of logistic regression as it relates to complex survey data will follow in section 3.4.

3.4 Logistic Regression Model for Complex Survey Data

The ordinary logistic regression approach detailed in section 3.2 assumes that the observations are independent and identically distributed. However, with practical survey data typically obtained using a complex sampling design that involves stratification, clustering, multistage sampling, or unequal probability, the assumption of independence between observations is often violated. Therefore it is necessary to adjust and modify the maximum likelihood method because the default parameter estimates could lead to an incorrect estimation of the standard errors and problems in the associated hypotheses tests Cassy et al. (2016). This modification of the maximum likelihood method is called the pseudo-maximum likelihood and is also known as weighted maximum likelihood method. The pseudo-maximum likelihood incorporates the sampling designs and the sampling weights in the estimation of β, as discussed by Hosmer and Lemeshow (2000), Lumley (2004), Lee and Forthofer (2006), Archera et al. (2007), Binder (1983), Skinner et al. (1989) and Chambers and Skinner (2003).

As per Archera et al. (2007), under simple random sampling elements are selected independently hence the covariance between the elements is zero, however under complex sampling there might be a number of primary sampling units (PSUs) implying that there are $j = 1,...,M$ clusters where PSUs are sampled and units within a cluster are often dependent. In addition, within each sampled PSU there are $i = 1,...,N_j$ units from which n_m are sampled. The pseudo-maximum likelihood of a binary observation is given by

$$\pi\left(x_{ji}\right)^{w_{ji}y_{ji}}\left[1-\pi\left(x_{ji}\right)^{w_{ji}\left(1-y_{ji}\right)}\right], \tag{3.12}$$

where $w_{ji} = \frac{1}{\pi_{ji}}$ is the sampling weight of each sampling unit and it is given by the inverse of its probability of inclusion in the sample and the subscript ji denotes the cluster observation number. The pseudo-maximum likelihood function for a design that includes clustering, for example, is constructed as the product of the individual contributions to the likelihood, but now it is the product over the m sampled clusters and n_m observations within the cluster and is expressed as

$$L_p(\beta) = \Pi_{j=1}^{M} \Pi_{i=1}^{N_j} \pi\left(x_{ji}\right)^{w_{ji}y_{ji}} \left[1 - \pi\left(x_{ji}\right)^{w_{ji}\left(1-y_{ji}\right)}\right]. \tag{3.13}$$

The pseudo-maximum likelihood estimator is the value that maximizes the pseudo-log-likelihood function, namly

$$ln\left\{L_p(\beta)\right\} = l(\beta) = \sum_{j=1}^{M} \sum_{i=1}^{N_j} \left[w_{ji}y_{ji}\right] ln\left[\pi\left(x_{ji}\right)\right] + \left[w_{ji}\left(1-y_{ji}\right)\right] ln\left[1 - \pi\left(x_{ji}\right)\right]. \tag{3.14}$$

The pseudo-maximum likelihood estimator of β is obtained by differentiating the pseudo-log-likelihood function given in equation 3.14, setting to zero and solving:

$$\frac{\partial}{\partial \beta} l(\beta) = 0.$$

This adjusted survey logistic regression model fitted in a GLM framework is applied in Chapter 7.

Chapter 4

Multilevel Modelling

4.1 Introduction

The multilevel models, also known as hierarchical models, are an extension of generalized linear models in which lower level units are nested within a hierarchy of successively higher-level units Goldstein (1991). For instance, individuals (level 1) can be nested within households (level 2) that are nested within communities (level 3). Multilevel models explain variations in random variables that vary between units at different levels of the hierarchy Goldstein (1991). Multilevel models are also designed to capture and account for the different-level induced sources of variability.

The multilevel modelling approach is commonly used to analyze clustered or grouped data from multistage complex sampling designs. Such sampling designs often use unequal probability of selection at each sampling stage as well as stratification and clustering, Rabe-Hesketh and Skrondal (2006). Complex surveys often employ these multistage sampling designs where clusters or primary sampling units (PSUs) are sampled in the first stage, subclusters in the second stage, elementary units are sampled in the final stage resulting in a multilevel dataset. Each stage corresponds to a level with elementary units at level 1 and progressively higher level units at higher levels.

With reference to Mason et al. (1984), Snijders and Bosker (1999), Goldstein (1991), Goldstein

(2011) suppose the data are observed at two levels, the individual (micro) level 1 and some higher (macro) level 2. The level 1 equation is expressed as

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{1ij} + \varepsilon_{ij}, \tag{4.1}$$

where $j = 1, ..., J$ denotes level 2 units and $i = 1, ..., n_j$ denotes level 1 observations within level 2 units. Equation 4.1 has fixed $\boldsymbol{X}$ values and full-column rank using the data supplied for each context, that is, for no context are the values of X_1 identical for all observations, and $n_j \geq 2$. Also $\varepsilon_{ij} \sim N\left(0, \sigma_j^2 \mathbf{I}_{n_j}\right)$ where ε_{ij} are random errors, where $i = 1, ..., n_j$, $j = 1, ..., J$, and σ_j^2 are all independently distributed.

In order to describe the macro part of the model, level 2, the following equations are considered,

$$\beta_{0j} = \beta_0 + u_{0j}, \tag{4.2}$$

$$\beta_{1j} = \beta_1 + u_{1j}, \tag{4.3}$$

where u_{0j} and u_{1j} are random variables with $E\left(u_{0j}\right) = E\left(u_{1j}\right) = 0$, $\text{Var}\left(u_{0j}\right) = \sigma_{u0}^2$ $\text{Var}\left(u_{1j}\right) = \sigma_{u1}^2$. Therefore $u_{0j} \sim N\left(0, \sigma_{u0}^2\right)$, $u_{1j} \sim N\left(0, \sigma_{u1}^2\right)$ and $\text{Cov}\left(u_{0j}, u_{1j}\right) = \sigma_{u01}$.

Each macro equation meets the rank condition using the data supplied for each context and the micro errors are independent of the macro errors, implying that for all i, j, and k, ε_{ij} is independent of u_{kj} for $k = 0, 1$.

A hybrid equation, or model, for the multilevel model is obtained by substituting equations 4.2 and 4.3 into equation 4.1 obtaining

$$Y_{ij} = \beta_0 + u_{0j} + \left(\beta_1 + u_{1j}\right) X_{1ij} + \varepsilon_{ij}$$

$$= \beta_0 + \beta_1 X_{1ij} + \left(u_{0j} + u_{1j} X_{1ij} + \varepsilon_{0ij}\right). \tag{4.4}$$

Goldstein (2011) further simplifies equation 4.4 by expressing the response variable Y_{ij} as the sum of a fixed part and a random part. The fixed part can be given in a matrix form as follows,

$$E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}_{ij}\boldsymbol{\beta} = (\boldsymbol{X}\boldsymbol{\beta})_{ij} \tag{4.5}$$

with $\boldsymbol{Y} = \left[Y_{ij}\right]$, $\boldsymbol{X} = \left[X_{ij}\right]$ as the matrix of the explanatory variables, $\boldsymbol{\beta}$ is the vector of the fixed part coefficients, parameters and the random variables in the model which are also known as the random effects.

The variances and covariances in equation 4.4, referred to as variance-covariance components are, σ^2_{u0}, σ^2_{u1}, σ_{u01} and σ^2_{e0}. For a simple level 2 model random parameters only include σ^2_{u0} and σ^2_{e0} since

$$Y_{ij} = \beta_0 + u_{0j} + \beta_1 X_{1ij} + \varepsilon_{0ij}.$$

Since $\text{Var}\left(\varepsilon_{0ij}\right) = \sigma^2_{e0}$, $\text{Var}\left(u_{0j}\right) = \sigma^2_{u0}$ the variance of the response about the fixed component is

$$\text{Var}\left(Y_{ij}|\beta_0, \beta_1, X_{ij}\right) = \text{Var}\left(u_0 + \varepsilon_{0ij}\right) = \sigma^2_{u0} + \sigma^2_{e0},$$

and the residuals are assumed to be mutually independent Goldstein (2011).

4.2 Parameter Estimation

There are several estimation techniques used in multilevel modelling. This is because the models comprise of different types of parameters, there are fixed effects, random effects or random coefficients and the variance-covariance components involved. Available methods include Maximum Likelihood (ML), Restricted Maximum Likelihood (REML), Iteratively Generalized Least Squares (IGLS) and various Bayesian methods. Bayesian estimation is usually used in simple multilevel modelling because the numerical integration for complex models is computationally intensive Ker (2014). The focus in this thesis will be on the other estimation methods.

As explained by Goldstein (2011), for the IGLS consider a level 2 multilevel model, as defined in section 4.1, given by

$$Y_{ij} = \beta_0 + \beta_1 X_{1ij} + \left(u_{0j} + u_{1j} X_{1ij} + \varepsilon_{0ij}\right).$$

Suppose that the values of the variances and covariances are known and contained in a matrix V. A Generalized Least Squares (GLS) estimation procedure can be used to obtain an estimator for the fixed,

$$\widehat{\beta} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$$

where

$$\mathbf{V} = \mathbf{X}\mathbf{G}\mathbf{X}' + \mathbf{R}.$$

The elements of G and R are called the variance-covariance components, namely σ^2_{u0}, σ^2_{u1}, σ_{u01} and σ^2_{e0}. These components are estimated by ML or REML. The variance of the estimator of $\hat{\beta}$ is estimated by

$$\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right) = \left(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}.$$

The REML method is preferred because it produces less biased estimates for the variance components part in case of small sample sizes Snijders and Bosker (1999). ML is obtained by maximizing the likelihood of the full dataset, whereas REML maximizes the likelihood of the residuals. The disadvantage of ML over REML is that the ML estimators of the variance components do not correct the degrees of freedom lost due to the variance components being too small. The REML estimates take into account the degrees of freedom used in estimating the fixed effects when estimating the covariance component which is especially useful if the number of the level 2 units is small, Hox (2002).

Iterative numerical procedures, like IGLS, are used to obtain the random estimates. Usually

these procedures are based on maximum likelihood estimation techniques. The IGLS procedure starts with an ordinary least squares (OLS) estimate of the fixed parameters which are then used to estimate the random part of the model. Then the estimate of the covariance matrix of the model is used to make an improved estimate of the fixed part. The IGLS method alternates iteratively between estimating the fixed and the variance components of the model until convergence. Goldstein (2011) proved that under normality of all the error terms, the parameter estimates resulting from IGLS procedure are equivalent to ML estimates. Therefore either of these methods can be used for parameter estimation in this research project. Maximum likelihood estimates of G and R are found by maximizing the log-likelihood function given by

$$l(G,R) = -\frac{1}{2}\log|V| - \frac{N}{2}\log r'V^{-1}r - \frac{N}{2}\left(1 + log\frac{2\pi}{N}\right),$$

where

$$r = Y - X\left(X'V^{-1}X\right)'X'V^{-1}Y,$$

and $N = \sum_{j=1}^{J} n_j$ referring to section 4.1, as discussed by Sullivan et al. (1999).

If the number of level 2 units, J, is large then the estimates generated through maximum likelihood are approximately equal to estimates generated through REML. The REML estimates of the covariance components are based on residuals which are computed after estimating the fixed effects that can be estimated using weighted least squares (WLS) or generalized least squares (GLS).

The REML estimates of G and R are found by maximizing the log-likelihood function

$$l(G,R) = -\frac{1}{2}\log|V| - \frac{1}{2}\log|X'V^{-1}X| - \frac{(N-p)}{2}\log r'V^{-1}r - \frac{(N-p)}{2}\left(1 + \log\frac{2\pi}{(N-p)}\right)$$

where $r = Y - X\left(X'V^{-1}X\right)'X'V^{-1}Y$ and $p = \text{rank}(X)$, Sullivan et al. (1999). Section 4.3 which follows, presents multilevel modelling with complex survey data.

4.3 Multilevel Modelling with Complex Survey Data

Complex sampling designs frequently incorporate unequal selection probabilities. Failing to account for this aspect of the design in the standard multilevel model can lead to biased parameter estimates. To correct this problem, design weights are incorporated in the likelihood function. To incorporate the complex sampling designs in the estimation method, the pseudo-maximum likelihood estimation method as outlined in Skinner et al. (1989) can be used. The level 2 version of the pseudo-maximum likelihood estimator, called the multilevel pseudo-maximum likelihood estimator (MPML), will be used in this study Rabe-Hesketh and Skrondal (2006). Pfeffermann et al. (1998) addressed the problem of weighting in the multilevel models using the probability weighted iterative generalized least squares method. The logistic regression in a MLM framework will be discussed in section 4.4 that follows.

4.4 Logistic Regression in the MLM framework

Multilevel modelling is applied to logistic regression in the same way as with linear regression Gelman and Hill (2007). Suppose Y_{ij} denotes the binary response of the level 1 unit i, $i = 1,...,n_j$ belonging to the level 2 unit j, $j = 1,...,J$ with explanatory variables denoted by X_{ij}, the vector with fixed effects is denoted by β and the vector with the random coefficients shared by all level 1 units belonging to the j^{th} level 2 unit by u_j. Y_{ij} is distributed as a Bernoulli random variable with probability π_{ij}. Then if $Y_{ij} \backsim \text{Bernoulli}\left(\pi_{ij}\right)$ then the logit link function is obtained given by

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_0 + \beta_1 X_{ij} + u_j \tag{4.6}$$

where π_{ij} denotes the probability that the i^{th} subject in the j^{th} level 2 cluster has the event, u_j is the random effect at level 2 model. Conditional on u_j, Y_{ij} are assumed to be independent and in the case of multilevel linear models, $u_j \sim N\left(0, \sigma_u^2\right)$ and equation 4.6 implies that the probability function is

$$\pi_{ij} = \frac{\exp\left(\beta_0 + \beta_1 X_{ij} + u_j\right)}{1 + \exp\left(\beta_0 + \beta_1 X_{ij} + u_j\right)},$$

as discussed by Wong and Mason (1985), Guo and Zhao (2000) and Rozi et al. (2017). The application of multilevel models is presented in Chapter 7.

Chapter 5

Mixed Effects Models

5.1 Introduction

Mixed effects models, also known as random effects models, are an extension of the corresponding regression models that have random effects to accommodate the between individual and within individual variation in the data, Wu (2010). The modification of the mixed effects model described by Laird and Ware (1982) for repeated-measures data are used in this research. A general linear regression model can be written as

$$y_i = \beta_0 + \beta_{1x_{i1}} + ... + \beta_p x_{ip} + \varepsilon_i$$

or equivalently

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{5.1}$$

for $i = 1, 2, ..., n$, where y_i is the response for individual i, $\beta_j's$ are unknown parameters, x_{ij} is the j^{th} covariate for individual i, $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})'$, $\beta = (\beta_0, ..., \beta_p)'$ and $\varepsilon_i's$ are random errors. The standard assumptions for equation 5.1 include that the errors, $\varepsilon_i's$ are independent, the errors $\varepsilon_i's$ having mean zero, constant variance σ^2 and are normally distributed. Therefore the distribution of $\boldsymbol{y}$ under the model assumptions is given by Wu (2010)

$$\boldsymbol{y} \sim N\left(X\beta, \sigma^2 I\right).$$

For complex survey data, the classical linear regression model in equation 5.1 may be inappropriate because the observations within each individual or cluster may be correlated, which leads to a violation of independence assumption. To incorporate correlation within individuals or clusters and the variation between individuals or clusters, equation 5.1 can be extended by introducing random effects in the model and thus obtain a linear mixed model.

Consider a longitudinal study. Let y_{ij} denote the response value for individual i at time t_j , $i = 1,2,...,n,\ j = 1,2,...,n_i$ and let $\boldsymbol{y}_i = (y_{i1}, y_{i2}, ..., y_{in_i})'$ denote the n_i repeated measurements of the response variable $\boldsymbol{y}$ on individual i, $i = 1,2,\cdots,n$. Thus a general linear mixed effects model can be written as

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{u}_i + \boldsymbol{\varepsilon}_i, \tag{5.2}$$

for $i = 1,2,...,n,\ \boldsymbol{u}_i \sim N(0,D)$ and $\varepsilon_i \sim N(0,\mathbf{R}_i)$ Wu (2010). In this context $\boldsymbol{\beta} = (\beta_1,...,\beta_p)'$ is a p×1 vector of fixed effects, $\boldsymbol{u}_i = \left(u_{i1},...,u_{iq}\right)'$ is a q×1 vector of random effects the n_i×p matrix $\boldsymbol{X}_i$ and the n_i×q matrix $\boldsymbol{Z}_i$ are known as the design matrices which may contain covariates, $\boldsymbol{e}_i = (e_{i1}, e_{i2}, ..., e_{in_i})'$ represents the random errors of the repeated measurements within individual i, $\boldsymbol{D}$ is a q×q covariance matrix of the random effects, and $\boldsymbol{R}_i$ is a $n_i \times n_i$ covariance matrix of the within-individual errors. It is often assumed that $\boldsymbol{R}_i = \sigma^2 I_{n_i}$ for simplicity, where $\boldsymbol{I}_{n_i}$ is the n_i×n_i identity matrix. The linear mixed model expressed in equation 5.2 specifically incorporates two sources of variability, the within individual variation and the between individual variation. This is similar to the two stage hierarchical model where stage one specifies the within individual variation, and stage two specifies the between individual variation. The model in equation 5.2 differs from the classical linear regression model in equation 5.1 by the term $\boldsymbol{Z}_i\boldsymbol{u}_i$, which links the random effects to the response.

5.2 Generalized Linear Mixed Models

The extension of the generalized linear models outlined in section 3.1 to include random effects results in a family of models called the generalized linear mixed model (GLMMs). GLMMs incorporate random effects into the linear predictor portion of a generalized linear model. This extension allows the accommodation of correlation in the context of a broad class of models for non-normally distributed data. As mentioned in section 3.1, the class of generalized linear models extends ordinary regression models in two ways, firstly by allowing non-normal responses and by allowing modelling a function of the mean rather than the mean itself. Generalized linear mixed models are a further extension that permits both fixed and random effects in the predictor rather than only fixed effects Agresti et al. (2000), McCulloch (2003) and McCulloch et al. (2008). In a GLMM it is assumed that correlation arises among repeated observations within a given individual or cluster because of the shared random effects, but these repeated observations are assumed to be conditionally independent given the random effects Wu (2010). The conditional independence assumption of the response variable, given the random effects, has an important impact in the formulation of the GLMM. GLMMs are typically constructed by incorporating random effects into the linear predictor of a conditionally independent exponential family model. The definition of a GLMM, as per McCulloch (2003), is

$$f_{Y_i|u}(y_i|u) = \frac{\exp\{[y_i\theta_i - b(\theta_i)]\}}{\phi^2} + c(y_i, \phi) \tag{5.3}$$

$$g(\mu_i) = x_i'\beta + z_i'u \tag{5.4}$$

where $E[Y_i|u] = \mu_i$ and u is distributed as $f_U(u)$. The definition of GLMM expressed in equation 5.3 resembles the usual components of a generalized linear model because firstly, the distribution of Y_i from an exponential family is assumed to hold conditional on the random effects u. Secondly the link function, $g(\cdot)$ is applied to the conditional mean of Y_i given u to obtain the conditional linear predictor. Finally the linear predictor is assumed to consist of two components, the fixed effects

portion, described by $x_i'\beta$ and the random effects portion, $z_i'u$, where a distribution is assigned to u. The model can be presented compactly in matrix notation as

$$g(\boldsymbol{\mu}) = \boldsymbol{X\beta} + \boldsymbol{Zu}$$

for $\boldsymbol{u} \sim N(0, \boldsymbol{D})$.

5.3 Parameter Estimation

Subject to the assumption of independence, the joint likelihood function of the fixed effects parameters, β, and the covariance parameter of the random effects, D, for a generalized linear mixed model is given as

$$L(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{D}) = \int L(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{D}, \boldsymbol{u}) f(\boldsymbol{u}) du, \tag{5.5}$$

where $\boldsymbol{Y} = \{y_{ij}\}$ is the vector of the response variable, with $i = 1, 2, ..., n$, and $j = 1, 2, ..., n_i$, $\boldsymbol{u} = (u_{1,} u_2, ..., u_{n_i})'$ is the vector of random effects and the integration is taken over a n_i dimensional space. Under the standard assumptions of GLMMs the distribution of the response variable, y_{ij}, given the random effect, u_j, belongs to a member of the exponential family of distributions. $f(\boldsymbol{u})$ is a specific probability density function which is often identically and independently distributed and converts the multivariate integral in equation 5.5 into a product of n_i univariate integrals McCulloch (1997).

With reference to McCulloch et al. (2008) the models described in equations 5.2, 5.3 and 5.5 can be expressed as

$$L(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{D}) = \log \int f_{Y_i|u}(y_i|u) f(u) du$$

$$= \log f_Y(y)$$

where the aim is to estimate the parameters $\boldsymbol{\beta}$ and $\boldsymbol{u}$. Estimates of these parameters are obtained by partially differentiating the likelihood function with respect to $\boldsymbol{\beta}$ to estimate the fixed effects parameters

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}} \int f_{Y_i|\boldsymbol{u}}(y_i|\boldsymbol{u}) f(\boldsymbol{u}) d\boldsymbol{u} / f_Y(y) \tag{5.6}$$

$$= \int \left[\frac{\partial}{\partial \boldsymbol{\beta}} f_{Y_i|\boldsymbol{u}}(y_i|\boldsymbol{u}) \right] f(\boldsymbol{u}) d\boldsymbol{u} / f_Y(y)$$

since $f(\boldsymbol{u}) du$ does not involve $\boldsymbol{\beta}$. Then

$$\frac{\partial}{\partial \boldsymbol{\beta}} f_{Y_i|\boldsymbol{u}}(y_i|\boldsymbol{u}) = \left(\frac{1}{f_{Y_i|\boldsymbol{u}}(y_i|u)} \frac{\partial f_{Y_i|\boldsymbol{u}}(y_i|u)}{\partial \boldsymbol{\beta}} \right) f_{Y_i|\boldsymbol{u}}(y_i|\boldsymbol{u})$$

$$= \frac{\partial \log f_{Y_i|\boldsymbol{u}}(y_i|\boldsymbol{u})}{\partial \boldsymbol{\beta}} f_{Y_i|\boldsymbol{u}}(y_i|\boldsymbol{u}).$$

As a result

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \int \frac{\partial \log f_{Y_i|\boldsymbol{u}}(y_i|\boldsymbol{u})}{\partial \boldsymbol{\beta}} f_{Y_i|\boldsymbol{u}}(y_i|\boldsymbol{u}) du / f_Y(y)$$

$$= \int \frac{\partial \log f_{Y_i|\boldsymbol{u}}(y_i|\boldsymbol{u})}{\partial \boldsymbol{\beta}} f_{U|y}(\boldsymbol{u}|y) du. \tag{5.7}$$

The random effects parameters are estimated using the maximum likelihood equation for the parameters in the distribution of $f_U(\boldsymbol{u})$. Let $\boldsymbol{\theta}$ denote the parameter in the distribution of $f_U(\boldsymbol{u})$ and then

$$\frac{\partial L}{\partial \theta} = \int \frac{\partial \log f_U(\boldsymbol{u})}{\partial \theta} f_{U|y}(\boldsymbol{u}|y) du$$

$$= E\left[\frac{\partial \log f_U(\boldsymbol{u})}{\partial \theta} | y \right].$$

In the next section we discuss how the mixed effects models can be fitted to complex survey data.

5.4 Mixed Effects Modelling with Complex Survey Data

The extension of the generalized linear model involves the addition of random effects and correlated errors. A pseudo-likelihood estimation procedure is developed to fit this class of mixed models based on an approximate marginal model for the mean response Wolfinger and O'Connell (1993). The pseudo-likelihood method estimates the model parameters of GLMMs using a linearization technique which employs a Taylor series expansions iteratively to approximate the initial generalized linear mixed model with a linear mixed model. Fitting the resulting linear mixed model is itself an iterative process which, upon convergence, leads to new parameter estimates that are then used to update the linearization. Pseudo-likelihood estimation is preferred because when the conditional distribution belongs to the exponential family the distributional assumption becomes explicit. The Pseudo-likelihood method conveys the idea that the approximating function has much of the structure of a Guassian log-likelihood, allowing the use of estimating equations which resembles the linear mixed models for covariance components and linear predictor effects, Stroup (2013). Wolfinger and O'Connell (1993) developed the pseudo-likelihood approach which is motivated from the perspective of a Laplace approximation. The process begins with a Taylor series expansion of the inverse link function evaluated at $\tilde{\eta}$, which follows from the objective of a GLMM, to model $E(y|u) = \boldsymbol{\mu}|\boldsymbol{u}$ by $h(\boldsymbol{X\beta} + \boldsymbol{Zu}) = h(\eta)$ as it is also expressed in equation 5.4.

The Taylor series expansion is

$$h(\eta) \cong h(\tilde{\eta}) + \frac{\partial h(\eta)}{\partial \eta}|_{\eta=\tilde{\eta}} (\eta - \tilde{\eta})$$

and $\boldsymbol{D}^{-1} = diag\left[\frac{\partial h(\eta)}{\eta}\right]$. The Taylor series expansion can be expressed as

$$h(\eta) \cong h(\tilde{\eta}) + \tilde{\boldsymbol{D}^{-1}} \left(\boldsymbol{X\beta} + \boldsymbol{Zu} - \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{Z}\tilde{\boldsymbol{u}}\right)$$

where $\tilde{\boldsymbol{D}}$ denoted $\boldsymbol{D}$ are evaluated at $\tilde{\eta} = \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{Z}\tilde{\boldsymbol{u}}$. When the terms are rearranged, it yields the expression

$$\tilde{\boldsymbol{D}}\left[h\left(\eta\right) - h\left(\tilde{\eta}\right)\right] + \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{Z}\tilde{\boldsymbol{u}} \cong \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}.$$

Thus the mixed model pseudo-variable y^* is

$$y^* = \tilde{\eta} + \tilde{\boldsymbol{D}^{-1}}\left[y - \left(\tilde{\mu}|\tilde{u}\right)\right].$$

Therefore

$$y^* = \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{Z}\tilde{\boldsymbol{u}} + \tilde{\boldsymbol{D}^{-1}}\left[y - h\left(\tilde{\eta}\right)\right].$$

The logistic regression for mixed effects models will be discussed in the section that follows.

5.5 Logistic Regression for Mixed Effects Models

Generalized linear mixed models are an extension of linear mixed models to allow response variables from different distributions, such as binary responses. Consider a clustered binary response y_{ij} taking only two possible values say 0 or 1, where $i = 1, ..., n$ and $j = 1, ..., n_i$. A simple logistic regression model with random intercept can be written as

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \boldsymbol{\beta}_{0i} + \boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 + \boldsymbol{b}_i + \boldsymbol{\beta}_1$$

where $\boldsymbol{b}_i \sim N\left(0, D\right)$, $\mu_{ij} = E\left(y_{ij}\right) = P\left(y_{ij} = 1\right)$ and $\boldsymbol{\beta}_{0i} = \boldsymbol{\beta}_0 + \boldsymbol{b}_i$, as reviewed by Wu (2010). A more general GLMM for binary clustered responses may be written as

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b}_i \tag{5.8}$$

as discussed in Rabe-Hesketh and Skrondal (2006). The inverse logit function is given by

$$\mu = E(\boldsymbol{Y}) = \frac{\exp\{\boldsymbol{X\beta} + \boldsymbol{Zb}\}}{1 + \exp\{\boldsymbol{X\beta} + \boldsymbol{Zb}\}}.$$

We apply a logistic regression model as a mixed effects model with simulated complex survey data on Chapter 7.

Chapter 6

Simulation of Data

6.1 Introduction to Data Simulations

Simulation can be defined as a method for using computer software to model the operation of "real-world" processes, systems, or events Davis et al. (2007). In studies involving simulation of this kind, the researcher develops a model of the phenomenon under investigation and then chooses an appropriate simulation method. The model is then run many times under various conditions to observe the outcomes. In this sense, simulations are seen as similar to virtual experiments although in a simulation process the researcher experiments with the model rather than the actual phenomenon Gilbert and Troitzsch (2005). Simulations can be described as a mathematical technique for conducting research on a computer to answer specific complex methodological and theoretical questions that cannot otherwise be estimated directly through classical methods.

The steps taken under a simulation study are arranged in chronological order to ensure that the simulation study achieves the aims of the researcher's investigation, Hallgren (2014). A set of assumptions about the nature and parameters of the dataset are specified. A dataset is generated under the pre-assumptions, then statistical analyses of interest are performed on this data set and the parameter estimates of interest from these analyses are attained with many newly generated datasets to obtain an empirical distribution of parameters or assumptions. Often the initial assumptions

specified are modified and the steps are repeated for datasets generated according to new parameters or assumptions. The distribution of parameter estimates is obtained from these simulated data sets and analyzed to answer the question of interest. Hallgren (2014) states that simulation studies are used because they allow researchers to answer specific questions about the data analysis, statistical power and best practices for obtaining accurate results in empirical research. For the current research the simulation procedure is used to obtain data in order to demonstrate the statistical methods discussed in preceding chapters.

6.2 Data for the Current Study

The data used in this research project are obtained from simulations to resemble real data often encountered in situations that arise in research studies for the human immunodeficiency virus (HIV) and acquired immune deficiency syndrome (AIDS) in a complex surveys framework. HIV is a virus that affects the immune system, the body's way of fighting off diseases. By killing or damaging cells of the body's immune system, HIV slowly destroys the body's ability to fight infections and certain cancers. AIDS is the disease that results from this breakdown of the immune system Duesberg (1989). The current study aims to explain the variation in the likelihood of an individual contracting HIV based on associated socio-economic, biological and behavioural risk factors. This involves fitting a model that explains how likely one is to be infected with HIV as a function of factors such as age, residential area, education, literacy, economic status, marital status, employment and gender.

Research indicates that a lack of socio-economic resources is associated with risky sexual behaviours, potentially increasing one's likelihood of contracting HIV Pellowski et al. (2013). These behaviours include transactional sex and power differences in sexual relationships, which can place an individual at risk of infection as highlighted by Pellowski et al. (2013). Studies have also shown gender disparities with regards to HIV infection in that women are at greater risk of HIV for reasons that range from socio-cultural to biological especially in sub-Saharan African countries Burgoyne

and Drummond (2008). Research shows that women have a limited role in sexual decision making and protection, because of a number of gender inequality and social norm issues. With the lack of sexual education and higher rates of transactional sex women become more susceptible to HIV as compared to men and biologically women are more exposed to HIV than men especially during sexual contact Burgoyne and Drummond (2008). Even though HIV is predominately prevalent in major urban areas, trends over the years suggest an increasing impact of the disease on rural communities particularly in sub-Saharan Africa Dyson (2003). Rural residents face unique challenges such as distance to health care centers, lack of health care facilities and health care providers with HIV/AIDS expertise, limited availability of supportive or ancillary services, stigma and discrimination and limited educational and economic infrastructure Schur et al. (2012).

Demographic values were generated for 20000 individuals (first level units variable) where the individuals are classified by their *Gender* where 1 represents a male and 0 a female all within *Residential Area* (second level units variable). The number of *Individuals* observed within *Residential Area* is selected at random with possible values 1 for suburbs, 2 for village and 3 for township. This study focuses on individuals with the *Ages* between 18 years and 50 years as the other variables focus specifically on adult's lifestyle and wellbeing. This study also considers the in the level of *Education* of an individual with 1 for primary phase, 2 for secondary phase and 3 for tertiary phase. Other binary variables of interest include *Employment* where 1 is employed and 0 represents non employed, as well as whether the family owns a home or not, *Own Home* in this case 0 represents does not own a home and 1 represents owns a home. The *Race* of an individual is considered is to see whether if has an impact on an individual's likelihood of getting HIV where 1 represents black, 2 for white, 3 for coloured and 4 for indian. Some of the risk factors of an individual's likelihood of getting HIV include the variable *Marital Status* where 1 represents single, 2 for married, 3 for divorced and 4 widowed. The *Wealth* variable with 1 being lower class, 2 is middle class and 3 represents the upper class.

6.3 Data Structure for Complex Surveys

Many survey data, including national surveys, have a hierarchical or clustered structure. As mentioned in the previous chapters, statistical models designed for analyzing data with hierarchical structure include mixed, hierarchical linear, random coefficient models. Burton et al. (2006) details the steps to simulating complex data. They indicate that it is important to include the level of dependence between simulations, the allowance for failures, the choice of random number generator, starting seeds and the software package that is used. The statistical software package must be able to handle the complexities involved in the proposed simulation study and have a reliable random number generator number generator. **R** version 3.6.2 was used to generate simulations for this study.

An important feature of complex survey data mentioned in the previous chapters is dependence of units which is due to clustering Pfeffermann (2011). Consequently the classical statistical methods such as the ordinary logistic regression become inappropriate, hence the models to be built require a structure that accounts for this complexity. As per the simplifications in Hosmer and Lemeshow (2000) and Lee and Forthofer (2006), suppose that a finite population $\boldsymbol{U} = \{1,2,...,N\}$ is divided into $\boldsymbol{C} = 1,...,M$ clusters, where each cluster is further divided into $i = 1,...,n_i$ PSUs (primary sampling units) each of which is found in $j = 1,...,n_{Ci}$ SSUs (secondary sampling units), each containing n_{cij} elements. As described in section 4.1, let $\boldsymbol{Y}_{ij}$ denote the binary response of the level 1 unit i, belonging to the level 2 unit j, with predictor variables denoted by $\boldsymbol{X}_{ij}$. The vector with fixed effects is denoted by $\boldsymbol{\beta}$ and the vector with the random coefficients shared by all level 1 units belonging to the j^{th} level 2 unit by $\boldsymbol{u}_j$.

The data are observed at two levels, the individual level 1 and the higher level 2. The clusters or levels mainly represent an existing population grouping. For instance, learners are in classes, classes are in schools, schools are in school districts and school districts are in provinces. When a model of this type is specified, inferences can be made on any of the levels that is school, class or district. An outcome is described for an individual student as a sum of the effects for the individual learner, for their class, for the school, for the district and for the provinces. Each of these effects

can often be regarded as one of a substitutable collection of effects for instance "all school-level effects" that are drawn from a distribution described by a variance component Pfeffermann (2011). In this example where learners are grouped into schools, the learners will be the level 1 units while schools will be the level 2 units and so forth.

6.3.1 Simulation for GLMM

As per section 5.2, GLMMs are an extension to the GLMs in which the linear predictor contains random effects in addition to the usual fixed effects. GLMMs can handle binary outcomes and they also extend linear mixed models to non-normal data just like GLMs.

Simulations to generate clustered data will be performed to obtain an outcome Y_{ij} and a predictor X_{ij}, where i denotes the cluster, and j denotes the subject within the cluster. Thus Y_{ij} is the outcome observed for subject j from cluster i. The binary independent variables in X_{ij} will be generated from a Bernoulli distribution. To generate the corresponding binary outcome variable Y_{ij}, first the probability of the outcome will be generated from the following logistic regression model

$$\text{logit}\,(\pi_i) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + u_i, \tag{6.1}$$

as expressed by Benedetti et al. (2014), where u_i is a random effect generated from a normal distribution with mean 0 and variance σ_u^2. By including u_i in the data generation step, correlation between observations in the same cluster is induced. The number of clusters, number of subjects per cluster, β_i, variance of the random effect, and proportion of subjects with the outcome will vary.

When simulating responses for a GLMM, values for the intercept and the regression coefficients, namely the fixed effects, variances, and the covariances of the random effects, must be assumed.

6.3.2 Simulation for Multilevel Modelling

The level 2 model for the continuous outcome variable Y_{ij} with explanatory variables X_{1ij}, X_{2ij}... X_{nij} on the level 1 will be examined. The random part of the model contains residual error terms at the level 2 and are given as $u_{0j} \sim N\left(0, \sigma^2_{u0}\right)$, $u_{1j} \sim N\left(0, \sigma^2_{u1}\right)$ and individual level 1 residuals $\varepsilon_{ij} \sim (0,1)$. The fixed part contains the $\beta_0, \beta_1, \beta_2$ coefficients. This model can be expressed as

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{1ij} + \beta_{2j} X_{2ij} + \varepsilon_{ij} \tag{6.2}$$

where $\beta_{0j} = \beta_0 + u_{0j}$ and $\beta_{1j} = \beta_1 + u_{1j}$, as reviewed by Goldstein (2011) and the model equation 6.2 is estimated using maximum likelihood.

6.4 Statistical Computations in R

Computations for the analysis for both simulations (that model fitting) are done using the statistical software **R** version 3.6.2. The **R** packages used for the statistical computations include the **lmerTest** package, **lme4** package that was developed by Douglas et al. (2014) and the **Matrix** package. The simulations used in the research utilize several **R** functions. In this study in particular, three functions for creating random numbers namely, *rnorm()*, *rbinom()* and *sample()*, will be used in order to generate values for random variables for the simulated data. An additional function for setting the randomization seed, *set.seed*, is useful for generating the same sets of random numbers each time a simulation study is run, allowing exact replications of results. The *lm()* command is used to create a simple regression model and returns an object with information about the fitted linear model. The *glm()* command is used as an extension of linear models designed to accommodate both normal and non-normal data and also obtains the Akaike information criterion. The Akaike information criterion (AIC) is important for comparisons of different models as well as an estimator of the relative quality of statistical models for a given set of data. A generalized linear model obtained from the *glm()* command produces model fit results through maximum likelihood estimation.

The **survey** package was used to provide functionalities in **R** for analyzing data from complex surveys. Generalized linear models, including the linear model in a survey framework, are estimated by *svyglm()*. This has almost the same arguments as *glm()*, the difference being that the data argument to *glm()* is replaced by a design argument in *svyglm()* to account for the sampling design. Model-based approaches to analytic objectives can make use of existing software procedures for fitting regression models. The important aspect of implementing these procedures is making sure that the design features have been carefully accounted for in the design matrices of the specified models West et al. (2018).

There are two commonly used **R** packages for fitting multilevel, hierarchical and mixed effects models which are **nlme** and **lme4**. For the current study all model fitting will be performed using the **lme4** package because numerous design-adjusted model evaluation tools have been developed to evaluate the fits of regression models based on complex sample survey data. In **R**, the function *lmer()* in the package **lme4** takes account of both the fixed and the random effects. The *lme()* function is defined in the form of the LMM where the outcome variable is expressed as a linear combination of the random and fixed effects. The only difference in treatment of fixed and random effects is that the random effects require information on the nesting structure for the parameter within which they vary. The setup of the structure of the model in the **R** function takes the form of a general linear regression model given, that is

$$\mathrm{M}^* = \mathrm{lmer}(Y \sim x_1 + x_2 + x_3 ... + (1|r_1) + (1|r_2) + (1|r_1/r_2), \mathrm{REML} = \mathrm{TRUE}, \mathrm{family}, \mathrm{data}). \quad (6.3)$$

In equation 6.3, M^* represents a mixed model frame, Y is the response variable and $x_1, x_2, x_3 ...$ are the fixed effects. Random effects are included in the explanatory variables as explained in Chapter 5 and shown in the **R** code presented in Appendix 8.2. The crossed random effects take the form $(1|r_1) + (1|r_2)$ and nested random effects take the form $(1|r_1/r_2)$. A specific subject intercept is accounted for by $(1|\mathrm{subject})$. The next argument is whether the mixed model will estimate the parameters using maximum likelihood (ML) or restricted maximum likelihood (REML). REML is

the default parameter estimation criterion for linear mixed models, REML = FALSE is set to ensure *lmer* uses ML estimates. However ML estimates are known to be biased, REML being usually less biased. REML estimates of variance components are generally preferred because REML assumes that the fixed effects structure is correct. ML is often used when comparing models with different fixed effects as it does not rely on the coefficients of the fixed effects Harville (1977). The last arguments are the data frame from which the variables come and setting the family if applicable. The **lme4** package includes deviance and REML estimated deviance values in the model fit statistics in addition to the AIC, Bayesian information criterion (BIC) and log likelihood. The **lme4** package does not include $p-$value for model coefficients. However, there are several methods that allow for one to obtain $p-$value for example installing *lmerTest* library as suggested by Bates (2006).

Chapter 7

Data Analysis

7.1 Introduction

This chapter presents the application of the models discussed in Chapters 3, 4 and 5 to explain the variation in the likelihood of an individual being infected with HIV using the associated socio-economic, biological and behavioural risk factors as explained in section 6.2. This is done by fitting the models discussed to the simulated data. Various illustrations of survey data analysis are presented in this chapter using **survey** package, **lme4** package and all found in **R** as detailed in section 6.4. The emphasis is on the demonstration of the analysis of data from complex surveys using model-based approach methods.

7.2 Distribution of the Variables

It is important to note that the distribution of the response variable, HIV status, can be examined in two ways, from the original data set, which effectively ignores the sampling design, and using the survey design object that is created. The results from the distribution of the response variable, show that about 18.71% of respondents are HIV positive. Taking the sampling design into account, a design-based estimate of the proportion of respondents who are HIV positive is 18.56% and a standard error of the proportion of 0.49%.

The respondent's sexual behaviour characteristics are assessed by considering two variables; the methods of contraception an individual uses and the number of sexual partners that an individual engages with. The South Africa Demographic and Health Survey (SADHS) 2016 explored that among women, the most popular method of contraceptive is hormonal contraception namely injectables and contraceptive pills, where 18% use 3 month injectables and 7% use the 2 month injectables and 7% use contraceptive pills. For male individuals it is common to use a barrier method like condoms (16%). For both genders less than 5% use long-acting contraceptive methods; 3% use implants, 1% use Intrauterine Device (IUDs) and less than 1% use vasectomy. This means that individuals are less likely to use a permanent contraception method. About 30% of the respondents have no sexual partner, more than 50% of respondents have one partner and about 15% of respondents have more that one partner. Hence it should be intriguing to investigate how the number of sexual partners has an impact on the likelihood of an individual being HIV positive. The distribution of the survey population shows that 16% are considered upper class, 18% being middle class and the largest class being lower class with 66%. In South Africa, urban households are more likely than non-urban households to fall into the upper wealth class, while non-urban households are more likely to fall into the lower wealth class. Wealth varies widely by province, residential area, race and many other factors. It will be interesting to see how one's wealth status affects their likelihood of being HIV positive.

7.3 The Inter-class Correlation Coefficient

As described by Koch (1982), the Inter-class correlation coefficient (ρ), is a measure of the strength of the statistical relationship or association between two numerical variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship. The Inter-class (or inter-cluster) correlation (ICC) under clustered survey sampling data or multilevel nested data is often used to measure correlation or homogeneity

between units within clusters, classes or nests and is often calculated together with the proportion of variation explained by random effects, as explained by Snijders and Bosker (1999), Guo and Zhao (2000) and Goldstein (2003). Under multilevel modelling, the ICC provides a measure of the variability ascribed to a particular level of the hierarchy. A non-zero ICC shows that observations that share the same hierarchy or nest are not independent. In particular the ICC is the correlation between two randomly selected units in one randomly selected group. Furthermore the ICC is the fraction of the variability that is due to groups or to nests. A special function of the ICC, particularly under multilevel modelling, is for justifying considering an hierarchical model instead of an ordinary linear or logistic regression model.

7.4 Data Analysis

The interpretation of a logistic regression model involves the use of odds ratios. An odds ratio (OR) is a quantitative measure of association between an exposure and an outcome. The OR represents the likelihood that an outcome will occur given a particular exposure of the variable of interest, compared to the likelihood of the outcome occurring in the absence of that exposure. When a logistic regression model is fitted, the regression coefficient is the estimated increase in the log odds of the outcome per unit increase in the value of the exposure for continuous predictor variables. In other words, the exponential function of the regression coefficient is the odds ratio associated with a one unit increase in the exposure Szumilas (2010). An odds ratio of greater than one implies an increasing effect of the predictor variable on the response variable. For a factor variable an odds ratio of greater than one indicates an increasing effect of the given factor level as compared to the reference level on the response variable.

With reference to section 3.2, for the current study we let Y denote a binary response variable denoting a respondent's HIV status where

$$Y_i = \begin{cases} 1 & \text{if the repondent is HIV positive,} \\ 0 & \text{if the repondent is HIV negative.} \end{cases}$$

When interpreting a logistic regression model, the concern is how the predictor variables predict the likelihood of a subject being in the success category (HIV positive) of the response variable. For this study a $p-\text{value} \leqslant \alpha = 0.05$ significance level is used as a cut-off value for assessing statistical significance of a given predictor variable.

7.5 Results for LM, GLM and SvyGLM

The results are shown in tables 7.1, 7.2 and 7.3 and the relevant **R** codes are shown in 8.2. Table 7.1 displays the results from an ordinary logistic regression model fitted using linear models (LM). Table 7.2 displays the results from the generalized logistic regression model fitted using a GLM. Table 7.3 displays the results from the regression model taking the survey design into consideration under a GLM framework.

7.5.1 Results of the Models: LM, GLM and SvyGLM

For the purpose of illustration, three variables are chosen for each model for interpretation, including variables which are not significant. In this case these include age, marital status and gender. The interpretations in this section will target these variables using the output centering on the odds ratios from the respective tables.

With reference to table 7.1 where an ordinary logistic regression was fitted to the data, with comparison to females the odds of an individual who is a male individual being HIV positive are 0.9904. This implies that it is less likely for a male individual than a female to be HIV positive holding the other variables constant. With reference to the individuals that are not employed, the odds of an employed person being positive is 0.9962. This implies that it is less likely for an employed individual to be HIV positive than for an unemployed individual. The results show that with reference to individuals who have more than one sexual partner, the odds of a person having one sexual partner being HIV positive are 0.9859. This implies that it is less likely for a person with one sexual partner to be HIV positive holding the effects of the other variables constant than a

Table 7.1: Parameter estimates, S.Es., $p-$values and odds ratios for an ordinary logistic regression model.

Coefficients	**Estimate**	**S.E.**	$p-$**value**	**OR**
Intercept	0.2203	0.0205	0.0000	1.2464
Age	−0.0004	0.0004	0.3535	0.9996
Gender: Male	−0.0096	0.0004	0.2253	0.9904
Employment: Yes	−0.0038	0.0086	0.6544	0.9962
Education: Secondary	−0.0052	0.0091	0.5686	0.9948
Education: Tertiary	−0.0124	0.0126	0.3253	0.9877
Wealth: Middle class	0.0076	0.0104	0.4645	1.0077
Wealth: Upper class	−0.0068	0.0108	0.5266	0.9932
Marital status: Married	0.0023	0.0084	0.7850	1.0023
Marital status: Divorced	0.0011	0.0205	0.9565	1.0011
Marital status: Widow	−0.0488	0.0319	0.1270	0.9524
Residential area: Village	0.0094	0.0095	0.3224	1.0094
Residential area: Township	0.0168	0.0096	0.0793	1.0169
Own home: Yes	−0.0043	0.0079	0.5853	0.9957
Contraception: Hormonal	0.0004	0.0093	0.9627	1.0004
Contraception: Barrier	0.0141	0.0115	0.2191	1.0142
Contraception: Emergency	−0.0161	0.0206	0.4352	0.9840
Contraception: Permanent	−0.0096	0.0152	0.5263	0.9904
Sexual partners: One	−0.0142	0.0088	0.1055	0.9859
Sexual partners: >1	−0.0036	0.0129	0.7816	0.9964
Race: White	−0.0179	0.0110	0.1041	0.9822
Race: Coloured	−0.0139	0.0109	0.2027	0.9862
Race: Indian	−0.0061	0.0111	0.5835	0.9939

Table 7.2: Parameter estimates, S.Es., $p-$values and odds ratios for an ordinary generalized logistic regression model.

Coefficients	**Estimate**	**S.E.**	$p-$**value**	**OR**
Intercept	-1.2540	0.1341	0.0000	0.2854
Age	-0.0025	0.0027	0.3520	0.9975
Gender: Male	-0.0630	0.0521	0.2265	0.9389
Employment: Yes	-0.0251	0.0561	0.6549	0.9752
Education: Secondary	-0.0338	0.0599	0.5726	0.9668
Education: Tertiary	-0.0824	0.0842	0.3274	0.9209
Wealth: Middle class	0.0495	0.0675	0.4638	1.0507
Wealth: Upper class	-0.0456	0.0718	0.5256	0.9554
Marital status: Married	0.0149	0.0552	0.7868	1.0150
Marital status: Divorced	0.0076	0.1350	0.9552	1.0076
Marital status: Widow	-0.3632	0.2375	0.1262	0.6954
Residential area: Village	0.0632	0.0633	0.3185	1.065
Residential area: Township	0.0076	0.1350	0.0788	1.0076
Own home: Yes	-0.0284	0.0521	0.5860	0.9719
Contraception: Hormonal	0.0027	0.0610	0.9643	1.0027
Contraception: Barrier	0.0904	0.0741	0.2223	1.0946
Contraception: Emergency	-0.1102	0.1403	0.4324	0.8956
Contraception: Permanent	-0.0650	0.1022	0.5244	0.9371
Sexual partners: One	-0.0932	0.0574	0.1046	0.9110
Sexual partners: >1	-0.0230	0.0841	0.7845	0.9773
Race: White	-0.1178	0.0724	0.1039	0.8889
Race: Coloured	-0.0908	0.0716	0.2044	0.9132
Race: Indian	-0.0390	0.0721	0.5882	0.9617

Table 7.3: Parameter estimates, S.Es., $p-$values and odds ratios for a survey generalized logistic regression model.

Coefficients	**Estimate**	**S.E.**	$p-$**value**	**OR**
Intercept	-1.0993	0.1699	0.0000	0.3331
Age	-0.0019	0.0033	0.5730	0.9981
Gender: Male	0.0159	0.0621	0.7977	1.0160
Employment: Yes	-0.0046	0.0694	0.9470	0.9954
Education: Secondary	-0.0740	0.0734	0.3131	0.9286
Education: Tertiary	-0.1554	0.1062	0.1435	0.8561
Wealth: Middle class	0.0551	0.0846	0.5147	1.0566
Wealth: Upper class	-0.1370	0.0897	0.1268	0.8719
Marital Status: Married	-0.0351	0.0683	0.6075	0.9655
Marital Status: Divorced	-0.1349	0.1661	0.4164	0.8738
Marital Status: Widow	-0.5544	0.2786	0.0467	0.5744
Residential Area: Village	-0.0248	0.0797	0.7555	0.9755
Residential Area: Township	0.0102	0.0796	0.8980	1.0102
Own Home:Yes	-0.0523	0.0618	0.3978	0.9490
Contraception: Hormonal	-0.1246	0.0766	0.1041	0.8828
Contraception: Barrier	-0.0069	0.0902	0.9393	0.9931
Contraception: Emergency	-0.2265	0.1881	0.2285	0.7973
Contraception: Permanent	-0.1023	0.1257	0.4159	0.9027
Sexual partners: One	-0.1802	0.0721	0.0125	0.8351
Sexual partners: > 1	-0.1143	0.10662	0.2837	0.8919
Race: White	-0.1188	0.0892	0.1831	0.8879
Race: Coloured	0.0117	0.0903	0.8965	1.0118
Race: Indian	-0.1022	0.0881	0.2460	0.9028

person with multiple sexual partners.

With reference to table 7.2 where an ordinary generalized logistic regression model was fitted to the data, when age increases by one unit, the odds of an individual being HIV positive are 0.9975, holding the other variables constant. With reference to an individual who resides in an urban area, the odds of a person residing in a township being HIV positive is 1.0076 times higher. This implies that it is more likely for a person residing in the townships to be HIV positive than a person who resides in an urban area holding the effects of the other variables constant.

With reference to the individuals who use no method of contraception, the odds of a person using a barrier method being HIV positive is 1.0946 times higher. This implies that individuals who use a barrier method as a method of contraception is more likely to be HIV positive when holding other variables constant.

With reference to table 7.3 where a survey generalized logistic regression model was fitted to the data, it is shown that the odds of an individuals owning their homes being HIV positive are 0.9490 times higher, meaning individuals owning a home are less likely to be HIV positive in relation to individuals who do not own homes, holding the other variables constant. With reference to a black individual, a white individual is less likely to be HIV positive with odds ratio of 0.8879 holding the effects of the other variables constant. Single unmarried individuals might possibly have a higher likelihood of having multiple sexual partners as compared to those who are married and are in more stable sexual relationships. Lastly, with reference to single individuals, the odds of a married person being HIV positive are 0.9655. This implies that it is less likely for a married person to be HIV positive than a single person holding the effects of the other variables constant.

7.6 Results for Logistic Mixed Models (LMMs)

7.6.1 The Null Intercept only Linear Mixed Model

Consider the results presented in table 7.4. The random effects that the residual variance on the *Residential Area* level is 0.000 and residual variance on the first level is 0.156. This means that

Table 7.4: Random effects for the null intercept only logistic mixed model.

Random effects			
Groups	**Name**	**Variance**	**Standard deviation**
Residential area	(Intercept)	0.000	0.000
Residual		0.156	0.395
Total variance		0.156	

Table 7.5: Fixed effects for the null intercept only logistic mixed model.

Fixed effects	**Estimate**	**Standard Error**	$t-$**value**	$p-$**value**
(Intercept)	0.1934	0.0028	69.25	0.0000

there is no slope variation for the residential areas. Under fixed effects the estimate of the intercept is 0.1934, in table 7.5, and the intercept is significant since $p-\text{value} < 0.05$.

7.6.2 Logistic Mixed Modelling with First Level Predictors

We now consider a mixed model with the first level (individual level) predictors which are *Age, Gender, Employment, Education, Wealth, Marital status, Own home, Methods of contraception, Number of sexual partners and Race* as fixed effects. Tables 7.6 and 7.7 present the results for the linear mixed model with first level predictors. In this case, the aim is to control the effects of *Residential Area* using the urban, village and township as the categories. The interest is not in the effect of each *Residential Area* on the *HIV status*, however the *HIV status* from within the *Residential Areas* might be correlated hence the control for that is important. The *HIV status* is an attempt to explain part of the variation in *HIV status* through fitting the rest of the explanatory variable as fixed effects. The response variable has some residual variation which is unexplained and associated with the *Residential Areas* and therefore by using random effects, that unexplained variation is modelled through the variance. In table 7.6 the variance for *Residential Area* $= 0.012$, the total variance is 0.1668 and hence $(0.012/0.1668 * 100) = 7.19\%$, meaning the differences between *Residential Areas* explain about 7% of the variance that is left over after the variance explained by the fixed effects.

From the results in table 7.7, the $p-$values column, which indicate all regression coefficients

Table 7.6: Random effects for a logistic mixed model with first level predictors.

Random effects			
Groups	**Name**	**Variance**	**Standard deviation**
Residential area	(Intercept)	0.012	0.1095
Residual		0.1548	0.3935
Total variance		0.1668	

Table 7.7: Fixed effects for a logistic mixed model with first level predictors.

Fixed effects	**Estimate**	**S.E.**	$t-$**value**	$p-$**value**	**OR**
(Intercept)	0.1951	0.0144	13.550	0.0000	1.2154
Age	−0.0001	−0.0003	−0.517	0.6049	0.9999
Gender: Male	−0.0055	0.0057	−0.981	0.3268	0.9945
Employment:Yes	−0.0078	0.0061	−1.281	0.2000	0.9922
Education: Secondary	0.0021	0.0065	0.330	0.7415	1.0021
Education: Tertiary	−0.0066	0.0088	−0.753	0.4516	0.9934
Wealth: Middle class	0.0015	0.0074	0.210	0.8340	1.0015
Wealth: Upper class	0.0065	0.0077	0.838	0.4022	1.0065
Marital status: Married	0.0189	0.0059	3.157	0.0016	1.0191
Marital status: Divorced	−0.0127	0.0154	−0.824	0.4098	0.9874
Marital status: Widow	−0.0099	0.0229	−0.433	0.6650	0.9901
Own home:Yes	0.0097	0.0057	1.713	0.0866	1.0097
Contraception: Hormonal	−0.0002	0.0066	−0.034	0.9730	0.9998
Contraception: Barrier	−0.0032	0.0081	−0.402	0.6873	0.9968
Contraception: Emergency	−0.0064	0.0147	−0.438	0.6616	0.9936
Contraception: Permanent	−0.0298	0.0110	−2.712	0.0066	0.9706
Sexual partners: One	0.0024	0.0063	0.383	0.7019	1.0024
Sexual partners: >1	0.0022	0.0092	0.247	0.8050	1.0022
Race: White	−0.0014	0.0079	−0.173	0.8622	0.9986
Race: Coloured	−0.0012	0.0079	−0.160	0.8727	0.9988
Race: Indian	0.0021	0.0079	0.262	0.7930	1.0021

Table 7.8: Random effects for a logistic mixed model with first and second level predictors.

Random effects			
Groups	**Name**	**Variance**	**Standard deviation**
Residential area	(Intercept)	0.012	0.1095
Residual		0.1547	0.3933
Total variance		0.1667	

are mostly not significant except for *Marital status: Married* which is a significant predictor of *HIV status.* With reference to individuals who are single, the chance of being HIV positive when an individual is married is 1.0095 times higher. This implies that individuals who are married are more likely to be HIV positive when holding other variables constant. When age increases by one year, the odds of an individual being HIV positive are 0.9999 times, holding the other variables constant. This means that as the age increases it is less likely for an individual to be HIV positive.

7.6.3 Logistic Mixed Modelling with First and Second Level Predictors

In addition to the first level variables, a predictor variable on the second level which is *Provinces* is added. With reference to table 7.8, the variance for *Residential Areas* $= 0.012$, the total variance is 0.1667 and therefore $(0.012/0.1667 * 100) = 7.19\%$ meaning the differences between *Residential Areas* explain about 7% of the variance that is left over after the variance explained by the fixed effects. This is the same as logistic mixed model with first level predictors even though there is an addition of a second level predictor.

The results in table 7.9 show that only *Marital status: Married* and *Contraception: Permanent* from the level 1 variable are significant. The level 2 variable *Province: Free State* is also significant, which implies that an individual's marital status, method of contraception they use and the province that they reside in is significantly related to the their *HIV status*. The odds of a married individual being HIV positive with reference to a single individual is 1.0191.With reference to an individual who resides in the Eastern Cape province, the odds of a person residing in Free State being HIV positive is 1.0292 times higher. This implies that it is more likely for a person residing in the Free State to be HIV positive than a person who resides in the Eastern Cape holding the effects of

Table 7.9: Fixed effects for a logistic mixed model with first and second level predictors.

Fixed effects	**Estimate**	**S.E.**	$t-$**value**	$p-$**value**	**OR**
(Intercept)	0.1902	0.01639	11.604	0.0000	1.2095
Age	−0.0001	−0.0003	−0.505	0.6134	0.9999
Gender: Male	−0.0054	0.0057	−0.958	0.3379	0.9946
Employment: Yes	−0.0082	0.0061	−1.343	0.1794	0.9918
Education: Secondary	0.0026	0.0065	0.400	0.6891	1.0026
Education: Tertiary	−0.0069	0.0088	−0.784	0.4331	0.9931
Wealth: Middle class	0.0018	0.0074	0.249	0.8037	1.0018
Wealth: Upper class	0.0062	0.0077	0.808	0.4190	1.0062
Marital status: Married	0.0189	0.0056	3.152	0.0016	1.0191
Marital status: Divorced	−0.0124	0.0154	−0.806	0.4200	0.9877
Marital status: Widow	−0.0107	0.0229	−0.470	0.6384	0.9893
Own home: Yes	0.0095	0.0057	1.690	0.0911	1.0095
Contraception: Hormonal	−0.0005	0.0066	−0.076	0.9397	0.9995
Contraception: Barrier	−0.0029	0.0081	−0.367	0.7134	0.9971
Contraception: Emergency	−0.0054	0.0147	−0.367	0.7135	0.9946
Contraception: Permanent	−0.0298	0.0109	−2.714	0.0066	0.9706
Sexual partners: One	0.0025	0.0063	0.402	0.6875	1.0025
Sexual partners: >1	0.0021	0.0092	0.225	0.8220	1.0021
Race: White	−0.0012	0.0079	−0.158	0.8748	0.9988
Race: Coloured	−0.0013	0.0079	−0.169	0.8661	0.9987
Race: Indian	0.0021	0.0079	0.271	0.7861	1.0021
Province: Free State	0.0288	0.0119	2.415	0.0157	1.0292
Province: Gauteng	0.0019	0.0119	0.159	0.8733	1.0019
Province: Kwazulu Natal	−0.0060	0.01182	−0.510	0.6104	0.9940
Province: Limpopo	0.0232	0.0119	1.954	0.0506	1.0235
Province: Mpumalanga	−0.0156	0.0118	−1.315	0.1884	0.9845
Province: Northern Cape	0.0120	0.0117	1.026	0.3049	1.0120
Province: North West	−0.0032	0.0118	−0.255	0.7986	0.9968
Province: Western Cape	0.0034	0.0118	0.295	0.7682	1.0034

the other variables constant. The chances of individuals who use a permanent contraception with comparison to those who uses no contraception is 0.9706. This implies that it is less likely for an individual using a permanent contraceptive to be HIV positive than a person using no contraceptive, holding the effects of the other variables constant.

7.6.4 Multilevel Logistic Modelling (MLM) with First and Second Level Predictors and Nested Random Effects

A multilevel model is defined in much the same way as the linear mixed first level model function, where the outcome variable is the sum or linear combination of all of the random and fixed effects. The only difference is the fact that the random effects require information on the nesting structure for the parameter within which they vary. For this section, the random effect specifies the nested effect of *Residential Area* within or under *Province*, as *Residential Area* would be considered the first level variable and *Province* being the second level variable.

Tables 7.10 and 7.11 displays the results from a multilevel logistic model with first and second level predictors and nested random effects. In particular, the results in table 7.10 provides estimates for the random effects in the form of variances and standard deviations. Notice that there are three values shown, the nested effect of *Residential Area* within *Province*, the random effect of the higher level variable, *Province* and the residual term which represents error. The variance estimates are of interest here, they can be added together to find the total variance of the random effects which is 0.1577 in this case. Then divide each random effect by that total to see what proportion of the random effect variance is attributable to each random effect which indicates whether or not this effect is significant. Dividing the nested effect variance by the total variance provides the proportion of variance accounted, namely $(0.002/0.1577) * 100 = 1.27\%$. This implies that only 1.27% of the total variance of the random effects is attributed to the nested effect. The effect of *Province* alone, $(0.001/0.1577) * 100 = 0.63\%$ is very small.

From table 7.11 it is evident that *Marital status: Married* is the only statistically significantly related to the *HIV status.* With reference to individuals who are single, the chance of being HIV

Table 7.10: Random effects for a multilevel logistic model with first and second level predictors and nested random effects.

Random effects			
Groups	**Name**	**Variance**	**Standard deviation**
Residential area:Province	(Intercept)	0.002	0.0447
Province	(Intercept)	0.001	0.0316
Residual		0.1547	0.3948
Total variance		0.1577	

Table 7.11: Fixed effects for a multilevel logistic model with first and second level predictors and nested random effects.

Fixed effects	**Estimate**	**S.E.**	$t-$**value**	$p-$**value**	**OR**
(Intercept)	0.1902	0.0210	9.020	1.0000	1.209
Age	−0.0001	0.0003	−0.505	0.6134	0.9999
Gender: Male	−0.0054	0.0057	−0.958	0.3379	0.9946
Employment:Yes	−0.0082	0.0061	−1.343	0.1794	0.9918
Education: Secondary	0.0025	0.0065	0.400	0.6891	1.0025
Education: Tertiary	−0.0069	0.0088	−0.784	0.4331	0.9931
Wealth: Middle class	0.0018	0.0074	0.249	0.8037	1.0018
Wealth: Upper class	0.0062	0.0077	0.808	0.4190	1.0062
Marital status: Married	0.0189	0.0059	3.152	0.0016	1.0191
Marital status: Divorced	−0.0124	0.0154	−0.806	0.4200	0.9877
Marital status: Widow	−0.0108	0.0229	−0.470	0.6384	0.9892
Own home: Yes	0.0096	0.0057	1.690	0.0911	1.0096
Contraception: Hormonal	−0.0005	0.0066	−0.076	0.9396	0.9995
Contraception: Barrier	−0.0029	0.0081	−0.367	0.7134	0.9971
Contraception: Emergency	−0.0054	0.0147	−0.367	0.7135	0.9946
Contraception: Permanent	−0.0298	0.0109	−2.714	0.0066	0.9706
Sexual partners: One	0.0025	0.0062	0.402	0.6875	1.0025
Sexual partners: > 1	0.0020	0.0091	0.225	0.8220	1.0020
Race: White	−0.0012	0.0079	−0.158	0.8748	0.9988
Race: Coloured	−0.0013	0.0078	−0.169	0.8661	0.9987
Race: Indian	0.0021	0.0079	0.271	0.7861	1.0021
Province: Free State	0.0288	0.0223	1.297	1.0000	1.0292
Province: Gauteng	0.0019	0.0222	0.086	1.0000	1.0019
Province: Kwazulu Natal	−0.0060	0.0221	−0.272	1.0000	0.9940
Province: Limpopo	0.0321	0.0222	1.046	1.0000	1.0326
Province: Mpumalanga	−0.0156	0.0222	−0.704	1.0000	0.9845
Province: Northern Cape	0.0120	0.0221	0.544	1.0000	1.0120
Province: North West	−0.0030	0.0221	−0.136	1.0000	0.9970
Province: Western Cape	0.0035	0.0221	0.158	1.0000	1.0035

Table 7.12: Random effects for a multilevel logistic model with first and second level predictors and random slopes.

Random effects			
Groups	**Name**	**Variance**	**Standard deviation**
Residential area	(Intercept)	0.0000	0.0000
	Sexual partners: One	0.0000	0.0005
	Sexual partners: >1	0.0000	0.0015
Residual		0.1560	0.3949
Total variance		0.1560	

positive when an individual is married is 1.0191 times higher, which is the same as per prior models. However it is interesting to look into the outcome of the other variables like *Age*, *Gender* and *Employment* to see how they relate to one's HIV status. In comparison to a female individual, the results shows that the chances of a male individual to be HIV positive are 0.9946. This implies that it is less likely for male individual than a female individual to be HIV positive. With reference to the individuals that are not employed, the odds of an employed person being positive is 0.9918. This implies that it is less likely for an employed individual to be HIV positive than for an unemployed individual. When age increases by one year, it corresponds to a 0.0001 decrease to the outcome of an individual's *HIV status* being HIV positive with the odds ratio of 0.9999, meaning as a person grows older they will have less chances of being HIV positive, holding the effects of the other variables constant.

7.6.5 Multilevel Logistic Modelling with First and Second Level Predictors and Random Slopes

For a multilevel logistic model with first and second level predictors and random slopes, the random slopes are included in this model to first level predictor variables *Methods of contraception* and *Number of sexual partners* as random slopes.

Table 7.12 above shows that the the variance for the slope of the both of the variables *Sexual partners: One* and *Sexual partners: >1* are 0.0000. This means that there is no slope variation of the number of sexual partners variables between the residential areas.

Table 7.13: Fixed effects for a multilevel logistic model with first and second level predictors and random slopes.

Fixed effects	**Estimate**	**S.E.**	$t-$**value**	$p-$**value**	**OR**
(Intercept)	0.1902	0.1639	11.604	0.0000	1.2095
Contraception: Hormonal	−0.0005	0.0066	−0.076	0.9394	0.9995
Contraception: Barrier	−0.0029	0.0081	−0.367	0.7135	0.9971
Contraception: Emergency	−0.0054	0.0147	−0.368	0.7130	0.9946
Contraception: Permanent	−0.0298	0.0109	−2.712	0.0067	0.9706
Age	−0.0001	0.0003	−0.505	0.6136	0.9999
Gender: Male	−0.0054	0.0057	−0.958	0.3381	0.9946
Employment: Yes	−0.0082	0.0061	−1.343	0.1791	0.9918
Education: Secondary	0.0026	0.0065	0.399	0.6900	1.0026
Education: Tertiary	−0.0069	0.0088	−0.784	0.4328	0.9931
Wealth: Middle class	0.0018	0.0074	0.248	0.8039	1.0018
Wealth: Upper class	0.0062	0.0077	0.808	0.4188	1.0062
Marital status: Married	0.0189	0.0059	3.151	0.0016	1.0191
Marital status: Divorce	−0.0124	0.0154	−0.806	0.4202	0.9877
Marital status: Widow	−0.0107	0.0229	−0.469	0.6388	0.9893
Own home: Yes	0.0096	0.0057	1.690	0.0910	1.0096
Sexual partners: One	0.0025	0.0064	0.395	0.6930	1.0025
Sexual partners: >1	0.0021	0.0092	0.225	0.8219	1.0021
Race: White	−0.0012	0.0079	−0.157	0.8753	0.9988
Race: Coloured	−0.0013	0.0079	−0.168	0.8662	0.9987
Race: Indian	0.0021	0.0079	0.272	0.7853	1.0021
Province: Free State	0.0288	0.0119	2.414	0.0158	1.0292
Province: Gauteng	0.0019	0.0119	0.157	0.8750	1.0019
Province: Kwazulu Natal	−0.0060	0.0118	−0.511	0.6095	0.9940
Province: Limpopo	0.0232	0.0118	1.954	0.0508	1.0235
Province: Mpumalanga	−0.0156	0.0119	−1.316	0.1882	0.9845
Province: Northern Cape	0.0119	0.0117	1.024	0.3059	1.0119
Province: North West	−0.0030	0.0118	−0.256	0.7978	0.9970
Province: Western Cape	0.0035	0.0118	0.294	0.7688	1.0035

Table 7.14: Random effects for a multilevel logistic model with first and second level predictors with random slopes and cross-level interaction.

Random effects			
Groups	**Name**	**Variance**	**Standard deviation**
Residential area	(Intercept)	0.0000	0.0040
	Contraception: Hormonal	0.0000	0.0050
	Contraception: Barrier	0.0000	0.0148
	Contraception: Emergency	0.0000	0.0220
	Contraception: Permanent	0.0000	0.0008
Residual		0.1547	0.3933
Total variance		0.1547	

For multilevel logistic model with first and second level predictors and random slopes, it is shown in table 7.13 that when age increases by one year, it corresponds to a 0.0001 decrease to the outcome of an individual's *HIV status* being HIV positive with the odds ratio of 0.9999, meaning as a person grows older they will have less chances of being HIV positive. Looking at *Sexual partners*: >1, with reference to an individual who has no sexual partner, the chances of an individual being HIV positive when they have more than one partner are 1.0021, therefore this implies that individuals with more than one partner are more likely to being HIV positive, holding the effects of the other variables constant. With reference to an individual who resides in the Eastern Cape province, the odds of a person residing in the Gauteng province being HIV positive is 1.0019 times higher. This implies that it is more likely for a person residing in Gauteng to be HIV positive than a person who resides in the Eastern Cape, holding the effects of the other variables constant.

7.6.6 Multilevel Logistic Modelling with First and Second Level Predictors, with Random Slopes and Cross-level Interactions

The multilevel logistic model with first and second level predictors with random slopes and cross-level interaction is done by adding a cross-level interaction between *Methods of contraception* and *Province*. The aim is to investigate whether the differences in the relation between *Methods of contraception* and *HIV status* in a specific *Residential area* can be explained by which *Province* that individual is from.

Table 7.15: Fixed effects for a multilevel logistic model with first and second level predictors with random slopes and cross-level interaction.

Fixed effects	**Estimate**	**S.E.**	$t-$**value**	$p-$**value**	**OR**
(Intercept)	0.1985	0.0142	14.005	0.0000	1.2196
Age	−0.0001	0.0003	−0.471	0.6377	0.9999
Gender: Male	−0.0056	0.0057	−0.993	0.3208	0.9944
Employment: Yes	−0.0079	0.0061	−1.282	0.1999	0.9921
Education: Secondary	0.0021	0.0065	0.326	0.7445	1.0021
Education: Tertiary	−0.0069	0.0088	−0.782	0.4340	0.9931
Wealth: Middle class	0.0016	0.0074	0.213	0.8315	1.0016
Wealth: Upper class	0.0064	0.0077	0.825	0.4095	1.0064
Marital status: Married	0.0191	0.0059	3.178	0.0015	1.0193
Marital status: Divorced	−0.0121	0.0154	−0.788	0.4306	0.9879
Marital status: Widow	−0.0098	0.0229	−0.428	0.6689	0.9902
Own home: Yes	0.0098	0.0057	1.723	0.0848	1.0098
Sexual partners: One	0.0026	0.0063	0.407	0.6839	1.0026
Sexual partners: >1	0.0025	0.0092	0.268	0.7887	1.0025
Race: White	−0.0013	0.0079	−0.166	0.8684	0.9987
Race: Coloured	−0.0013	0.0079	−0.166	0.8684	0.9987
Race: Indian	0.0021	0.0079	0.261	0.7944	1.0021
Contraception:Province	−0.0014	0.0004	−3.426	0.0014	0.9986

From results in table 7.14, the explained slope variance *Methods of contraception* using the *Province* as second level variable is 0 for all the methods of contraception. Therefore 0% of the variance of the regression coefficients of the *Methods of contraception* slopes can be explained by the *Province.*

Looking at results in table 7.15, a married person will increase their chances of being HIV positive by 1.0193 point for every method of contraception they use. With reference to a wealth index of an individual coming from a lower class to one from a middle class the odds of being HIV positive increased are 1.0016 times higher. This implies that an individual who is in middle class is more likely to be HIV positive compared to an individual who is from a lower class wealth index. In comparison to females the odds of an individual who is a male individual being HIV positive are 0.9944. This implies that it is less likely for a male individual than a female to be HIV positive holding the other variables constant. This is in agreement with the results of the ordinary linear and GLM and survey logistic regression given in section 7.5.

Table 7.16: Random effects for a generalized logistic mixed model with first and second level predictor variables.

Random effects			
Groups	**Name**	**Variance**	**Standard deviation**
Residential area	(Intercept)	0.6744	0.8212
Residual		0.2174	0.4662
Total Variance		0.8918	

7.7 Results for Generalized Logistic Mixed Models (GLMMs)

We now consider the generalized linear mixed models (GLMMs), where the GLMMs serve as an extension to linear mixed models in order to allow the specification of the residual distribution and link function, but also allows the inclusion of random effects as explained in section 5.2. The difference between logistic models and multilevel logistic models is that a link function must be added under the multilevel model.

7.7.1 Generalized Logistic Mixed Models with First and Second Level Predictor Variables

The interpretation of GLMMs is similar to GLMs; however, there is an added complexity because of the random effects. The mixed effects logistic model is estimated to predict the *HIV status* of an individual $(\text{HIVpositive} = 1, \text{HIVnegative} = 0)$ from *Age, Gender, Employment, Education, Wealth, Marital status, Methods of Contraception, Number of sexual partners* and so on, where the intercept is allowed to vary randomly by *Residential Area.* The results presented in tables 7.16 and 7.17 are from a generalized logistic mixed model with first and second level predictor variables.

Table 7.16 provides estimates for the random effects in the form of variances and standard deviations. Compared to the total variance $(0.6744/0.8918) * 100 = 75.62\%$ is the effect of *Residential Area* alone and it is quite substantial.

The estimates in table 7.17 show that when age increases by one year, it corresponds to a -0.0009 decrease in the outcome of an individual's HIV status being HIV positive with the odds ratio of 0.9991, meaning as a person grows older they will have less chances of being HIV posi-

Table 7.17: Fixed effects for a generalized logistic mixed model with first and second level predictor variables.

Fixed effects	**Estimate**	**S.E.**	$t-$**value**	$p-$**value**	**OR**
(Intercept)	−1.4510	0.1062	−13.662	0.0000	0.2343
Age	−0.0009	0.0019	−0.503	0.6148	0.9991
Gender: Male	−0.0350	0.0366	−0.956	0.3389	0.9656
Employment: Yes	−0.0529	0.0395	−1.340	0.1803	0.9485
Education: Secondary	0.0167	0.0418	0.399	0.6896	1.0168
Education: Tertiary	−0.0451	0.0575	−0.785	0.4326	0.9559
Wealth: Middle class	0.0120	0.0479	0.251	0.8020	1.0121
Wealth: Upper class	0.0400	0.0495	0.808	0.4189	1.0408
Marital status: Married	0.1205	0.0383	3.146	0.0016	1.1281
Marital status: Divorce	−0.0849	0.1035	−0.820	0.4119	0.9186
Marital status: Widow	−0.0731	0.1533	−0.477	0.6333	0.9295
Own home: Yes	0.0618	0.0366	1.689	0.0913	1.0637
Contraception: Hormonal	−0.0033	0.0424	−0.077	0.9383	0.9967
Contraception: Barrier	−0.0193	0.0527	−0.367	0.7134	0.9809
Contraception: Emergency	−0.0349	0.0959	−0.364	0.7155	0.9657
Contraception: Permanent	−0.2033	0.0747	−2.723	0.0065	0.8160
Sexual partners: One	0.0165	0.0408	0.404	0.6864	1.0166
Sexual partners: >1	0.0133	0.0594	0.223	0.8233	1.0138
Race: White	−0.0079	0.0512	−0.155	0.8772	0.9921
Race: Coloured	−0.0085	0.0511	−0.167	0.8672	0.9915
Race: Indian	0.0141	0.0512	0.275	0.7831	1.0142
Province: Free State	0.1802	0.0759	2.372	0.0177	1.1974
Province: Gauteng	0.0125	0.0778	0.161	0.8722	1.0126
Province: Kwazulu Natal	−0.0401	0.0776	−0.516	0.6056	0.9607
Province: Limpopo	0.1463	0.0759	1.928	0.0538	1.1575
Province: Mpumalanga	−0.1064	0.0789	−1.349	0.1775	0.8991
Province: Northern Cape	0.0775	0.0756	1.025	0.3055	1.0806
Province: North West	−0.0201	0.0775	−0.259	0.7955	0.9801
Province: Western Cape	0.0229	0.0772	0.297	0.7666	1.0232

Table 7.18: Random effects for a generalized logistic mixed model with first and second level predictors and nested random effects.

Random effects			
Groups	**Name**	**Variance**	**Standard deviation**
Residential area:Province	(Intercept)	0.0040	0.0638
Province	(Intercept)	0.0012	0.0035
Total variance		0.0052	

tive holding the other variables constant. Considering *Race: White*, a white individual is expected to have -0.0079 log odds of being HIV positive than a black individual, also odds ratio of 0.9921 making it less likely that a white individual's HIV status will be HIV positive. The variable *Employment: Yes* shows that people who are employed are expected to have -0.0529 log odds of being HIV positive than people who are not employed, with an odds ratio of 0.9485. This means that those who are employed are less likely to be HIV positive. *Marital status: Married* and *Province: Free State* are significant predictors of *HIV status* which is the same case as the logistic mixed model with level one and second level predictor variables.

7.7.2 Generalized Logistic Mixed Models with First and Second Level Predictors and Nested Random Effects

With reference to table 7.18 for a generalized mixed logistic model with first and second level predictors and nested random effects, the results show that the total variance of the random effects is 0.0052. Dividing each random effect by the total to get the proportion of the random effect variance attributable to each random effect. By dividing the nested effect variance by the total variance yields the proportion of variance accounted, is $(0.0040/0.0052) * 100 = 76.92\%$. This implies that 76.92% of the total variance of the random effects is attributed to the nested effect. The effect of *Province* alone is $(0.0012/0.0052) * 100 = 23.08\%$.

Table 7.19 shows results from the estimates of the fixed effects. Variable *Education: Tertiary* shows that the odds of an individual who has tertiary education being HIV positive is 0.9559, mean-

Table 7.19: Fixed effects for a generalized logistic mixed model with first and second level predictors and nested random effects.

Fixed effects	**Estimate**	**S.E.**	$t-$**value**	$p-$**value**	**OR**
(Intercept)	-1.4510	0.1062	-13.662	0.0000	0.2343
Age	-0.0009	0.0019	-0.503	0.6148	0.9991
Gender: Male	-0.0350	0.0366	-0.956	0.3389	0.9656
Employment: Yes	-0.0529	0.0395	-1.340	0.1803	0.9485
Education: Secondary	0.0167	0.0418	0.399	0.6896	1.0168
Education: Tertiary	-0.0451	0.0575	-0.785	0.4326	0.9559
Wealth: Middle class	0.0120	0.0479	0.251	0.8020	1.0121
Wealth: Upper class	0.0400	0.0495	0.808	0.4189	1.0408
Marital status: Married	0.1205	0.0383	3.146	0.0016	1.1281
Marital status: Divorce	-0.0849	0.1035	-0.821	0.4119	0.9186
Marital status: Widow	-0.0731	0.1533	-0.477	0.6333	0.9295
Own home: Yes	0.0618	0.0366	1.689	0.0913	1.0637
Contraception: Hormonal	-0.0033	0.0424	-0.077	0.9383	0.9967
Contraception: Barrier	-0.0193	0.0527	-0.367	0.7134	0.9809
Contraception: Emergency	-0.0349	0.0959	-0.364	0.7155	0.9657
Contraception: Permanent	-0.2033	0.0747	-2.723	0.0065	0.8160
Sexual partners: One	0.0165	0.0408	0.404	0.6863	1.0166
Sexual partners: >1	0.0133	0.0594	0.223	0.8233	1.0134
Race: White	-0.0079	0.0512	-0.155	0.8772	0.9921
Race: Coloured	-0.0085	0.0512	-0.167	0.8672	0.9915
Race: Indian	0.0141	0.0512	0.275	0.7831	1.0141
Province: Free State	0.1802	0.0759	2.372	0.0177	1.1974
Province: Gauteng	0.0125	0.0778	0.161	0.8722	1.0125
Province: Kwazulu Natal	-0.0401	0.0776	-0.516	0.6056	0.9607
Province: Limpopo	0.1463	0.0759	1.928	0.0538	1.1575
Province: Mpumalanga	-0.1064	0.0789	-1.349	0.1775	0.8991
Province: Northern Cape	0.0775	0.0756	1.025	0.3055	1.0806
Province: North West	-0.0201	0.0775	-0.259	0.7955	0.9801
Province: Western Cape	0.0229	0.0772	0.297	0.7666	1.0232

Table 7.20: Comparing logistic mixed models (LMM).

	Df	**AIC**	**BIC**	**logLik**	**deviance**	**Chisq**	**Chi Df**	**Pr(>χ^2)**
M_0	3	19459	19482	−9726.3	19453			
M_1	23	19471	19653	−9712.6	19425	27.434	20	0.1235
M_2	31	19465	19710	−9701.3	19403	22.733	8	0.0037

ing individuals having tertiary education are less likely to be HIV positive in relation to individuals who have primary education, holding the other variables constant. The odds of an individual owning a home being HIV positive is 1.0637 times higher, meaning individuals owning a home are more likely to be HIV positive in relation to individuals who do not own homes, holding the other variables constant. When age increases by one year, it corresponds to a -0.0009 decrease to the outcome of an individual's HIV status being HIV positive with the odds ratio of 0.9991, meaning as a person grows older they will have less chances of being HIV positive holding the other variables constant. The results for *Age* variable is actually the same for all the logistic mixed models, multilevel logistic models and generalized logistic mixed models.

7.8 Comparing Model Fits

Models can be compared using the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) which are measure of model quality, followed by the log-likelihood, the deviance for the ML criterion and the deviance for the REML criterion where smaller deviances indicate better fits. The comparison between two or more models is used to check if they are or not significantly different.

Table 7.20 shows the results comparing three logistic mixed models, namely a null intercept only mixed model (M_0), a logistic mixed model with first level predictors (M_1) and a mixed model with first and second level predictors (M_2). Comparing all three models, the null intercept only model seems to be a better fit with the least AIC and BIC value, but with the interest being on the different levels we shall focus on the other two models. The mixed logistic model with first and second level predictors is favourable and shows a significantly improved model fit with $\chi^2_{(1)} =$

Table 7.21: Comparing logistic multilevel models (MLM).

	Df	**AIC**	**BIC**	**logLik**	**deviance**	**Chisq**	**Chi Df**	**Pr(>χ^2)**
M_1	32	19467	19719	−9701.3	19403			
M_2	36	19475	19759	−9701.3	19403	18.596	2	0.0000
M_3	34	19489	19758	−9710.6	19421	0.0000	2	1

Table 7.22: Comparing Generalized logistic mixed models (GLMM).

	Df	**AIC**	**BIC**	**logLik**	**deviance**	**Chisq**	**Chi Df**	**Pr(>χ^2)**
M_1	30	19310	19547	−9625.0	19250			
M_2	31	19550	19795	−9744.2	19488	0	1	1

22.733 and $p-\text{value} < 0.05$. For the logistic mixed model with first and second level predictors, the AIC is 19465 and $p-\text{value} < 0.05$, whereas for the logistic mixed model with first level predictors the AIC is 19471 and $p-\text{value} > 0.05$. For this data, it is better to fit a mixed model with first and second level predictors.

Table 7.21 shows the results for comparing three logistic multilevel models, a multilevel model with first and second level predictors and nested random effects (M_1), a multilevel model with first and second level predictors and random slopes (M_2) and a multilevel model with first and second level predictors, with random slopes and cross-level interaction (M_3). Since the values for both the AIC and BIC are smaller for a multilevel model with first and second level predictors and nested random effects compared to the other two models, we would conclude that it provides a better fit to the data. Lower AIC and BIC are better, since higher deviances mean that the model is not fitting the data well.

Table 7.22 displays the results from comparing two generalized logistic mixed models, a generalized logistic mixed model with first and second level predictor variables (M_1) and a generalized linear mixed model with first and second level predictors and nested random effects (M_2). Since the values for both the AIC and BIC are smaller for a generalized logistic mixed model with first and second level predictor variables, we would conclude that it provides a better fit to the data compared to the other model.

Chapter 8

Conclusion and Recommendations

8.1 Conclusion

Complex sampling methods were defined and explained in the literature chapter, especially how these methods may cause complications in the structure of the data and their analysis thereof. The study then evaluated several approaches that are proposed for modelling and analysing complex survey data. A modelling approach was adopted for this research study and thus survey logistic regression models were fitted. These models include the mixed logistic models, multilevel logistic models and generalized mixed logistic models. These models were designed to account for the complex survey design responsible for possible violation of the independence assumption in the data rendering the conventional statistical methods inappropriate. Survey data were simulated to illustrate, explore and fit different models to complex survey data based on the theory of using a modelling approach for analysis of complex survey data. The models used in the study, particularly the generalized linear models and their extensions multilevel and mixed effects models are important in order to explicitly take into account for the multi-level structured data and clusters nature of most practical survey data.

As per the results presented and discussion given in Chapter 7, the following conclusions were reached:

- A logistic mixed model with first and second level predictors had a better fit compared to a logistic mixed model with first level predictors. This means that an addition of the second level predictors variables made the model better.

- A logistic multilevel model with first and second level predictors and nested random effects provided a better fit to the data, compared to the other two models where there is a multilevel model with first and second level predictors and random slopes and a multilevel model with first and second level predictors, with random slopes and cross-level interaction. The later two models did not fit the data better due to their high deviances.

- The outcomes with regards the generalized logistic mixed models, namely a generalized logistic mixed model with first and second level predictor variables and a generalized linear mixed model with first and second level predictors and nested random effects, showed that the generalized logistic mixed model with first and second level predictor variables provides a better fit to the data compared to the other model.

8.2 Recommendations and Future Research

The results from this study can be used as a basis for evaluating more approaches and strategies to the analysis of complex survey data. Issues that need to be further developed and investigated in future research include:

- There are two fundamental approaches used for analyzing complex survey data, the design-based and the model-based approach, for this study only considers the model-based approach. For future research it would be interesting to compare and contrast both these approaches.

- Demonstrate other complexities of survey data besides the complex sampling designs.

- For further analysis it would have been ideal to run more simulations using a larger population and sample size.

- One of the possible limitation of the study would be in the use of simulated data that might not completely account for all the dynamics inherent in natural population.

- Another limitation and possible future research point is that the nature of the relationship between the response and some of the covariates can also be explored further. For instance, age may have a nonlinear relationship with HIV status, and thus semi-parametric models can be considered.

References

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. Wiley, 2nd edition.

Agresti, A., Booth, J. G., Hobert, J. P., and Caffo, B. (2000). Random-Effects Modeling of Categorical Response Data. *Sociological Methodology*, 30:27–80.

Archera, K. J., Lemeshow, S., and Hosmer, D. W. (2007). Goodness-of-Fit Tests for Logistic Regression Models when Data are Collected Using a Complex Sampling Design. *Computational Statistics & Data Analysis*, 51:4 450–4 464.

Bates, D. (2006). lmer, p-value and all that. Technical report.

Benedetti, A., Platt, R., and Atherton, J. (2014). Generalized Linear Mixed Models for Binary Data: Are Matching Results from Penalized Quasi-Likelihood and Numerical Integration Less Biased? *PlosOne*, 9(1).

Binder, D. A. (1983). On the Variances of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Institute*, 51(3):279–292.

Brick, J. and Kalton, G. (1996). Handling Missing Data in Survey Research. *Statistical Methods in Medical Research*, 5:215–238.

Burgoyne, A. D. and Drummond, P. D. (2008). Knowledge of HIV and AIDS in Women in Sub-Saharan Africa. *African Journal of Reproductive Health*, 12(2):14–31.

Burton, A., Altman, D. G., Royston, P., and Holder, R. L. (2006). The Design of Simulation Studies in Medical Statistics. *Statistics in Medicine*, 25:4 279–4 292.

Cassel, C., Sarndal, C. E., and Wretman, J. H. (1977). *Foundations of Inference in Survey Sampling*. Wiley.

Cassy, S. R., Natario, I., and Martins, M. R. (2016). Logistic Regression Modelling for Complex Survey Data with an Application for Bed Net Use in Mozambique. *Open Journal of Statistics*, 6:898–907.

Chambers, R. and Skinner, C. (2003). *Analysis of Survey Data*. John Wiley & Sons, Ltd, 1st edition.

Chinomona, A. and Mwambi, H. (2015). Multiple Imputation for Non-Response when Estimating HIV Prevalence Using Survey Data. *BMC Public Health*.

Cochran, W. (1977). *Sampling Techniques*. New York: John Wiley and Sons, 3rd edition.

Davis, J., Eisenhardt, K., and Bingham, C. (2007). Developing Theory Through Simulation Methods. *Academy of Management Review*, 32(2):480–499.

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*. Chapman & Hall/CRC, 2nd edition.

Douglas, B., Martin, M., Ben, B., and Steve, W. (2014). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*.

Duesberg, P. H. (1989). Human Immunodeficiency Virus and Acquired Immunodeficiency Syndrome: Correlation but not Causation. *Proceedings of the National Academy of United States of America*, 86:755–764.

Dyson, T. (2003). HIV/AIDS and Urbanization. *Population and Development Review*, 29(3):427–442.

Elsayir, H. A. (2014). Comparison of Precision of Systematic Sampling with some other Probability samplings. *American Journal of Theoretical and Applied Statistics*, 3:111–116.

Faraway, J. J. (2016). *Linear Models with R*. CRC Press Taylor & Francis Group, 2nd edition.

Fienberg, S. (1989). Modeling Considerations: Discussion from a Modeling Perspective. in Panel Surveys. *Wiley: New York*, pages 566–574.

Finch, W. H., Bolin, J. E., and Kelley, K. (2014). *Multilevel Modeling Using R*. CRC Press.

Fuller, W. A. and Rao, J. N. K. (1978). Estimation for a Linear Regression Model with Unknown Diagonal Covariance Matrix. *The Annals of Statistics*.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gilbert, N. and Troitzsch, K. (2005). *Simulation for the Social Scientist*. Maidenhead: Open University Press, 2nd edition.

Goldstein, H. (1991). Multilevel Modelling of Survey Data. *Journal of the Royal Statistical Society*, 40(2):235–244.

Goldstein, H. (2003). *Multilevel Statistical Models*. Hodder Headline Group, 3rd edition.

Goldstein, H. (2011). *Multilevel Statistical Models*. John Wiley & Sons Ltd, 4th edition.

Graubard, B. I. and Korn, E. L. (2002). Inference for Superpopulation Parameters Using Sample Surveys. *Statistical Science*, 17(1):73–96.

Green, P. J. (1984). Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some Robust and Resistant Alternative. *Royal Statistical Society*, 46(2):149–192.

Guo, G. and Zhao, H. (2000). Multilevel Modeling for Binary Data. *Annual Reviews*, 26:441–462.

Hallgren, K. A. (2014). Conducting Simulation Studies in the R Programming. *Tutorials in Quantitative Methods for Psychology*, 9(2):43–60.

Hansen, M., Madow, W., and Tepping, B. (1983). An Evaluation of Model-Dependent and Probability Sampling Inferences in Sample Surveys. *Journal of the American Statistical Association*, 78:776–793.

Hargovan, K. A. (2007). Inference from Finite Population Sampling a Unified Approach. Master's thesis, University of KwaZulu Natal.

Harville, D. A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association*, 72(358):320–338.

Heeringa, S., West, B., and Berglund, P. (2010). *Applied Survey Data Analysis*. Chapman and Hall / CRC Press; Boca Raton.

Hoem, J. (1989). The Issue of Weights in Panel Surveys of Individual Behavior. In Panel Surveys. *Wiley*, (540):512–539.

Holt, D., Smith, T. M. F., and Winter, P. D. (1980). Regression Analysis of Data from Complex Surveys. *Journal of the Royal Statistical Society*, 143:474–487.

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

Hox, J. (2002). *Multilevel Analysis: Techniques and Applications.* Lawrence Erlbaum Associates.

Jennrich, R. I. and Sampson, P. F. (1976). Newton-Raphson and Related Algorithms for Maximum Likelihood Variance Component Estimation. *Technometrics*, 18(1):11–17.

Johnson, P., Barry, J. E. S., Ferguson, H., and Muller, P. (2014). Power Analysis for Generalized Linear Mixed Models in Ecology and Evolution. *Methods in Ecology and Evolution*, 6:133–142.

Jorgensen, B. (1983). Maximum Likelihood Estimation and Large-Sample Inference for Generalized Linear and Nonlinear Regression Models. *Biometrika*, 70(1):19–28.

Kalton, G. (1989). Modeling Considerations: Discussion from a Survey Sampling Perspective. In Panel Surveys. *Wiley*, pages 575–587.

Ker, H.-W. (2014). Application of Hierarchical Linear Models/Linear Mixed-Effects Models in School Effectiveness Research. *Universal Journal of Educational Research*, 2(2):173–180.

Kish, L. (1965). *Survey Sampling*. John Wiley & Sons: New York.

Koch, G. G. (1982). *Intraclass Correlation Coefficient*. John Wiley & Sons: New York.

Korn, E. L. and Graubard, B. I. (1999). *Analysis of Health Surveys*. John Wiley & Sons, Inc.

Laird, N. M. and Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics*, 38(4):963–974.

Lee, E. S. and Forthofer, R. N. (2006). *Analyzing Complex Survey Data*. SAGE Publications, 2nd edition.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc, 2nd edition.

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Brooks/Cole Publishing Company, California.

Lohr, S. L. (2010). *Sampling: Design and Analysis*. Brooks/Cole: Boston, 2nd edition.

Lohr, S. L. and Liu, J. (1994). A Comparison of Weighted and Unweighted Analyses in the National Crime Victimization Survey. *Journal of Quantitative Criminology*, 10:343–360.

Lumley, T. (2004). Analysis of Complex Survey Samples. *Journal of Statistical Software*, 9:1–19.

Lumley, T. (2010). *Complex Surveys: A guide to Analysis using R*. John Wiley & Sons, Inc., New Jersey.

Maas, C. J. M. and Hox, J. J. (2005). Sufficient Sample Sizes for Multilevel Modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 1:85–91.

Mason, W. M., Wong, G. Y., and Entwisle, B. (1984). Contextual Analysis Through the Multilevel Linear Model. *Wiley*, 14:72–103.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, 2nd edition.

McCulloch, C. E. (1997). Maximum Likelihood Algorithms for Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 92(437):162–170.

McCulloch, C. E. (2003). Generalized Linear Mixed Models. *NSF-CBMS Regional Conference Series in Probability and Statistics*, 7:1–84.

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008). *Generlalized, Linear and Mixed Models*. Wiley, 2nd edition.

Millar, R. B. (2011). *Maximum Likelihood Estimation and Inference with Example in R, SAS and ADMB*. John Wiley and Sons Ltd. United Kingdom, 1st edition.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society*, 135:370–384.

Pellowski, J., Kalichman, S., Matthews, K., and Andler, N. (2013). A Pandemic of the Poor: Social Disadvantage and the U.S. HIV Epidemic. *American Psychologist*, 68:197–209.

Pfeffermann, D. (1993). The Role of Sampling Weights when Modelling Survey Data. *International Statistical Review*, 5:317–37.

Pfeffermann, D. (2011). Modelling of Complex Survey Data: Why model? Why is it a problem? How can we approach it? *Statistics Canada,*, 37(2):115–136.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., and Rasbash, J. (1998). Weighting for Unequal Selection Probabilities in Multilevel Models. *Journal of the Royal Statistical Society*, 60(1):23–40.

Rabe-Hesketh, S. and Skrondal, A. (2006). Multilevel Modeling of Complex Survey Data. *Journal of the Royal Statistical Society*, 169:805–827.

Royall, R. (1976). Likelihood Functions in Finite Population Sampling Theory. *Biometrika*, 63:605–614.

Rozi, S., Mahmud, S., Lancaster, G., Hadden, W., and Pappas, G. (2017). Multilevel Modeling of Binary Outcomes with Three-Level Complex Health Survey Data. *Open Journal of Epidemiology*, 7:27–43.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrik*, 63:581–592.

Rueda, M. and Sanchez-Borrego, I. R. (2009). A Predictive Estimator of Finite Population Mean using Nonparametric Regression. *Computational Statistics*, 24:1–14.

Samphath, S. (2005). *Sampling Theory and Methods*. Alpha Science International Ltd., 2nd edition.

Schur, C. L., Berk, M., Dunbar, J. R., Shapiro, M. E., Cohn, M. E., and Bozzette, S. A. (2012). Where to seek care: An examination of people in rural areas with HIV/AIDS. *The Journal of Rural Health*, 18:337–347.

Skinner, C. J., Holt, D., and Smith, T. (1989). *Analysis of Complex Surveys*. Wiley.

Snijders, T. and Bosker, R. (1999). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. SAGE publications.

Stirling, W. D. (1984). Iteratively Reweighted Least Squares for Models with a Linear Part. *Journal of the Royal Statistical Society*, 33(1):7–17.

Stroup, W. W. (2013). *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. Taylor & Francis Group.

Sullivan, L. M., Dukes, K. A., and Losina, E. (1999). An introduction to hierarchical linear modelling. *Statistics in Medicine*, 18:855–888.

Szumilas, M. (2010). Explaining Odds Ratios. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, 19:227–229.

Vallient, R. (2009). Model-Based Prediction of Finate Population Totals. *Sample Surveys: Inference And Analysis*, 29B:11–31.

West, B. T. (2008). Statistical and Methodological Issues in the Analysis of Complex Sample Survey Data: Practical guidance for trauma researchers. *Journal of Traumatic Stress*, 21:440–447.

West, B. T., Sakshaug, J. W., and Aurelien, G. A. S. (2018). Accounting for Complex Sampling in Survey Estimation: A Review of Current Software Tools. *Journal of Official Statistics*, 34(3):721–752.

Wolfinger, R. and O'Connell, M. (1993). Generalized Linear Mixed Models a Pseudolikelihood Approach. *Journal of Statistical Computation and Simulation*, 48:3–4.

Wong, G. Y. and Mason, W. M. (1985). The Hierarchical Logistic Regression Model for Multilevel Analysis. *Journal of the American Statistical Association*, 80(391):513–524.

Wu, L. (2010). *Mixed Effects Models for Complex Data*. Chapman and Hall/CRC: New York:, 1st edition.

Zhang, L.-C. (2008). On some Common Practices of Systematic Sampling. *Journal of Official Statistics*, 24:557–569.

Appendix A: R Code

RStudio®

```
LMMs,GLMMs, MLMs with simulated complex survey data
library(sampling)
library(survey)
library(lme4)
library(lmerTest)# to get pvalues for lmer & glmer set.seed(12345)
Section 6:Create Simulated Data
x=rnorm(20000) p<- exp(x)/(1+exp(x))
y_HIVstatus<- rbinom(20000,1,prob = 0.19)
#1 Positive, 0 Negative
mean(y_HIVstatus)
x1_Age <- sample(18:50, 20000, replace = TRUE)
#Simulate ages between 18-50
x2_gender <- rbinom(20000,1,p)
#1 Male, 0 Females
x3_Employment <- rbinom(20000,1,prob = 0.71)
#1 Yes, 0 No
x4_Education <-sample(1:3, size=20000, replace=TRUE,
prob =c(0.62,0.26,0.115))
# 1-Primary phase, 2- Secondary phase,3- tertiary Phase
```

```
x5_Wealth <- sample(1:3, size = 20000, replace = TRUE,
prob =c(0.66,0.18,0.16))
#1-lower class,2-middle class, 3-upper class
x6_MaritalStatus <-sample(1:4,size = 20000,replace = TRUE,
prob =c(0.62,0.33,0.035,0.015))
#1-Single,2-Married,3-Divorced,4-widow
x7_ResidentialArea <- sample(1:3, size=20000, replace=TRUE)
# 1-urbans, 2-Village,3- Township
x8_OwnHome <- rbinom(20000,1,p)
# Does the family own their home?: 0 = no, 1 = yes
x9_Methods_of_contraception<-sample(0:4, size = 20000,replace=TRUE,
prob =c(0.40,0.32,0.16,0.04,0.08))
#0-no contraception,1-hormonal contraception,2-barrier method,
3- emergency contraception,4-permanent contraception
x10_Sexual_partners<-sample(0:2,size =20000,replace=TRUE,
prob=c(0.30,0.57,0.13))#0-none, 1-one partner,2-more than 1 partner
x11_Race<-sample(1:4, size = 20000, replace=TRUE)
#1-black, 2- white=, 3-coloured,4-indian,
x12_Province<-sample(1:9, size = 20000, replace=TRUE)
#1-EC,2-FS,3-GP,4-KZN,5-LP,6-MP,7-NC,8-NW,9-WC
Simulated_data <-data.frame(y_HIVstatus,x1_Age,x2_gender,
x3_Employment,x4_Education,x5_Wealth,x6_MaritalStatus,
x7_ResidentialArea,x8_OwnHome,x9_Methods_of_contraception,
x10_Sexual_partners,x11_Race,x12_Province)
View(Simulated_data)
#Section 7.5: Fittiting the data using Linear Models (lm) and
Generalised Linear Model(glm WITHOUT the svyglm)f
```

```
or a Binomial family
model_lm <-lm(y_HIVstatus~x1_Age+factor(x2_gender)+
factor(x3_Employment)+factor(x4_Education)+factor(x5_Wealth)+
factor(x6_MaritalStatus)+factor(x7_ResidentialArea)+factor(x8_OwnHome)+
factor(x9_Methods_of_contraception)+factor(x10_Sexual_partners)+
factor(x11_Race), data=Simulated_data)
summary(model_lm)
model_glm <-glm(y_HIVstatus~x1_Age+factor(x2_gender)+
factor(x3_Employment)+factor(x4_Education)+factor(x5_Wealth)+
factor(x6_MaritalStatus)+factor(x7_ResidentialArea)+factor(x8_OwnHome)+
factor(x9_Methods_of_contraception)+factor(x10_Sexual_partners)+
factor(x11_Race),family="binomial", data=Simulated_data)
summary(model_glm)
#Section 7.5: Svg_GLMS with simulated complex survey data
#Fitting a generalised linear model to data from a "complexsurveydesign",
#Stratified Independent Sampling design design_strat<-svydesign
(id=~1,probs = ~p, data=Simulated_data)
model_svyglm<-svyglm(y_HIVstatus~x1_Age+factor(x2_gender)+
factor(x3_Employment)+factor(x4_Education)+factor(x5_Wealth)+
factor(x6_MaritalStatus)+factor(x7_ResidentialArea)+factor(x8_OwnHome)+
factor(x9_Methods_of_contraception)+factor(x10_Sexual_partners)+
factor(x11_Race),family="quasibinomial",design=design_strat)
summary(model_svyglm)
#Logistic Mixed Modelling (LMM)
#Section 7.6:The logistic Mixed Model (LMM) null intercept only model
model_lmm0<-lmer(y_HIVstatus ~ 1+(1|factor(x7_ResidentialArea)),
data=Simulated_data)
```

```
summary(model_lmm0)
#Section 7.6: logistic Mixed Model (LMM)with First Level Predictors
model_lmm1<-lmer(y_HIVstatus ~ (1|factor(x7_ResidentialArea))+ x1_Age+
factor(x2_gender)+factor(x3_Employment)+factor(x4_Education)+
factor(x5_Wealth)+factor(x6_MaritalStatus)+factor(x8_OwnHome)+
factor(x9_Methods_of_contraception)+factor(x10_Sexual_partners)+
factor(x11_Race),data=Simulated_data)
summary(model_lmm1)
#Section 7.6:Logistic Mixed Modelling with First and Second Level Predictors
model_lmm2<-lmer(y_HIVstatus ~ (1|factor(x7_ResidentialArea))+ x1_Age+
factor(x2_gender)+factor(x3_Employment)+factor(x4_Education)+
factor(x5_Wealth)+factor(x6_MaritalStatus)+factor(x8_OwnHome)+
factor(x9_Methods_of_contraception)+factor(x10_Sexual_partners)+
factor(x11_Race)+factor(x12_Province),data=Simulated_data)
summary(model_lmm2)
#MULTILEVEL logistic modelling (MLM)
#Section 7.6: Using the lmer function for multilevel model (MLM)
#mlm lmer model with First and Second Level Predictors
#and nested random effects
model_mlm1<-lmer(y_HIVstatus ~ (1|x12_Province/x7_ResidentialArea)+
x1_Age+factor(x2_gender)+factor(x3_Employment)+factor(x4_Education)+
factor(x5_Wealth)+factor(x6_MaritalStatus)+factor(x8_OwnHome)+
factor(x9_Methods_of_contraception)+factor(x10_Sexual_partners)+
factor(x11_Race)+factor(x12_Province),REML=TRUE, data=Simulated_data)
summary(model_mlm1)
#Section 7.6:MLM First and Second Level Predictors with Random Slopes
model_mlm2<-lmer(y_HIVstatus ~ 1+(1+factor(x9_Methods_of_contraception))+
```

```
(factor(x10_Sexual_partners)|factor(x7_ResidentialArea))+x1_Age+
factor(x2_gender)+factor(x3_Employment)+factor(x4_Education)+
factor(x5_Wealth)+factor(x6_MaritalStatus)+factor(x8_OwnHome)+
factor(x9_Methods_of_contraception)+factor(x10_Sexual_partners)+
factor(x11_Race)+factor(x12_Province),REML=TRUE, data=Simulated_data)
summary(model_mlm2)
#Section 7.6:MLM with First and Second Level Predictors,
#with Random Slopes and Crosslevel Interaction
model_mlm3<-lmer(y_HIVstatus ~ 1+x9_Methods_of_contraception:x12_Province+
(1+factor(x9_Methods_of_contraception)|factor(x7_ResidentialArea))+x1_Age+
factor(x2_gender)+factor(x3_Employment)+factor(x4_Education)+factor(x5_Wealth)+
factor(x6_MaritalStatus)+factor(x8_OwnHome)+factor(x10_Sexual_partners)+
factor(x11_Race),REML=TRUE, data=Simulated_data)
summary(model_mlm3)
#Generalized logistic mixed models (GLMMs)
#Section 7.7: Using glmer function for generalized mixed models with
#with first and second level predictor variables
model_glmer1 <- glmer(y_HIVstatus ~(1|factor(x7_ResidentialArea))+x1_Age+
factor(x2_gender)+factor(x3_Employment)+factor(x4_Education)+
factor(x5_Wealth)+factor(x6_MaritalStatus)+factor(x8_OwnHome)+
factor(x9_Methods_of_contraception)+ factor(x10_Sexual_partners)+
factor(x11_Race)+factor(x12_Province), data=Simulated_data,
family=binomial(link="logit"))
summary(model_glmer1)
#Section 7.7: glmer model- generalized logistic mixed model
#with first and second level predictors and nested random effects
model_glmer2<- glmer(y_HIVstatus ~(1|x12_Province/x7_ResidentialArea)+
```

```
x1_Age+factor(x2_gender)+factor(x3_Employment)+factor(x4_Education)+
factor(x5_Wealth)+ factor(x6_MaritalStatus)+factor(x8_OwnHome)+
factor(x9_Methods_of_contraception)+ factor(x10_Sexual_partners)+
factor(x11_Race)+factor(x12_Province),family=binomial(link="logit"),
data=Simulated_data)
summary(model_glmer2)
Section 7.8: Comparing models
The anova() command computes analysis of variance (or deviance) tables.
anova(model_lmm0,model_lmm1,model_lmm2)
anova(model_mlm1,model_mlm2,model_mlm3)
anova(model_glmer1,model_glmer2)
```